Integrated models for predicting start-ups evolution

Mariia Garkavenko

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'information, Informatique
Spécialité : Mathématiques et Informatique
Unité de recherche : Laboratoire d'Informatique de Grenoble

Modèles intégrés pour prévoir l'évolution des entreprises en démarrage

Integrated models for predicting start-ups evolution

Présentée par :

Mariia GARKAVENKO

Direction de thèse :

Eric GAUSSIER	Directeur de thèse

Rapporteurs :

CHRISTINE LARGERON
Professeur des Universités, UNIVERSITE DE SAINT-ETIENNE - JEAN MONNET
CELINE ROBARDET
Professeur des Universités, INSA LYON

Thèse soutenue publiquement le **22 juin 2022**, devant le jury composé de :

ERIC GAUSSIER Professeur des Universités, UNIVERSITE GRENOBLE ALPES	Directeur de thèse
CHRISTINE LARGERON Professeur des Universités, UNIVERSITE DE SAINT-ETIENNE - JEAN MONNET	Rapporteure
CELINE ROBARDET Professeur des Universités, INSA LYON	Rapporteure
VERONIQUE BLUM Maître de conférences HDR, UNIVERSITE GRENOBLE ALPES	Examinatrice
JEAN-MICHEL DALLE Professeur des Universités, SORBONNE UNIVERSITE	Examinateur
SIHEM AMER-YAHIA Directeur de recherche, CNRS DELEGATION ALPES	Présidente

Invités :

CÉDRIC LAGNIER
Ingénieur Docteur, Société Skopai
HAMID MIRISAEE
Ingénieur Docteur, Société Skopai

Université Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Informatique**

Arrêté ministériel : 25 mai 2016

Présentée par

Mariia Garkavenko

Thèse dirigée par **Eric Gaussier**, Professeur, Université Grenoble Alpes

préparée au sein du **Laboratoire d'Informatique de Grenoble**
dans **l'École Doctorale Mathématiques, Sciences et technologies de l'information, Informatique**

Valorisation et levée de fonds de start-up

Startup valuation and fundraising

Thèse soutenue publiquement le **22 juin 2022**,
devant le jury composé de :

Madame Christine Largeron
Professeure, Université Jean Monnet, Rapporteuse

Madame Céline Robardet
Professeure, Institut National des Sciences Appliquées de Lyon, Rapporteuse

Madame Véronique Blum
Maître de conférence, Université Grenoble Alpes, Examinatrice

Monsieur Jean-Michel Dalle
Professeur, Sorbonne Université, Examinateur

Madame Sihem Amer-Yahia
Directrice de Recherche Première Classe, Centre National de Recherche scientifique, Examinatrice

Monsieur Cédric Lagnier
CTO, Skopai, co-encadrant de thèse, Invité

Monsieur Hamid Mirisaee
Data Scientist, Skopai, co-encadrant de thèse, Invité

Monsieur Eric Gaussier
Professeur,Université Grenoble Alpes, Directeur de thèse

Abstract / Résumé

Abstract

Startups play an increasingly important role in the modern economy. In this thesis, we study startup valuation and fundraising problems with machine learning and causal discovery methods. After reviewing the existing machine learning approaches to startup success prediction and the literature on startup valuation factors, we present a domain adaptation-based approach to predict startup valuations in funding rounds with known funding amounts. We show that funding rounds in which startup valuation is announced to the public are statistically different from those in which the valuation is kept secret. We mine a novel data source, Companies House, to learn the startup valuation in the later funding rounds and show that domain adaptation methods yield the best results for our task. Further, we collect a rich dataset of United Kingdom startups and their valuations and discover which variables make the best valuation predictors. Also, we apply causal discovery methods to learn which variables, directly and indirectly, affect startup valuation. We draw the connection to the previous startup valuation factors research and provide evidence for further theoretical studies. Finally, we propose a method for predicting whether a startup will secure a funding round based on publicly freely available information on the web. We propose methods to collect information about the startups and their funding rounds from different sources. Since it is impossible to collect the information about all the funding rounds, we propose to tackle the funding round prediction problem in the positive-unlabeled setting and show that this setting is beneficial for the neural network model.

Résumé

Les startups jouent un rôle de plus en plus important dans l'économie moderne. Dans cette thèse, nous étudions la valorisation et le financement des startups à l'aide de méthodes d'apprentissage automatique et de découverte causale. Après avoir étudié les méthodes d'apprentissage automatique existantes pour prédire le succès des startups et la littérature sur les facteurs impactant la valorisation des startups, nous présentons une approche basée sur l'adaptation de domaine pour prédire la valorisation des startups au moment d'une levée de fonds dont le montant est connu. Nous montrons que les levées de fonds dans lesquelles la valorisation de la startup est annoncée publiquement sont statistiquement différentes de celles dans lesquelles la valorisation est gardée secrète. Nous exploitons une nouvelle source de données, "Companies House", pour apprendre à estimer la valorisation des startups lors des derniers tours de financement et nous montrons que les méthodes d'adaptation de domaine donnent les meilleurs résultats dans ce contexte. Cette source nous a permis de collecter un important jeux de données sur les startups du Royaume-Uni et leurs valorisations, nous permettant de découvrir quelles caractéristiques constituent les meilleurs prédicteurs de valorisation. De plus, nous appliquons des méthodes de découverte causale pour apprendre, à partir de ces données, quelles caractéristiques affectent, directement et indirectement, la valorisation des startups. Nous établissons un lien avec les précédentes recherches sur les facteurs de calcul de valorisation des startups et fournissons des preuves pour des études théoriques supplémentaires. Enfin, nous proposons une méthode pour prédire si une startup obtiendra une levée de fonds en se basant sur des informations publiques, librement disponibles sur le web. Nous proposons des méthodes pour collecter des informations sur les startups et leurs levées de fonds à partir de différentes sources. Puisqu'il est impossible de collecter des informations sur toutes les levées de fonds, nous proposons d'aborder le problème de leur prédiction comme un cas d'apprentissage positifs et indéterminés et montrons que ce type d'approche est bénéfique pour des modèles de réseaux de neurones.

Contents

Abstract / Résumé i

1 Introduction 1

1.1 Practical methods of startup valuation 4

1.2 Problem statement 5

1.3 Thesis outline 7

1.4 Corresponding Articles 7

2 Background 9

2.1 Machine learning for startup studies 9

2.2 Startup valuation factors 14

3 Startup Valuation 19

3.1 Data Collection 19

3.1.1 CrunchBase 20

3.1.2 Companies House 21

3.1.3 Twitter API 24

3.1.4 Google Search API 24

3.2 Predicting European startups valuation in a funding rounds via domain adaptation. 24

3.2.1 Data analysis and problem formulation 25

3.2.2 Approach 31

3.2.3 Experiments 38

3.2.4 Concluding remarks 43

3.3 Assessing the Determinants of Start-up Valuation through Prediction and Causal Discovery. 44

3.3.1 Variables . . . 46
3.3.2 Machine Learning Model . . . 49
3.3.3 Causal Discovery . . . 53
3.3.4 Experiments . . . 58
3.3.5 Discussion . . . 73
3.4 Conclusion . . . 79

4 Startup Fundraising **81**
4.1 Data Collection . . . 82
4.1.1 Features . . . 82
4.1.2 Data labeling . . . 88
4.2 Prediction models . . . 90
4.2.1 Positive-Unlabeled setting . . . 90
4.2.2 Positive-Negative setting . . . 93
4.3 Evaluation . . . 94
4.3.1 Data split and metrics . . . 95
4.3.2 Positive-Unlabeled results . . . 96
4.3.3 Positive-Negative results . . . 97
4.3.4 Ablation Analysis . . . 99
4.3.5 Feature Importance . . . 100
4.4 Conclusion . . . 102

5 Conclusion **105**
5.1 Summary of contributions . . . 105
5.2 Future Work . . . 106

6 Appendix **109**
6.1 Business register sources . . . 109
6.2 Some examples of sources for data labeling . . . 111

Bibliography **113**

1

Introduction

In recent years, startups have played an increasingly prominent role in world economics, bringing disruptive innovation to various markets. Innovative new ventures are essential contributors to the national economy and the sources of economic growth, job creation, innovation, and technological change. Startups intended to achieve high growth can boost the economy with revolutionary technologies and business models and create new markets over time. Often, startups rely on venture capitalists, business angels, equity crowdfunding platforms, and other equity investors to develop new technologies, hire teams and scale aggressively. The volume of Venture Capital (VC) invested in startups is astonishing and rapidly growing - $643 billion in 2021 compared to $335 billion for 2020 and less than $ 100 billion in 2014, according to Crunchbase[1]. High growth firms make up only several percent of the firms' population, but they create up to 60% of new jobs, which makes them a major driving force of job creation (OECD and Commission, 2021).

In this thesis, we study the problem of startup valuation with machine learning and causal discovery methods. Before going into more details, we provide the following definitions that will be frequently used throughout the paper:

- *Startup*: initially a small company with an innovative idea that potentially can disrupt the market and get large revenue. Most startups seek external funding in order to develop prototypes, test ideas, and scale up their business aggressively.

- *Startup valuation*: the process of determining how much a startup is valuated economically.

- *Equity*: percentage of ownership in a company.

[1] https://news.crunchbase.com/news/global-vc-funding-unicorns-2021-monthly-recap/ last visited on 11/04/2022

- *Funding amount*: the amount of money invested in a funding round.
- *Funding round*: a discrete fundraising event for a company, during which the company raises financing at a certain valuation. In early stages called Angel and Seed funding rounds, startups often obtain funding from friends, family, and wealthy individuals called business angels. The funding amount in Angel and Seed rounds is typically between $10k and $2M. At the later stages, such as Series A, Series B, and so on, Venture Capital professional firms come into play with much more significant amounts of money - tens of million dollars and more strict due diligence and security checks.

Access to financial capital is crucial for startups to test ideas, develop a team, and fund early-stage projects. To raise money, a startup needs to be valued. This, together with the amount of money invested, determines the proportion of shares of the company owned by the investors. For founders and investors, valuation allows tracking the effectiveness of strategic decision-making processes and venture performance in terms of the estimated change in value. Valuation also drives the motivation of entrepreneurs and sets a value to the efforts and resources they put into a new business (Miloud, Aspelund, and Cabrol, 2012). Furthermore, the ability to distinguish successful projects among many business ideas is of economic and social importance, as it would enable society to allocate funds to those projects that have the potential to be the most profitable in the future (Csaszar, Nussbaum, and Sepulveda, 2006).

Equity fundraising is intrinsically very different from debt financing familiar to us in everyday life. The essence of debt financing is that the person or firm must return the loaned money with interest within a certain time. In the equity funding process, an investor buys a share of a startup in the hope that in the (usually distant) future, this share could be sold at a much higher cost. While creditor almost always gets the money back, a business angel or venture capitalist, more often than not, will witness the failure of the funded startup and lose the invested money. Indeed, according to a widely accepted rule of thumb, nine out of ten early-stage startups fail. On the other hand, successful startups yield to their early investors giant returns.

For these reasons, startup valuation is a challenging task for investors intertwined with the startup success prediction task. The startup valuation

and success prediction problems can be approached from many different angles using instruments from various domains. In particular, with the recent advances in machine learning, many problems have been reformulated as prediction tasks which can then be solved by machine learning approaches. Over the last few years, various tasks related to startups have been studied with machine learning methods. For instance (Xiang, Zheng, Wen, Hong, Rose, and Liu, 2012) aims to predict Mergers & Acquisition deals using news data, Antretter, Blohm, Grichnik, and Wincent, 2019 predicts young venture survival, and Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke, 2018 predicts whether a startup will go from Seed to Series A funding stage within a year.

While many researchers investigated startup success problem, most works were conducted in the binary setting, namely failure/success prediction. However, in reality, there are more shades. A startup with a multi-billion valuation at IPO is undoubtedly a success for investors. However, a startup acquired by a large company for a dozen million might also be a success for a business angel that invested in it at a valuation of a million. Therefore the primary goal of equity investors is that the valuation of a startup they invest in increases in the future.

How to value a new venture is critical in entrepreneurial finance (Blohm, Antretter, Sirén, Grichnik, and Wincent, 2020; Miloud, Aspelund, and Cabrol, 2012). Startup valuation is intrinsically different from the valuation of established companies. "Because startups encounter so many hazards and because they have short-track records by which outsiders can evaluate their potential, there is considerable uncertainty about their value" (Baum and Silverman, 2004, p.415). Because of the high level of risk and often no or little revenues, traditional valuation methods based on quantitative analysis of a company's past financial performance are of little use. Hence, startup valuations are often determined based on various qualitative characteristics.

1.1 Practical methods of startup valuation

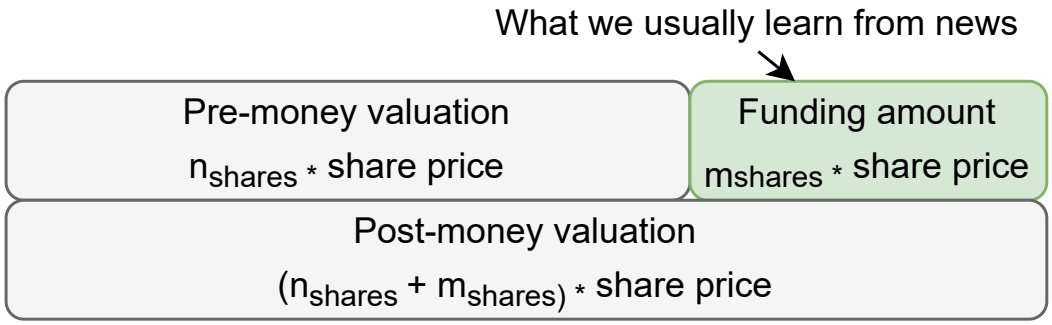
In business practice, various approaches are utilized by investors to value a start-up. For pre-revenue start-ups, the focus is on qualitative criteria and financial projections of performance since the revenue and earnings data are not available. For example, Venture Capital Method estimates the value of a start-up based on terminal value, or the expected selling price for the company at some point in the future, and the expected Return on Investment (ROI) required by the investors (Sahlman and Scherlis, 1987). The resulting valuation shows what is required for investors to meet their investment goals. However, it does not look explicitly at such factors as team, product, market, or risks. At the same time, few start-ups manage to meet or exceed the projected revenues in the periods planned. The other methods such as Berkus (Amis and Stevenson, 2001), scorecard, and risk factor summation methods were developed specifically for the early stage investments and without relying upon the founder's financial forecasts. They evaluate start-ups based on a list of criteria, such as quality of the team, sound idea, size of the opportunity, product, and technology, competitive environment, established relationships, and others, as well as various types of risks that may reduce the success of a new venture. Furthermore, the valuation of a start-up is often adjusted to an average valuation of similar companies in the same industry or geographic region.

Once a company is making revenues for any period of time, evaluation agents can use actual revenues to project its value and apply such methods as Discounted Cash Flow (DCF), comparables method, or First Chicago method. These methods are based on projecting the start-up's future cash flows and discounting them to get their present value (DCF method), comparing referential information on indicators of other similar funded start-ups with a target start-up to estimate its value (comparables method), or a combination of these methods, accounting also for the best- and worst-case scenarios (First Chicago Method). Overall, given a multitude of different approaches for establishing a valuation of early-stage companies and the variation in valuation results, a combination of multiple valuation factors and methods may yield more robust estimations.

1.2 Problem statement

In our work, we aim to approach the problem of startup valuation with Artificial Intelligence methods. While stocks of public companies are traded daily, and the value of a company can be calculated at any moment, the shares of a startup are rarely sold, and the valuation of a startup is documented only when particular events occur. These events include funding rounds, Merger and Acquisition deals (M&A), and Initial Public Offerings (IPO). In this thesis, we focus on the valuations obtained during the funding rounds since they are much more frequent than IPO and M&A.

Figure 1.1a gives an overview of funding round process. As well as public companies, startups have shares. In funding round, a startup issues new shares, and the investor buys them at some share price. To enter a funding round investor and startup must agree on the pre-money valuation of the startup($= n_{shares} \times$ share price). This agreement will determine how much equity the investor will receive for its money.



(a)


Hugging Face raises $40 million for its natural language processing library

Romain Dillet @romaindillet / 4:11 PM GMT+1 • March 11, 2021

(b)

Figure 1.1: (a) Funding round: pre-money and post-money valuation difference. (b) News about the funding round example: we learn that Hugging Face raised $ 40 million.

The problem of startup valuation in a funding round can be approached from different angles: the first is to infer the hidden valuation in a funding round, for which only the funding amount was announced publicly. This setting is interesting because the vast majority of funding rounds reported in news miss valuation data as in exemplary funding round announcement in the news on Figure 1.1b. Knowledge of startup valuation in a funding round might be useful for researchers as a proxy for valuation information to model temporal change of valuation, for example. At the same time, it might be interesting for practitioners. For instance, both investors and entrepreneurs might look at similar startups' funding rounds valuations to estimate the market. The practitioners' interest in the estimate of startup valuation in funding rounds is highlighted by the fact that popular databases dedicated to startups, such as Crunchbase, provide the information about the estimated startup valuation for an additional fee to its users.

Another angle to approach the startup valuation problem with machine learning would be to try to build an automated startup valuation method that investors could use. However, building such an algorithm would be extremely difficult because the investors in a funding round usually have access to information about the startup that is not readily available to the greater public. On the other hand, even the not-hundred percent accurate model could be used by investors to perform, for example, the initial screening of the startups. Indeed, typically investors have some constraints on the funding amount they are willing to invest – the ones focusing on early stage ventures do not have the amount of money sufficient to invest in a unicorn, and the investors that focus on late stage startups will not want to bother with a nascent startup. At the same time, venture capitalists spend a considerable amount of time identifying and monitoring startups. Thus it might be of great use to an investor to filter out the startups with valuations far outside the investor's deal range.

Finally, instead of predicting startup valuation, we might ask ourselves what the crucial factors for startup valuation are. To get insight from the data, we may use explainable machine learning techniques as well as the methods designed specifically for identifying causal relations from data *i.e.* causal discovery methods.

1.3 Thesis outline

The present manuscript is organized as follows:

Chapter 2 gives an overview of the literature on the applications of artificial intelligence methods to startup success and valuation studies.

Then, in Chapter 3, we study the problem of startup valuation with machine learning and causal discovery methods. In particular, Section 3.2 addresses the problem of inferring unknown startup valuation in a funding round with a known amount via domain adaptation framework. Then, in Section 3.3 we investigate the factors that affect startup valuation via both machine learning and causal discovery methods.

Finally, Chapter 4 explores the possibility of building a model for startup success prediction from publicly available web data without the use of proprietary databases such as CrunchBase.

1.4 Corresponding Articles

The contribution of this thesis includes the following articles, prepared during the postgraduate studies:

Journals

- Mariia Garkavenko, Tatiana Beliaeva, Eric Gaussier, Hamid Mirisaee, Agnès Guerraz, and Cédric Lagnier (2022). "Assessing the Determinants of Start-up Valuation through Prediction and Causal Discovery". In: *Entrepreneurship Theory and Practice*, in press

Peer-reviewed international conferences

- Mariia Garkavenko, Hamid Mirisaee, Eric Gaussier, Agnès Guerraz, and Cédric Lagnier (2021). "Valuation of Startups: A Machine Learning Perspective". In: *Proceedings of the 43rd European Conference on Information Retrieval*. Springer, pp. 176–189

Other contributions

- Tatiana Beliaeva, Mariia Garkavenko, Hamid Mirisaee, Eric Gaussier, Agnès Guerraz, and Cédric Lagnier (2020). "Estimating startup valuation with AI: Evidence from green technology startups in Europe". In: *European Centre for Alternative Finance Research Conference*. Utrecht University

- Mariia Garkavenko, Hamid Mirisaee, Eric Gaussier, Agnès Guerraz, and Cédric Lagnier (2022). *Where Do You Want To Invest? Predicting Startup Funding From Freely, Publicly Available Web Information*. arXiv: 2204.06479

2 Background

2.1 Machine learning for startup studies

This chapter gives an overview of the literature that studies the application of predictive machine learning techniques to the problems related to startups. Over the last few years, various tasks related to startups have been studied with machine learning methods by researchers from both the computer science field, which seek new applications for the machine learning methods, and the social science researchers, notably from the field of entrepreneurship that aim to understand the process that the startups undergo better with novel methods. Table 2.1 gives overview of some most prominent studies in the area.

Notably, the existing research primarily aims to predict whether a particular startup will succeed in the future and thus solve binary classification task. However, defining what startup success corresponds to is not straightforward, and there is no way to measure success directly. Therefore a number of proxies for startup success were proposed in the literature. These proxies are detailed below:

- *Startup exit: Merger and Acquisition (M&A) or Initial Public Offering (IPO)*. Startup founders and investors can normally transform their ownership in the startup into money once the company stock becomes publicly traded. This happens when the startup performs an initial public offering or when it is acquired by a public company. The ultimate goal of a startup's investor is to buy a share in a startup in a funding round and sell it for a much higher price after the startup's exit. For this reason, M&A and IPO make good proxies for startup success. However, reaching the IPO stage typically takes more than ten years, and building a machine learning model with such a time horizon is

very challenging. On the other hand, M&A acquisitions often yet not always mean financial success. The investor's profit largely depends on the amount of money for which a startup was acquired, which in most cases is not disclosed to the public. For these reasons, most studies do not aim to predict startup exit with machine learning.

- *Survival:* Some authors argue that survival is the best success measure for early-stage ventures (Soto-Simeone, Sirén, and Antretter, 2020). Survival is often measured by checking whether a startup's website is active. Although survival is a necessary attribute of success, without financial support from VCs and other investors, startups might not be able to test their ideas and scale up aggressively.

- *Fundraising:* Given the importance of access to financial capital for startup development, it is not surprising that many studies choose raising a funding round as a proxy for success. As can be seen in Table 2.1 some authors measure whether a startup manages to achieve a particular fundraising stage, e.g., Series A or Series B.

The startup success prediction studies vary enormously in terms of dataset size used for building the machine learning model - from hundreds to hundred thousand startups. Features used in the studies are mostly limited to tabular data, although some authors attempt to analyze text and the startup's position in the investor-startup graphs. The prediction approaches used to distinguish between successful startups and the failed ones can broadly be divided into two groups:

- Methods for tabular data: Bayesian network, fully-connected neural network, random forest, gradient-boosted trees.

- Graph neural networks that use the startup's position in a graph of business entities to predict its success.

Xiang, Zheng, Wen, Hong, Rose, and Liu (2012) study is perhaps one of the first attempts to dive into the field of using predictive models for assessing the "success" of companies. In that work, the authors explore the prediction of Merger & Acquisition (M&A) as a proxy for startup success. They consider news pertaining to companies and individuals on TechCrunch. The feature

Table 2.1: AI-based Studies on startups' Investment and Valuation.

Author(s)	Domain	Target	Features | Data sources	Dataset size	Method	GOF
Xiang, Zheng, Wen, Hong, Rose, and Liu (2012)	CS	M&A	22 features about startup team, competitors products, offices and financial rounds and 5 topic features extracted from TechCrunch news texts | CrunchBase, TechCrunch	59631	LDA+BN	0.68-0.95 (AUC)
Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke (2018)	CS	Fundraising (Series A)	49 features about startup team, funding rounds, industry + large number of web visibility features, i.e. domains that refer to the startup's website | CrunchBase, LinkedIn, a detailed crawl of the observable web by Yandex	37075	GBT+MLP	0.85 (AUC)
Gastaud, Carniel, and Dalle (2019)	CS	Success: Fundraising, M&A or IPO	20 features about startup team, funding rounds, industry, competitors and investor network | Crunchbase	65957	RF, GNN	0.63 (AUC)
Antretter, Blohm, Grichnik, and Wincent (2019)	BMA	5-year Survival	35 features about startup's activity in Twitter, such as number of likes, tweets, tweet length and emotion | Twitter, unknown angel investment platform	253	GBT	0.82 (F1)
Blohm, Antretter, Sirén, Grichnik, and Wincent (2020)	BMA	Survival	41 features about startup team, industry, social media activity | Twitter, LinkedIn, Google trends, unknown angel investment platform	623	GBT	0.60 (AUC)
Zhang, Zhong, Yuan, and Xiong (2021)	CS	Success: Fundraising (Series A) or M&A	Startup's position in startup-investor-employee network | CrunchBase	6741	GNN	0.71 (AUC)
Żbikowski and Antosiuk (2021)	BMA	Success: Fundraising (Series B), M&A or IPO	9 features about startup team, industry and geographical location | CrunchBase	213171	GBT	0.41 (F1)

Notes. BMA – Business, Management and Accounting, CS – Computer Science, LDA - Latent Dirichlet Allocation, BN - Bayesian network, GBT - gradient-boosted trees, MLP – multilayer perceptron, RF - random forest, GNN - graph neural network.

set they used includes company-specific features, such as managerial and financial features, combined with topic-dependent features that have been extracted via Latent Dirichlet Allocation (LDA) from the text of news. The authors measured performance across startup categories, e.g., biotech, software, and found that for the categories with a sufficiently large number of startups AUC score varied from 0.68 to 0.95. The authors also report that the most predictive feature is the number of revisions on the company CrunchBase profile. It is not entirely clear whether the revisions that happened after the target M&A event are counted or not. In the former case, it is possible that successful companies might have received more revisions *because* of their success: the phenomenon is known in the literature as look-ahead bias.

In a more recent study, Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke (2018) proposed a method named Web-Based Startup Success Prediction (WBSSP). The authors' goal is to predict whether a startup will secure a Series A funding round within a year, given that it has secured Seed or Angel funding during the previous year. In addition to the standard feature set extracted from CrunchBase, which includes information about a startup team, funding rounds, etc., the authors analyze the company's web presence via "a detailed crawl of the observable web used in building the web index of Yandex, a major Russian search engine." Their prediction approach combines logistic regression, a fully-connected neural network, and gradient-boosted trees, and it achieves an AUC metric value of 0.85. In addition, the authors provide importance ranking or the feature groups. Their analysis shows that the most predictive is the feature group, which contains information about previously raised funding rounds and the startup's investors, while the feature group that contains information about the startup team is the least predictive.

Another study in this context is Gastaud, Carniel, and Dalle (2019), where the authors study the factors that contribute to different fundraising series. The authors focus on fundraising as a proxy of startup success, although they also consider a company successful in case of exit, i.e., M&A or IPO. In addition to the standard features that can be directly extracted from CrunchBase, they also find startup's competitors applying word2vec (Mikolov, Sutskever, Chen, Corrado, and Dean, 2013) to the descriptions provided by CrunchBase and measure startup investors' centrality in the startup-investor network. They build random forest and graph neural network models that take into account these features and achieve a 0.63 AUC score. They find that competitors features are important in predicting early-stage fundraising while investor network features are more helpful in predicting the growth stage fundraising.

The idea of applying graph neural networks to predict startup success was further developed by Zhang, Zhong, Yuan, and Xiong (2021). The ability to reach Series A funding or to get acquired is used as a proxy of startup success in this study. The authors argue that the traditional approach for extracting features describing startups requires domain expertise and, instead, they propose to make a prediction based only on the startup's relations with other entities. To this end, they first construct the startup-investor-employee

network, then calculate the number of different metapaths (e.g., startup-employee-startup if a person worked in both startups) between the startups and obtain a parameterized summated adjacent graph consisting of startups only. This graph is then used to train a graph neural network via Maximum A Posterior inference. The authors name the proposed method Scalable Heterogeneous Graph Markov Neural Network. Although this method significantly outperforms baselines proposed in the paper, it achieves an AUC score of 0.71 while Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke (2018) achieve 0.85 on a very similar task. It would be interesting to see in the future what accuracy can be achieved by a model that uses a comprehensive set of features designed by domain experts while taking into account the startup position in the startup-investor-employee graph.

The same year Żbikowski and Antosiuk (2021) approached startup success prediction problem with a focus on the model's practical applicability. To this end, the authors aim to avoid look-ahead bias and intentionally focus on a small set of features that do not change in time, such as a startup's location, industry, and founder's education. As a proxy for a startup's success, they choose Series B fundraising, contrary to the previous studies that considered Series A sufficient to say that a company succeeded. The startups that performed IPO or were acquired are also labeled as successful. The authors compare logistic regression, support vector machine, and gradient boosted trees. The best performing gradient boosted trees model achieves an F1 score of 0.43. In addition, the authors provide feature importance analysis which shows that the three most important features are the country and region where the startup is located and the startup's industry.

Machine learning has also been applied in other studies for investigating phenomena of new venture survival. Antretter, Blohm, Grichnik, and Wincent (2019) is another example where the authors aim to predict the 5-year venture survival using the information about the startup's activity on Twitter. They apply the gradient-boosted trees model and achieve a 0.82 F1 score. Analyzing the variables' importance in the model, the authors find that, surprisingly, the average tweet length makes the best survival predictor among all the features in the analysis.

The startup survival problem was further studied by Blohm, Antretter, Sirén, Grichnik, and Wincent (2020) where the authors combine the features of a startup's team, industry, and social media activity and use gradient-boosted trees to predict new venture survival. To avoid the necessity to choose a particular time window, the authors use Cox Proportional-Hazards Model (Cox, 1972) and achieve an AUC score of 0.60. Interestingly, despite the proposed model's modest performance, it would still outperform angel investors in terms of portfolio returns according. According to the analysis presented in the paper, only the experienced business angels achieve higher returns than the proposed model.

2.2 Startup valuation factors

Startup valuation broadly describes the process of determining the value of a startup company considering key internal factors as well as market forces of the industry to which the company belongs. Startup valuation comes to the forefront when raising capital from investors, getting acquired by another company, or making an initial public offering (IPO). Methods for valuing startups are often applied to early-stage companies that are currently at a pre-revenue stage, which presents a challenge to financial valuation methods when applied to valuating a startup. The commonly used methods in corporate finance (e.g., discounted cash flow, earnings multiple) are based on strict assumptions and require accounting information that new ventures often cannot provide (Damodaran, 2001). When unambiguous measures of performance do not exist or cannot be observed, investors look for other signs of future promise and quality (Baum and Silverman, 2004) and base their estimation on the value of ideas, know-how, and human potential of the team.

The literature on startup investment decision-making and valuation is focused on different factors related to the selection and valuation of startups by venture capital (VC) firms or individual investors (e.g., Block, Fisch, Vismara, and Andres, 2019; Csaszar, Nussbaum, and Sepulveda, 2006; Maxwell, Jeffrey, and Lévesque, 2011; Tumasjan, Braun, and Stolz, 2021; Yin and Luo,

Table 2.2: Exemplary Studies of the Factors of Startups' Investment and Valuation.

Author(s)	Main factors: Financial Capital	Human Capital	Product and technology	Industry and market	Social capital	Online legitimacy	Method	Key findings
Tumasjan et al. (2021)	X				X		4,600 VC financing rounds of US startups; secondary data; regression	Twitter sentiment is positively associated with venture valuation, however, does not correlate with long-term investment success
Dhochak and Doliya (2020)	X		X	X			25 VCs in India and abroad; survey; fuzzy analytic hierarchy process technique	Comparison of the relative importance of criteria (internal, industry, and network-based resources) and subcriteria in explaining the valuation of a new venture
Block et al. (2019)	X	X	X				749 private equity investors; survey; experimental conjoint analysis	Revenue growth is the most important investment criterion, followed by the value-added of product/service, the management team's track record, and profitability
Yin and Luo (2018)	X	X	X	X			1003 startup application profiles to the Singapore-based JFDI accelerator; non-parametric test, regression	Identification of decision criteria of accelerator managers using a real-win-worth framework, and a shift of these criteria across the initial screening of startups and the final selection
Festel et al. (2013)	X	X	X				16 early stage high-tech Swiss and German startups; business plans	The individual adjustment of the beta coefficient applicable to early-stage startups is proposed within a discounted cash flow valuation based on the data in a business plan. The adjustment is based on technological, organizational, financial, and other characteristics
Miloud et al. (2012)	X		X	X			184 rounds of early-stage VC investments in 102 French startups; secondary data; regression	Attractiveness of the industry, the quality of the founder and top management team, and external relationships of a new venture positively affect its valuation by VCs
Maxwell et at. (2011)	X	X	X	X			150 interactions between entrepreneurs and BAs from the Canadian version of a reality TV show "Dragons' Den"; observational interaction technique	BAs use elimination-by-aspects heuristic to reduce the available investment opportunities in their initial decision-making process. After the selection stage, BAs consider different factors (product adoption, product status, protectability, customer engagement, route to market, market potential, relevant experience, financial model).

Continued on next page

Table 2.2. Continued

Author(s)	Financial Capital	Human Capital	Product and technology	Industry and market	Social capital	Online legitimacy	Method	Key findings
Zheng et al. (2010)		X			X		170 US biotechnology startups; secondary data; panel regression	The impact of network status declines, while the impact of innovative capability increases with firm age. Innovative capability and network heterogeneity have complementary effects
Csaszar et al. (2006)	X	X	X	X			Conceptual study; case illustration	A methodology is proposed the combines strategic and cognitive evaluation of startups. The items are related to strategy, team, and finance
Baum and Silverman (2004)		X	X			X	1093 Canadian biotechnology startups, 1991-2000; secondary data; regression	VCs finance startups that have strong technology and relationships, but are in need of management expertise
Mason and Stark (2004)	X	X	X	X			3 bankers, 3 VCs, 4 BAs in UK; verbal protocol analysis	Bankers emphasize the financial aspects of business plans. VCs and business angels stress both market and finance issues. Compared to VCs, business angels give more emphasis to the entrepreneur and 'investor fit' considerations
Knight (1994)	X	X	X	X			100 US, 81 Canadian, 195 European and 53 Asian Pacific investors; survey; non-parametric tests	The criteria were grouped into five categories: the entrepreneur's personality; the entrepreneur's experience; characteristics of the product or service; characteristics of the market; and financial considerations. Country groups ranked the criteria in a similar way with few exceptions
Hall and Hofer (1993)		X	X				16 verbal protocols; semi- structured interviews; verbal protocol analysis	In initial proposal screening, key criteria include fit with the VC's policies and the long-term growth and profitability of the industry
MacMillan et al. (1985)	X	X	X	X			100 US VCs; survey; factor and cluster analyses	The quality of the entrepreneur determines the funding decision. VCs assess ventures in terms of six categories of risk to be managed: risk of losing the entire investment; risk of being unable to bail out if necessary; risk of failure to implement the venture idea; competitive risk; risk of management failure; and risk of leadership failure

2018; Zheng, Liu, and George, 2010). Table 2.2 provides exemplary studies of the factors of startups' investment and valuation.

The different factors are commonly concerned with financial capital, human capital, industry and market, and product and technology, among others. In explaining the startup valuation, scholars built upon various theoretical perspectives such as the resource-based view, industrial organization economics, network theory, signaling theory, or a combination of those. The resource-based perspective focuses on internal resources and capabilities of a venture to understand its value (e.g., Dhochak and Doliya, 2020). Financial capital is considered to be an essential and flexible resource for entrepreneurial firms since a higher amount of capital allows them to experiment with new projects and explore new opportunities, protecting from uncertain outcomes and fostering risk-taking (Cooper, Gimeno-Gascon, and Woo, 1994). Human capital combines characteristics related to knowledge, skills, and experience of founders and the management team, and it is regarded as an important resource and contributor to the performance of a new venture (Macmillan, Siegel, and Narasimha, 1985; Smart, 1999). Investors rely on information about technological and innovative capabilities and the intensity of research and development activities to reveal the quality of a new venture and its ability to generate commercially successful products (Zheng, Liu, and George, 2010). The industry-based perspective emphasizes industry structure and market characteristics in determining the value of a firm (e.g., Miloud, Aspelund, and Cabrol, 2012). Industry characteristics such as size, growth and profitability, environmental threats, and the level of competition are associated with the accessibility of a new venture to the market and market potential for its products (Mason and Stark, 2004). Network theory underlines the role of inter-firm relationships to exchange information and other resources with the firms' environment and impact firm valuation (e.g., Zheng, Liu, and George, 2010). Social capital encompasses the networks and relationships of a firm with external partners, is valuable for knowledge diffusion and transfer, and provides new ventures with a source of competitive advantage (Florin, Lubatkin, and Schulze, 2003). Signaling theory provides an overarching framework focusing on firms' attributes as important signals to investors of the quality and promise of a new venture (e.g., Baum and Silverman, 2004). In recent years, investors started to increasingly look at

social media to access new technologies, trends, and online visibility of the new venture to support their investment decisions (Tumasjan, Braun, and Stolz, 2021).

The various approaches covered by previous studies have provided valuable contributions towards the understanding of factors relevant to startup valuation. However, the existing empirical research in the field has used a restricted set of independent variables, stems from the excessive reliance on surveys and self-reported measures, and "is dominated by regression analysis providing sufficient leeway for future research to use emerging methods" (Köhn, 2018, p.31). Relying on subsets of variables challenges the comparison between the studies and estimations of the variables' relative importance for startup valuation. Although having examined a wide variety of factors, current research has largely overlooked to study the role of human capital variables related to team heterogeneity for startup valuation (Köhn, 2018). Lastly, previous studies have analyzed correlations between variables whereas the empirical estimation of their causal relationships with startup valuation is less clear. This thesis aims to address these research gaps by assessing the factors of startup valuation through prediction and causal discovery.

3 Startup Valuation

In previous chapters, we reviewed the motivation for studying startups operations with various approaches, including data science methods. As mentioned before, startups play a huge role in the modern economy, and they need to be valued in order to raise funding that allows them to grow and develop. We also reviewed both known factors that affect startup value and the previous research on the machine learning methods applied to the startup success prediction problem. In this Chapter, we explore the startup valuation prediction problem and the variables that make good valuation predictors, as well as the variables that affect startup valuation.

3.1 Data Collection

While stocks of public companies are traded daily, and the value of a company can be calculated at any moment, the shares of a startup are rarely sold, and the valuation of a startup is documented only when particular events occur. These events include funding rounds, Merger and Acquisition deals (M&A), and Initial Public Offerings (IPO). In this Chapter, we focus on the valuations obtained during the funding rounds since they are much more frequent than IPO and M&A. Besides, the information about the raised funding amount gives a vital clue about the startup valuation. Based on the literature about startup valuation factors described in Chapter 2 we aim to collect information about startups team, industry, social network activity, and web visibility. To this end, we mine various sources. In the rest of this section, we describe our main repositories for data collection and then illustrate the compatibility between the information taken from different sources.

Similarly, when extracting web visibility variables, we use Google Search API's maximal date parameter to obtain only the results dated before the valuation date. Such a procedure is essential for the ML prediction task since a model that uses information from the future to make predictions cannot be used in practice. The same temporal limitation of variables is suitable for our causal discovery analysis since we are only interested in the determinants of start-up valuation. The analysis of the variables affected by the valuation is outside the scope of this study.

3.1.1 CrunchBase

Crunchbase is a well-established data source for entrepreneurship studies (Ferrati and Muffatto, 2020; Żbikowski and Antosiuk, 2021), which contains a wide range of information about funding rounds, team members, and industries in which a startup operates. For some funding rounds Crunchbase has startup valuation information and in this case we collect it.

The vast majority of studies investigating the field of startups via ML methods leveraged this database as discussed in Section 2.1. In this thesis, the data from Crunchbase plays a critical role as well.

Crunchbase database is populated mainly by community contributors and investment firms that monthly upload their portfolios on Crunchbase in exchange to free access to the database[1]. Crunchbase provides free research access for academia, which allows retrieving information about any startup in the database via API.

When extracting variables about a start-up to predict its valuation, we aim to use only the information available at the valuation date. In Crunchbase, start-ups are characterized both by information that is unlikely to change over time, e.g., industry, and by time-sensitive information such as employees. To avoid using information from the future, we consider only the employees with a start date of employment before the valuation date.

[1] https://support.crunchbase.com/hc/en-us/articles/360009616013-Where-does-Crunchbase-get-their-data- last visited 11/04/2022

3.1.2 Companies House

Companies House[2] is an official United Kingdom government registrar that contains a variety of firms' records as well as information about the people in charge of the company. Companies House database has been previously used in a variety of studies, for example, in the context of small firms accounting and financial management (Collis, 2012; Collis and Jarvis, 2002), or to study gender diversity in the management of UK companies (Martin, Warren-Smith, Scott, and Roper, 2008).

Valuation Extraction from Companies House

In this thesis, the outline of the valuation data collection is the following: we first extract and analyze the information about the funding rounds present in the largest open-access startup database Crunchbase. If a funding round in Crunchbase contains information about valuation, we collect it. In the opposite case, we look for the valuation in documents from Companies House registrar and calculate it via the procedure described below.

In the UK, whenever a company issues shares, it is obliged to file the SH01 form, which contains, among other things, the following information: the number of shares allotted, the amount paid on each share, and the total number of shares of the company. If the document corresponds to a funding round, the amount of money raised by the startup can be calculated as the number of shares allotted times the amount paid on each share; the valuation of the startup is the total number of shares times the amount paid on each share, and the investor's equity is equal to the number of shares allotted divided by the total number of shares. For example, in Figure 3.1 we study a document filed by a British AI drug discovery startup, "Exscientia," filed on May 18, 2020. We can see that the company raised £ $855 \times 57295 =$ £ 49 million ($ 60 million), giving the investor 26% equity. The corresponding post-money valuation of the company is £ $855 \times 217695 =$ £ 186 million ($

[2] https://www.gov.uk/government/organisations/companies-house

Shares Allotted (including bonus shares)

Date or period during which shares are allotted	From	To
	18/05/2020	18/05/2020

Class of Shares:	**SERIES C**	Number allotted	**57295**
	PREFERENCE	Nominal value of each share	**0.001**
Currency:	**GBP**	Amount paid:	**855.093466**
		Amount unpaid:	**0**

(a)

Statement of Capital (Totals)

Currency:	**GBP**	Total number of shares:	**217695**
		Total aggregate nominal value:	**217.695**
		Total aggregate amount unpaid:	**0**

(b)

Figure 3.1: Example of the filed SH01 form pages containing the information required to infer the fundraising amount, the valuation of the company and the equity of the investor.
Notes. Retrieved from Companies House.

227 million) and the pre-money valuation is $ 167 million. This funding round was disclosed in the company's blog post[3].

Although Companies House provides convenient API access to the filed documents, the data collection task posed several technical challenges. The first challenge we faced was finding the Companies House ID of a startup given its entry in Crunchbase. To tackle this task, we first used search by the name Companies House API call and then chose one of the three criteria to choose the correct company entry:

- The exact match of the "Legal name" field,
- Address match,

[3]https://investors.exscientia.ai/press-releases/press-release-details/2020/exscientia-raises-60-million-in-series-c-financing-round-led-by-novo-holdings last visited 11/04/2022

- Co-occurrence of the same name in Crunchbase "team" field and the Companies House "officers" field.

The second problem was reading the documents since they are stored as PDF files without a text layer. For this task, we used the Tesseract OCR package[4]. The code for the Companies House data collection is available.[5]. It is also worth noticing that if one aims to use the Companies House as the sole source of information about the startup's funding round, an additional challenge would be to separate the documents corresponding to the funding rounds from the other types of shares allotment. For this reason, in this thesis, we analyze only the documents that can be aligned by the date and funding amount with a CrunchBase funding round.

Startup team information

In addition to the valuation data, the UK registrar is a rich data source for the information about people managing the firm, which are referred to as "officers" in the records. Moreover, firms are legally obliged to enter the officers' information into the registrar. Thus, Companies House always contains some information about the startup team, while this information can be missing in Crunchbase.

A record of company officers contains demographic information such as age, gender, and nationality. Additionally, it contains each officer's role in the company, e.g., CEO, engineer, the dates at which the officer joined and left the company, and the list of the companies where the officer was previously listed as an officer. Thus it is possible to extract a reach set of features pertaining to the startup team size, diversity, and experience. It is also possible to avoid look-ahead bias by considering only the officers whose starting date in the company is before the valuation date. When collecting the information about a person's experience, we also consider that a startup officer's appointments that started after the valuation date should be excluded from the analysis.

[4]https://github.com/tesseract-ocr/tesseract last visited 11/04/2022
[5]https://github.com/garkavem/Company-House-SH01-Parsing

3.1.3 Twitter API

Information about a venture's Twitter activity is among the most commonly used features for startup success prediction with machine learning. Following this trend, we extract the information about ventures' Twitter activity using Twitter nicknames given on Crunchbase and Twitter API. Twitter API allows one to get the information about the last 2300 tweets of a user along with the tweet text, users mentioned in the tweet, the number of likes and retweets received and the publishing date. The later allows to filter out tweets published after the valuation date avoiding look ahead bias.

3.1.4 Google Search API

The importance of startup web visibility for success prediction was discussed in Chapter 2. In our work, we measure startups' visibility via Google Search API[6]. This instrument retrieves the top ten Google search results for a given query. In our study, we search for startups' names and derive features based on the domains to which the search results belong. In particular, we check whether the startup's own website is found, as well as whether major innovative business news outlets websites such as TechCrunch are present in the search results. To avoid look-ahead bias, we filter the search results by date so as to exclude the web pages that appeared online only after the valuation date.

3.2 Predicting European startups valuation in a funding rounds via domain adaptation.

Sometimes, funding round announcements include not only the amount of money received by the startup, but also the valuation of the startup. Reading such news, one might be wondering how exactly the entrepreneurs and the

[6]https://developers.google.com/custom-search/v1/introduction last visited 11/04/2022

VCs come to an agreement about the startup valuation, *i.e.*, how much equity the VC firms get for a certain funding amount.

In this section, we approach startup valuation from a machine learning perspective focusing on European startups due to the data availability. As discussed in Section 3.1, it is possible to learn the undisclosed to public UK startup valuation, and our experiments provided later in this section suggest that the model trained on UK startups can be effectively used for European startups.

Our goal is to infer the *the undisclosed valuation of a startup corresponding to a funding round with an announced funding amount*. To do that, we leverage both a large-scale Crunchbase dataset and the Great Britain government registrar, Companies House, which were discussed in previous sections. Our choice to study only European startups is based on the data availability and the possibility of knowledge transfer between countries, which will be discussed in detail in Sections 3.2.1 and 3.2.2. We then solve our problem in a Domain Adaptation setting by building a machine learning model which takes into account the discrepancy between the dataset on which the training is performed and the dataset for which we aim to make predictions. Overall, our approach outperforms previously proposed methods by a large margin.

Contribution: contribution of this section is thus two-fold: *(i)* we study a novel problem of great practical importance, namely the prediction of startup valuation, *(ii)* we show that the labeled and unlabeled objects are not aligned and, accordingly, propose to employ a Domain Adaptation setting to train different predictive models.

3.2.1 Data analysis and problem formulation

`Source` and `Target` data: Crunchbase

We adopt the following strategy to collect data from Crunchbase: First, we extract information about the funding rounds present in the Crunchbase snapshot on July 1, 2020, and then collect the corresponding startups' information. Since we are mostly interested in the traditional venture capital

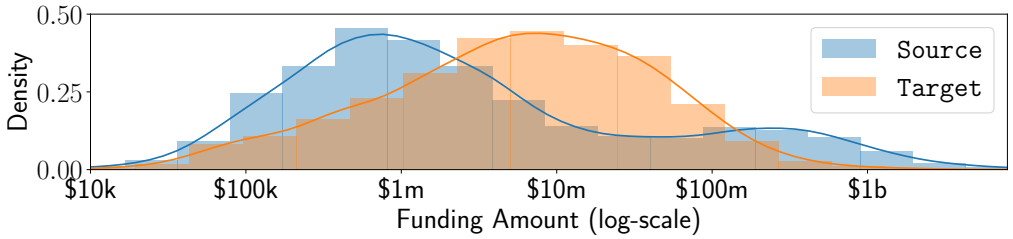


Figure 3.2: Comparison of funding amounts between `Source` and `Target`.

deals for the startups that have not yet gone public, we only collect the following funding rounds: Angel, Pre-Seed, Seed, Series {A, B, C, D, E, G, F, H, I}, Venture, Corporate Round, Private Equity, Undisclosed and Convertible note. Additional information on the startup funding types can be found in Crunchbase Glossary of Funding Types [7]. Such procedure leaves us with:

- 11994 funding rounds with known corresponding startup valuation, which will be referred to as `Source` and
- 185943 funding rounds for which the corresponding startup valuation is not disclosed, which will be referred to as `Target` and for which we aim at predicting the valuation.

Distribution shift Initial comparison of the funding amount distributions of the `Source` and the `Target` can be seen in Fig. 3.2. In the case of announced valuations, i.e. `Source`, the distribution is bimodal with the first mode corresponding roughly to $600K raised and the second mode at $250M. Simultaneously, the funding round sizes with undisclosed valuation, i.e. `Target`, have a single mode at $10M. Our goal is to predict the startup valuation on the `Target`, which is for now entirely unlabeled, *i.e.* the valuations are unknown for this set. Given the shift shown in Fig. 3.2, one needs at least a small portion of `Target` to be annotated. This annotated data then can be used for evaluating the trained models, or even partially for the training purposes, as we will see in Section 3.2.2. Nevertheless, annotating this kind of data is very difficult. As explained in Chapter 1, determining the valuation of startups requires a wide range of domain expertise. What is even more

[7] https://support.crunchbase.com/hc/en-us/articles/115010458467-Glossary-of-Funding-Types last visited 13/04/2022

important is that different investors use different processes to perform the valuation, leading to different valuation numbers for the same startup.

To alleviate this issue, we exploit here the Companies House data which, to the best of our knowledge, has not previously been exploited in the startup research literature. In the following section, we briefly describe how the data is collected from Companies House and then illustrate that this data can indeed be used as an additional source of data for the current study.

$\texttt{Target}_{LAB}$ data: Companies House

As discussed in Section 3.1, in UK the law obliges companies to file certain documents that allow one to calculate the valuation of a startup in a funding round. Thus, for startups in the $\texttt{Target}$ which are present in Companies House, one can readily obtain annotations. In the remainder, the annotated part of the $\texttt{Target}$ set will be denoted as $\texttt{Target}_{LAB}$ (LAB stands for labeled).

To make sure that $\texttt{Target}_{LAB}$ can be used safely in our study (be it in the training or testing part of the model), one needs to check if $\texttt{Target}_{LAB}$ has the same characteristics as $\texttt{Target}$ and, as a result, can be used as a proper evaluation (or further training) data. This point is investigated below.

Geographical transfer To measure the difference between two distributions we use D statistic of Two-Sample Kolmogorov-Smirnov test defined as:

$$D_{a,b} = \sup_x |F_a(x) - F_b(x)| \tag{3.1}$$

where F_a and F_b are the empirical distribution functions of the first and the second sample respectively. Our preliminary studies show that the UK funding rounds amounts differ from those of China or the USA, $D_{\text{UK,USA}} = 0.27$ and $D_{\text{UK,CHN}} = 0.73$. A reasonable suggestion might be that the investment context in the UK and other countries of the region, namely Europe, might be similar. The following countries are included in Europe countries list:

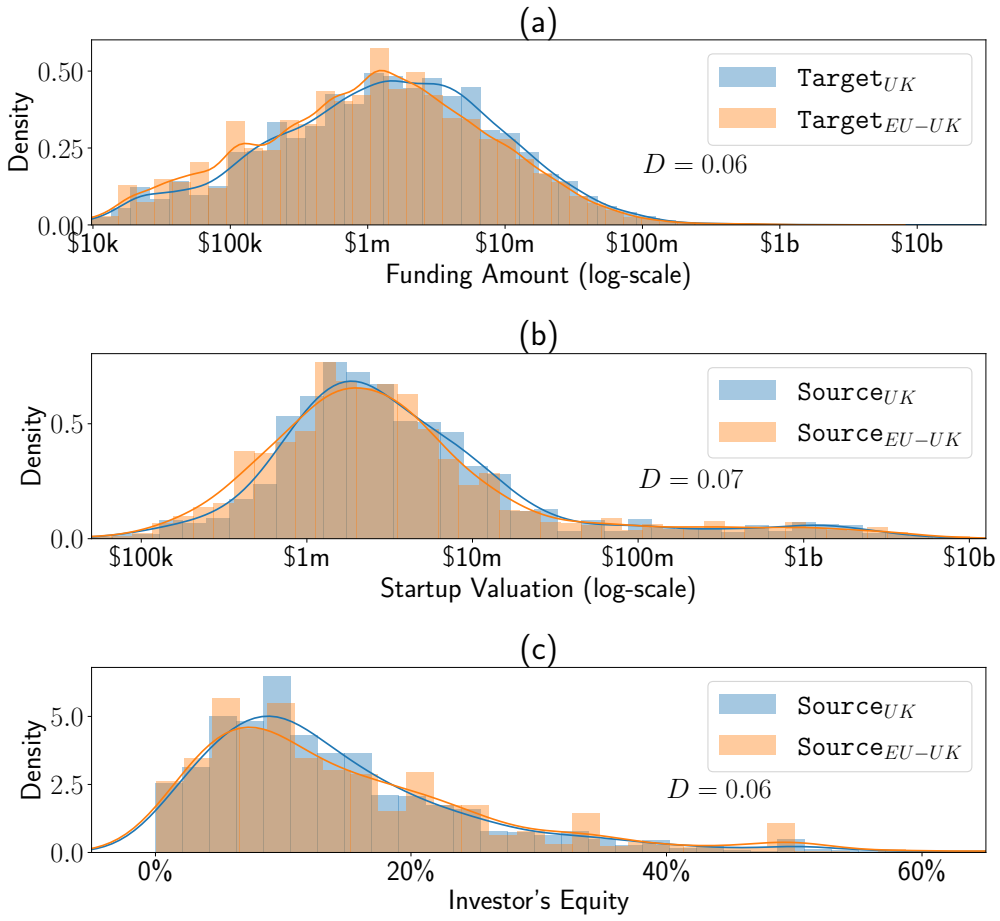


Figure 3.3: EU vs. the UK data: (a) funding amounts, (b) valuations, and (c) investors' equities in the funding rounds with announced valuation. The D statistics of the Kolmogorov-Smirnov test is provided in each case.

Andorra, Albania, Austria, Åland Islands, Bosnia and Herzegovina, Belgium, Bulgaria, Belarus, Switzerland, Cyprus, Czech Republic, Germany, Denmark, Estonia, Spain, Finland, Faroe Islands, France, United Kingdom, Guernsey, Greece, Croatia, Hungary, Ireland, Isle of Man, Iceland, Italy, Jersey, Liechtenstein, Lithuania, Luxembourg, Latvia, Monaco, Moldova, Macedonia, Malta, Netherlands, Norway, Poland, Portugal, Romania, Russian Federation, Sweden, Slovenia, Svalbard and Jan Mayen, Slovakia, San Marino, Ukraine, Vatican City.

To illustrate this point, we compared three different axes: We first investigated the funding amount distribution difference between the UK startups of `Target`, denoted as $\texttt{Target}_{UK}$, and all other European startups of `Target`,

denoted as $\texttt{Target}_{EU-UK}$. This comparison can be seen in Fig. 3.3 (a). We also compared, in Fig. 3.3 (b), the valuation of UK startups from the Source, denoted as $\texttt{Source}_{UK}$, with those of all other European countries, denoted as $\texttt{Source}_{EU-UK}$. Finally, Fig. 3.3 (c) illustrates the investor's equity for the same data. As one can note, on the three (sub-)figures, the distributions are very similar. This is confirmed by the D statistics of the Kolmogorov-Smirnov test which amounts to at most 0.07. In contrast, it amounts to 0.2 when comparing Source with $\texttt{Source}_{UK}$. These findings lead us to consider that one can treat UK based startups and European startups as similar in terms of funding and valuation. In other words, $\texttt{Target}_{LAB}$ shares the same characteristics as $\texttt{Target}_{EU}$ and, accordingly, can be used in the European startup valuation prediction task. The fact that these two sets are similar in terms of funding amount is crucial to design a valuation model, as we will see in Section 3.2.3.

We compare in Fig. 3.4 (a) the properties of funding amounts of $\texttt{Source}_{EU}$, $\texttt{Target}_{EU}$ and $\texttt{Target}_{LAB}$. This plot shows that $\texttt{Target}_{EU}$ and $\texttt{Target}_{LAB}$ are quite similar to each other ($D = 0.08$), and both are different from $\texttt{Source}_{EU}$ ($D = 0.32$). Such similarity supports our hypothesis that $\texttt{Target}_{LAB}$ is much closer to $\texttt{Target}_{EU}$ than $\texttt{Source}_{EU}$ and, thus, a machine learning model's performance on $\texttt{Target}_{EU}$ is better approximated by the model's performance on $\texttt{Target}_{LAB}$ than on hold-out $\texttt{Source}_{EU}$.

Additionally, in Fig. 3.4 (b) and (c) we illustrate the comparison of $\texttt{Source}_{EU}$ and $\texttt{Target}_{LAB}$ in terms of valuation and investor's equity. The properties of $\texttt{Target}_{LAB}$ in terms of startup valuation and investor's equity allow us to get some insight into the differences between the funding rounds with announced and unannounced valuations. An interesting observation is that the differences in funding amounts and investor's equity distributions partially compensate each other, and thus the difference in startup valuation distribution is slightly less prominent. This is not really surprising as startups want to be seen as successful and valuable. Thus, when they raise a relatively small amount of money for an unusually small investor's equity, they are more motivated to report its valuation in addition to the funding amount.

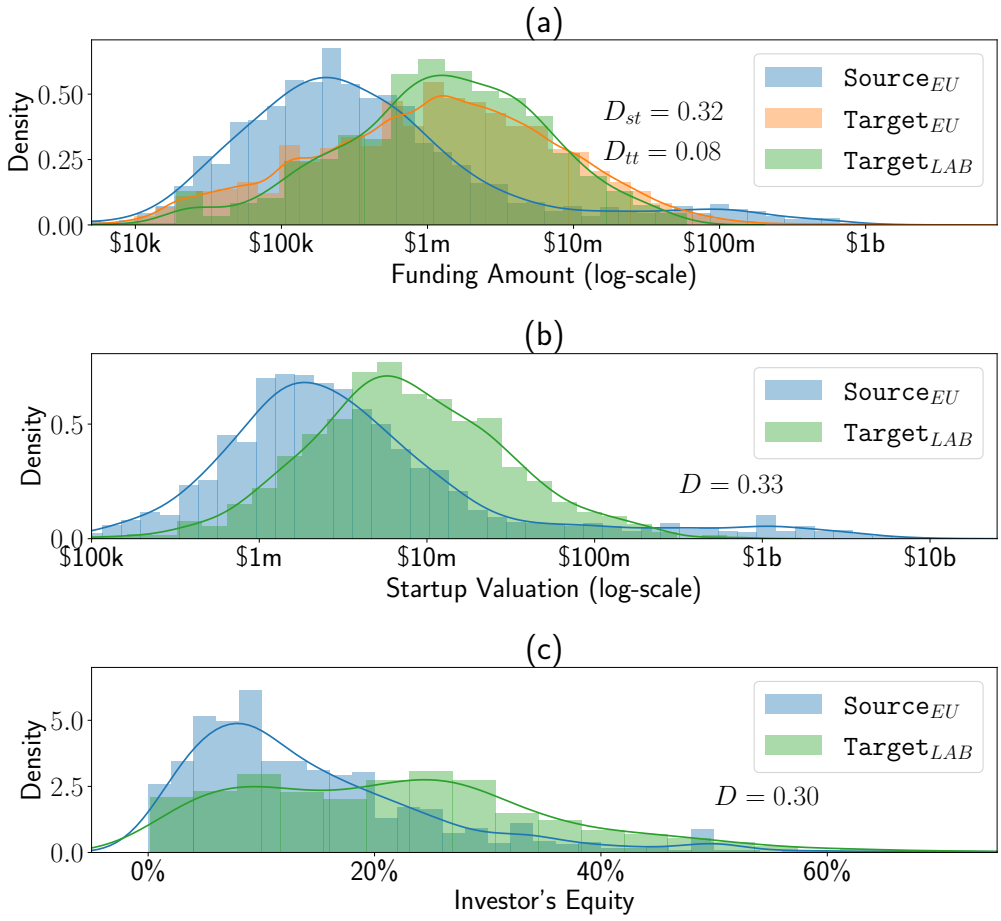


Figure 3.4: Comparison of European funding rounds with announced valuation ($\texttt{Source}_{EU}$), unannounced valuation ($\texttt{Target}_{EU}$) and the set of funding rounds for which the valuation was extracted from Companies House ($\texttt{Target}_{LAB}$). (a) funding amounts (b) startups' valuations (c) obtained investors' equities. The D statistics of the Kolmogorov-Smirnov test is provided in each case.

Summary of dataset

Table 3.1 summarizes the different sets used in our analysis. It is worth noticing that, according to what has been shown previously, there is no particular reason to restrict the training set to $\texttt{Source}_{EU} \in \texttt{Source}$. In our experiments, we report the results on $\texttt{Target}_{LAB}$. The training sets for different models include samples from Source, Target and $\texttt{Target}_{LAB}$. We will give more details on the training approach and experiments in Sections 3.2.2 and 3.2.3.

Table 3.1: Summary of the data. CB: Crunchbase, CH: Companies House.

Zone	Valuation announced in CB	Valuation undisclosed in CB	Valuation undisclosed in CB Computed from CH
World	11994 (`Source`)	185943 (`Target`)	
Europe	3177 ($\texttt{Source}_{EU}$)	34622 ($\texttt{Target}_{EU}$)	
UK	1438	12047	969 ($\texttt{Target}_{LAB}$)

3.2.2 Approach

Domain Adaptation

As explained in Section 3.2.1 and illustrated in Fig. 3.2, there is a significant shift between the `Source` and the `Target` distributions. The described problem typically corresponds to a Domain Adaptation (DA) setting. The core of the DA field is to deal with such scenarios where the source and target data come from different distributions. Suppose we are solving a regression task where X is the input space and $Y \in \mathbb{R}$ is the continuous label variable. We have two different distributions over $X \times Y$ called the source domain $D_s = p_s(x, y)$ and target domain $D_t = p_t(x, y)$. We are provided with labeled source sample S from D_s of size n and with either an unlabeled target sample T from D_t^X of size N, labeled target sample T^L from D_t of size m or both.

$$S = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^{n} \sim D_s \tag{3.2}$$

$$T = \{\boldsymbol{x_i}\}_{i=n}^{N+n} \sim D_t^x \tag{3.3}$$

$$T^L = \{(\boldsymbol{x_i}, y_i)\}_{i=n+N}^{n+N+m} \sim D_t \tag{3.4}$$

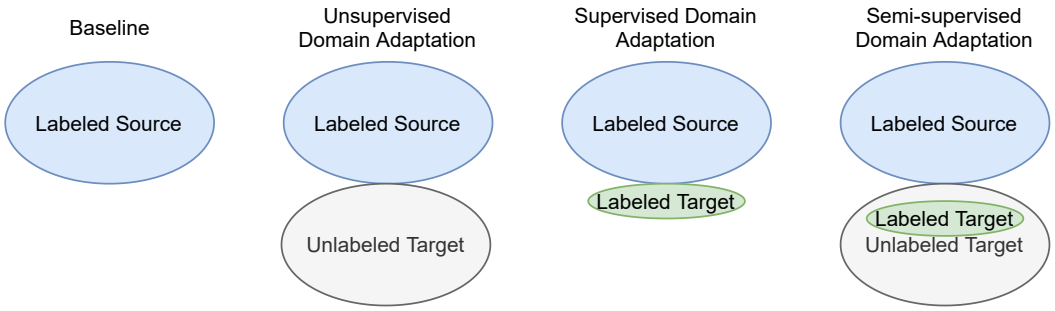


Figure 3.5: Unsupervised, Semi-Supervised and Supervised Domain Adaptation.

our goal is to build a model h with low target risk:

$$R_t(h) = \int_{y \in Y} \int_{x \in X} l(h(x), y) p_t(x, y) dx dy \tag{3.5}$$

In the literature, there are mainly three types of DA approaches illustrated in Figure 3.5: unsupervised, semi-supervised, and supervised. Unsupervised Domain Adaptation (UDA) refers to a setting in which the model is trained on the labeled data from source domain S and unlabeled data from target domain T. The setting in which a portion of the target data is annotated and the learning is performed using labeled source data S and both labeled T^L and unlabeled target data T is known as Semi-Supervised Domain Adaptation (SSDA). Finally, the Supervised Domain Adaptation (SDA) corresponds to the scenario in which both source S and target data T^L are labeled and they are both used in the training phase.

Another axis among which the domain adaptation algorithms can be differentiated concerns the assumptions made about the nature of domain shift (Moreno-Torres, Raeder, Alaiz-Rodríguez, Chawla, and Herrera, 2012). The three commonly recognized types of domain shift are the following:

- Covariate shift: $p_s(y|x) = p_t(y|x)$, $p_s(x) \neq p_t(x)$ when $X \rightarrow Y$
- Prior shift: $p_s(x|y) = p_t(x|y)$, $p_s(y) \neq p_t(y)$ when $Y \rightarrow X$
- Concept shift: $p_s(y|x) \neq p_t(y|x)$, $p_s(x) \neq p_t(x)$ when $X \rightarrow Y$ or $p_s(x|y) \neq p_t(x|y)$, $p_s(y) \neq p_t(y)$ when $Y \rightarrow X$

The latter scenario is not well studied in the literature because it is theoretically hard and not very common in practice. The difference between the first two, namely covariate shift and prior shift, comes down to the causal relations between input feature vector X and output variable y. Covariate shift characterizes the case when X affects y while prior shift describes the situation when output variable y affects X. It is worth noticing that in our task of startup valuation prediction, we can safely assume that X affects y because we collect only the information available before the valuation date as features to make our model applicable for the new funding rounds.

Unsupervised Domain Adaptation

Unsupervised domain adaptation is a long-standing topic in the literature. Some of the early methods for of unsupervised domain adaptation include data importance-weighting (Shimodaira, 2000; Sugiyama and Müller, 2005). The core idea of this approach is to relate source disctribution to the target risk (Eq. 3.5) in the following way:

$$\begin{aligned} R_t(h) &= \int_{y \in Y} \int_{x \in X} l(h(x), y) p_t(x, y) dx dy \\ &= \int_{y \in Y} \int_{x \in X} l(h(x), y) \frac{p_t(x, y)}{p_s(x, y)} p_s(x, y) dx dy \end{aligned} \tag{3.6}$$

Then under the covariate shift assumption we can rewrite $p_t(x, y) = p(y|x)p(x)$ and:

$$R_t(h) = \int_{y \in Y} \int_{x \in X} l(h(x), y) \frac{\cancel{p_t(y|x)} p_t(x)}{\cancel{p_s(y|x)} p_s(x)} p_s(x, y) dx dy \tag{3.7}$$

the goal is then to estimate sample weight $w(x) = \frac{p_t(x)}{p_s(x)}$. A large weight means that sample's probability is much higher in the target distribution than in the source distribution and thus more relevant to the task of building a model that works well on the target distribution. To estimate $w(x)$ various procedures were proposed, for example solving the task assuming that x_t and x_s have Gaussian distribution (Shimodaira, 2000) or via Kullback-Leibler Importance Estimation Procedure(KLIEP) (Sugiyama and Müller, 2005).

Another important family of unsupervised domain adaptation methods is based on the assumption that there exists a transformation that maps source data into target data. It has been theoretically shown that such procedure would allow to obtain a tighter bound on the generalization error (Ganin, Ustinova, Ajakan, Germain, Larochelle, Laviolette, Marchand, and Lempitsky, 2016). Some early works were tackling this problem under the assumption that the mapping transformation is linear. For example technique called Subspace Alignment (Fernando, Habrard, Sebban, and Tuytelaars, 2013)

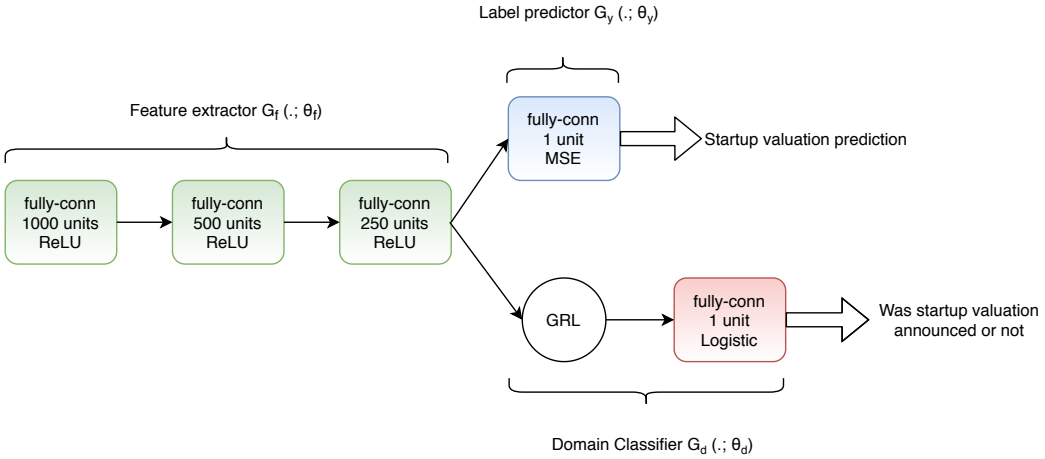


Figure 3.6: DANN architecture for the task of startup valuation prediction. Label is startup valuation. Source domain S is the distribution of the funding rounds for which the startup valuation was announced and target domain T is the distribution of the funding rounds for which the startup valuation was not announced (`Source` and `Target` sets respectively, introduced in Section 3.2.1)

computes the first d principal components in source and target domains, C_s and C_t respectively and a linear transformation matrix $M = C_s^\top C_t$ then aligns source components to target components. Then a model is trained on the transformed source data. More recent techniques from this family include deep domain adaptation method called Domain-Adversarial training of Neural Networks (Ganin, Ustinova, Ajakan, Germain, Larochelle, Laviolette, Marchand, and Lempitsky, 2016). In our thesis we choose this technique for unsupervised domain adaptation setting because it has shown outstanding results on different datasets. An overview of the model is given in Figure 3.6. This method learns a representation that is informative for the main learning task on the source domain and is invariant with respect to the shift between the domains. To this end, the domain classifier is trained to discriminate between the domains. However, a Gradient Reversal Layer incorporated into it passes the signal without a change on the forward pass but reverses the gradients on the backward pass. Thus, the feature extractor parameters are updated in the direction opposite to the one desirable for the domain discrimination task.

More formally, the gradient reversal layer can be treated as a "pseudo-function" $\mathcal{R}(x)$ that passes the signal without a change on the forward pass

but reverses the gradients on the backward pass: $\mathcal{R}(x) = x$, $\frac{d\mathcal{R}(x)}{dx} = -I$, where I is identity matrix. Training DANN then consists in optimising:

$$E(\theta_f, \theta_y, \theta_d) = \frac{1}{n}\sum_{i=1}^{n} L_y^i(\theta_f, \theta_y) - \lambda\left(\frac{1}{n}\sum_{i=1}^{n} L_d^i(\theta_f, \theta_d) + \frac{1}{N}\sum_{i=n}^{n+N} L_d^i(\theta_f, \theta_d)\right) \tag{3.8}$$

where θ_f, θ_y, θ_d are the parameters of the feature extractor, label predictor and domain classifier respectively. L_y and L_d are the main learning task means squared error loss and the domain classifier binary cross entropy loss respectively and λ is the adaptation factor that gradually changes from 0 to 1:

$$L_y^i(\theta_f, \theta_y) = \left(G_y(G_f(x_i; \theta_f)\theta_y) - y_i\right)^2 \tag{3.9}$$

$$L_d^i(\theta_f, \theta_d) = -d_i \log G_d(G_f(x_i; \theta_f)\theta_d) + (1 - d_i)\left(1 - \log G_d(G_f(x_i; \theta_f)\theta_d)\right) \tag{3.10}$$

$$\lambda = \frac{2}{1 + \exp(-\gamma p)} - 1 \tag{3.11}$$

where G_f is feature extractor neural network, G_y is label prediction layer, G_d is domain classifier layer, $d_i = 1$ if $x_i \in \texttt{Target}$ and $d_i = 0$ if $x_i \in \texttt{Source}$, and p is learning epoch.

Supervised Domain Adaptation

Supervised Domain Adaptation is a setting in which the labeled examples from the source domain are used along with only the labeled examples from the target domain. Usually, the number of labeled examples from source is much larger than the number of labeled examples from target. That is true in our case as well since $n = |\texttt{Source}| \gg |\texttt{Target}_{LAB(train)}| = m$ (11994 vs. 242 examples).

The most popular approaches to supervised domain adaptation including cost-sensitive learning (Elkan, 2001), synthetic minority over-sampling technique(SMOTE) (Chawla, Bowyer, Hall, and Kegelmeyer, 2002) and class imbalance learning(Jacobusse and Veenman, 2016) stem from the idea of class importance-weighting very similar to the data importance-weighting framework described in the previous section. Consider target empirical risk Eq. 3.5 under the prior shift assumption:

$$R_t(h) = \int_{y \in Y} \int_{x \in X} l(h(x), y) \frac{\cancel{p_t(x|y)} p_t(y)}{\cancel{p_s(x|y)} p_s(y)} p_s(x, y) dx dy \tag{3.12}$$

the goal is to estimate the weight $w(y) = p_t(y)/p_s(y)$ which will be used to correct the change in label prior. The problem with applying this type of approach for our task is two-fold: first, prior shift assumption seems highly unrealistic in our case, because data generation process $Y \rightarrow X$ would mean that startup valuation defines startup properties, which were collected before the valuation date. The second, practical concern is that class importance-weighting methods are in most cases designed specifically for classification task, and adapting them for the regression task is not straightforward.

For this reason we choose the most straightforward approach for supervised domain adaptation setting which is to train a supervised machine learning model θ on the concatenation of source data and the labeled part of target domain data optimizing the following objective:

$$E(\theta) = \frac{1}{n} \sum_{i=1}^{n} L_y^i(\theta) + \alpha \frac{1}{m} \sum_{i=n+N}^{n+N+m} L_y^i(\theta) \tag{3.13}$$

where target sample weight $\alpha \geq 1$ aims to partially alleviate the problem of huge difference in data sample sizes. The advantage of such an approach is that it can be applied to any base learning model. It has also been shown that even in the presence of abundant unlabeled target domain data and a tiny amount of labeled target data, UDA methods sometimes cannot outperform this simple approach (Saito, Kim, Sclaroff, Darrell, and Saenko, 2019). For this reason, we rely on several supervised machine learning models which will be described in Section 3.2.3.

Semi-Supervised Domain Adaptation

Semi-supervised Domain Adaptation remains a topic slightly less covered in the literature than UDA. Among the recent methods, one could highlight the minmax entropy method proposed by (Saito, Kim, Sclaroff, Darrell, and Saenko, 2019) or the domain adaptive adversarial perturbation scheme from (Kim and Kim, 2020). Despite these methods' impressive performance on various benchmarks, adapting them to the regression problem is not straightforward as they rely on class prototypes. Overall, our literature study did not lead to any SSDA method easily adaptable for our task, and we have directly adapted the DANN algorithm for this setting. This adaptation considers, at every iteration, two mini-batches, one consisting of unlabeled target examples and the other of labeled examples, half of which randomly selected from `Source` and the other half from $\texttt{Target}_{LAB(train)}$. Such an adaptation is quite standard, as described in (Saito, Kim, Sclaroff, Darrell, and Saenko, 2019), and allows one to bias the model learned towards the target domain.

In the case of Semi-Supervised Domain Adaptation the DANN objective function given in Eq. 3.8 is modified as follows:

$$E(\theta_f, \theta_y, \theta_d) = \frac{1}{n}\sum_{i=1}^{n} L_y^i(\theta_f, \theta_y) + \alpha\frac{1}{m}\sum_{i=n+N}^{n+N+m} L_y^i(\theta_f, \theta_y) - \lambda\left(\frac{1}{n}\sum_{i=1}^{n} L_d^i(\theta_f, \theta_d) + \frac{1}{N+m}\sum_{i=n}^{n+N+m} L_d^i(\theta_f, \theta_d)\right) \quad (3.14)$$

Features

Our choice for the features used for the task of startup valuation prediction was based on previous studies (Miloud, Aspelund, and Cabrol, 2012; Zhang, Ye, Essaidi, Agarwal, Liu, and Loo, 2017; Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke, 2018) as well as on the available data. Table 3.2 provides an overview of the features we finally retained, categorised into

Table 3.2: Startup features used in this study.

Group name	Features	Source
General	Country, age of the startup, number of founders, number of current team members, number of past team members, number of founders with previous experience as founder or top-manager at other companies, number of news talking about the startup	Crunchbase
Funding Round	The amount raised in the funding round corresponding to the target valuation, series of the funding round corresponding to the target valuation	Crunchbase
Financial History	Number of previously secured funding rounds, previous funding amount, time since the previous funding round, mean of funding amount raised during the previous funding rounds, max of funding amount raised during the previous funding rounds, funding amount at each series: Seed, Series A, etc.	Crunchbase
Social Networks	Number of tweets, mean/max number of likes of tweet, mean/max number of retweets of tweet, number of different users to which startup replied, number of different hashtags used by the startup	Twitter API

four main groups: General, Funding Round, Financial History and Social Networks.

The *General* group presents generic features such as age of startup, country of origin, number of founders and employees. The *Funding Round* group merely includes the series and the amount raised during the funding round for which we aim at predicting the valuation. The *Financial History* group includes statistics about the previous funding rounds. The *Social network* features, extracted from Twitter, represent the "importance" of startups on social media. Since many entrepreneurs dedicate a considerable amount of time on online networks in order to reach potential customers, partners or investors, we hypothesise that some characteristics of the startup's activity on social media might be correlated to its maturity and possibly valuation. Although it would be interesting to use other information from other social networks, in this study, we narrow down our monitoring to startups' activities on Twitter since its API is readily available to researchers, contrary to other platforms such as Linkedin or Facebook.

3.2.3 Experiments

In this section, we present our experimental results performed on the approaches explained in the previous section as well as some other baselines. We then provide some insight into the contributions of the different features.

Baselines

The following is the list of baselines that we use in order to illustrate the adaptability of the DA setting to the problem under consideration. Note that to train these baselines, we only use the `Source` data, i.e., we consider the problem as a classical regression problem.

- **EPoSV** (An Empirical Perspective on Startup Valuations(Quintero, 2019)): to the best of our knowledge, it is the only data-driven approach for startup valuation prediction. It consists in finding the best coefficient binding logarithm of the funding amount and the logarithm of startup valuation for each fundraising series. It is worth noticing that if we follow the procedure proposed by the authors, i.e., build and test the model on the startup valuations available at Crunchbase `Source` we achieve a R^2 score of 0.876 which is close to the one reported in the Quintero (2019). However, in practice, it would be of more interest to predict the undisclosed valuations, and unfortunately, facing this domain shift, the model's performance degrades.

- **CatBoost**: CatBoost (Prokhorenkova, Gusev, Vorobev, Dorogush, and Gulin, 2018) is a popular gradient boosting library. We choose gradient boosting for two reasons: *(i)* it achieves state-of-the-art results on many practical tasks (Caruana and Niculescu-Mizil, 2006; Roe, Yang, Zhu, Liu, Stancu, and McGregor, 2005; Zhang and Haghani, 2015), and *(ii)* this particular implementation has been shown to work well in the startup fundraising prediction taskSharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke, 2018. Although in (Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke, 2018) the authors use CatBoost as the principal component of a task-specific framework that combines several different models, applying a stand-alone CatBoost model to our data also seems appropriate.

- **MLP**: we also use a classical multilayer perceptron with three fully connected hidden layers of 1000, 500 and 250 neurons, ReLU (Nair and Hinton, 2010) nonlinearities followed by a batch normalization layer(Ioffe and Szegedy, 2015).

Experimental setup and metrics

We apply $\log_{10}$ *transformation to target values* so as to have them in a reasonable range. For evaluation, we make use of the coefficient of determination R^2 and the root mean squared error (RMSE):

$$R^2 = 1 - \frac{\sum_i (y_i - h(x_i))^2}{\sum_i (\bar{y} - y_i)^2} \tag{3.15}$$

$$\text{RMSE} = \sum_i (y_i - h(x_i))^2 \tag{3.16}$$

where h is the trained model, and $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ is the mean label value.

Note that, once only a portion of the target domain data is labeled, one can either employ SDA, by ignoring the unlabeled part of the target, or use SSDA by taking into account both the unlabeled and labeled parts of the target. It is also possible to solve the problem in UDA setting using only the unlabeled target data. In order to adapt our data to all these variants, we divide $\texttt{Target}_{LAB}$ into three sets (25%-25%-50% partitions respectively):

- $\texttt{Target}_{LAB(train)}$, with 242 examples, which will be used for the training in SSDA and SDA,
- $\texttt{Target}_{LAB(dev)}$, with 242 examples, which will be used for hyperparameters tuning in SSDA and SDA,
- $\texttt{Target}_{LAB(test)}$, with 485 examples, which will be used to evaluate the models and to report the results on all methods.

All the models in our experiments are tested on $\texttt{Target}_{LAB(test)}$data since we have shown that this dataset approximates that data on which we would like to apply our model in practice. The baselines either only on `Source`data – large but shifted with respect to the `Target`dataset or only on $\texttt{Target}_{LAB(train)}$. For DANN in the unsupervised setting, we do not use $\texttt{Target}_{LAB(train)}$ or $\texttt{Target}_{LAB(dev)}$ for training and parameter tuning, since our goal is to find out what is the best performance that one could achieve

Table 3.3: Experimental results on $\texttt{Target}_{LAB(test)}$set. Bold numbers are used for models statistically significantly better than the other models (Wilcoxon signed-rank test with $p < 0.001$). Baselines are separated with a vertical line. S, T and T_L denote `Source`, `Target` and $\texttt{Target}_{LAB}$respectively.

Train data	**EPoSV** S	**CatBoost** S	**MLP** S	**EPoSV** T_L	**CatBoost** T_L	**DANN** S+T (UDA)	**EPoSV** S+T_L (SDA)	**CatBoost** S+T_L (SDA)	**MLP** S+T_L (SDA)	**DANN** S+T+S_L (SSDA)
$R^2 \Uparrow$	0.617	0.738	0.769	0.759	0.767	0.788	0.720	**0.817**	0.807	0.807
RMSE$\Downarrow$	0.347	0.293	0.275	0.276	0.268	0.263	0.300	**0.245**	0.251	0.250

using only the `Source` (case of baselines) and unlabeled `Target` (case of unsupervised DANN) readily available in Crunchbase.

The essential CatBoost parameters, such as learning rate and the number of estimators, were chosen on cross-validation (CV) on `Source`. In the SDA setting, we use the same learning rate; the number of estimators is chosen based on the $\texttt{Target}_{LAB(dev)}$ metrics. The weights of the $\texttt{Target}_{LAB(train)}$ samples are set to 10 to partially compensate for the differences in `Source` and $\texttt{Target}_{LAB(train)}$ sizes.

The MLP architecture, as well as the training parameters, including the optimizer, learning rate scheduler, batch size, and the number of epochs, were chosen using CV on `Source`. The same parameters were used for DANN method. To reduce the hyperparameters influence, all these parameters (except for the number of epochs) are used in SDA and SSDA settings as well. The number of epochs in SDA and SSDA settings is defined by performance on $\texttt{Target}_{LAB(dev)}$.

To robustly estimate the performance of different methods, we repeat this procedure for 20 random splits of the $\texttt{Target}_{LAB}$ set into test and training/dev parts. For MLP and DANN, we repeat the experiment with five different random seeds used for the initialization of weights for each split.

Results

The results of our experiments are shown in Table 3.3. In each column, we specify if the method uses only `Source` data (the first three columns), only

Target data (columns 4 and 5) or if it is supervised, semi-supervised, or unsupervised domain adaptation (SDA, SSDA, and UDA, respectively). As one can observe, among all approaches, EPoSV performs significantly worse. This observation mainly suggests that using a rich set of features and a more powerful model is required for solving the startup valuation task, which is not the case for EPoSV. We can also see that EPoSV works best when trained on the $\texttt{Target}_{LAB}$ data exclusively. A linear regression model that only fits several linear coefficients cannot benefit from a large but shifted Source data set.

The second observation is that all DA based approaches outperform the baselines, both the ones trained only on Source and the ones trained only on Target. This point illustrates that DA setting is indeed a more appropriate approach for solving such a problem. Among all baselines, one can notice that MLP performs the best. The next observation is that in the absence of target domain information, MLP can generalize better to the target domain data. However, once target domain information is introduced, CatBoost achieves better results than MLP. Such improvement is due to the ability of boosting based methods in dealing with complex input data. The SDA version of CatBoost also achieves the best results even among all other DA based approaches.

Another finding is that even in the absence of labeled data in the target domain, i.e. $\texttt{Target}_{LAB}$, a UDA approach is a better match than the methods not benefiting from DA. Indeed, DANN in the UDA setting performs better than all baselines, which use only Source. The last observation is that DANN in SSDA setting does not improve the results over MLP(SDA). This result is surprising given the significant performance gain that DANN achieves over MLP in the absence of labeled data from target domain. However, a similar outcome, *i.e.* DANN failure in SSDA setting, has been reported previously on different benchmarks (Saito, Kim, Sclaroff, Darrell, and Saenko, 2019).

Feature group contributions

In this section, we aim to get some insight into the contributions of the feature groups described in Table 3.2. We are going to use feature ablation

Table 3.4: Contribution of different feature groups. Bold numbers are used for models statistically significantly better than the other models (Wilcoxon signed-rank test with $p < 0.05$).

	Full data	**⊖General**	**⊖Funding**	**⊖Financial**	**⊖Social Net.**
$R^2 \Uparrow$	**0.817**	0.789	0.499	0.793	0.811
RMSE⇓	**0.245**	0.263	0.405	0.261	0.250

(Bengtson and Roth, 2008) approach to estimate importance of each group. The feature ablation approach consists in comparing the performance of a model trained on the full set of features and the feature set containing all variables except the studied one. Thus, we train our best performing model, i.e. CatBoost(SDA), on different versions of the dataset, each of which containing all the feature groups except for one. Such an approach is known in literature as . The results of this experiment are illustrated in Table 3.4. The first column of the table (Full data) shows the performance of CatBoost(SDA) on the complete set of features.

As one can expect, the most significant impact comes from the Funding group, which makes sense since the valuation prediction that we considered in this study relies mainly on fundraising events. Nevertheless, even in the absence of information about the funding round, *ca.* 50% of the variability of the dependent variable is accounted for. Another observation is that the second most important group of features is the General group, comprising features such as the startup's age and its country of origin. Without this group, the model loses around 3% and 6.5% in terms of R^2 and RMSE respectively. This group is closely followed by the Financial group. As to the Social network group, its impact is relatively modest, though still statistically significant.

3.2.4 Concluding remarks

In this section, we investigated a real-world task of great importance: finding the undisclosed valuation of startups. To do that, we first collected data from Crunchbase and showed that there is a significant distributional shift between the labeled and the unlabeled data. We then used Companies House to partially annotate the unlabeled data and illustrated that these annotations

are compatible with the Crunchbase data distributions. We then proposed to solve this problem in a Domain Adaptation (DA) setting and illustrated that DA based methods perform much better than other baselines. We also provided some insight into the impact of the different feature groups on the model's performance, which shows that, if the funding features are of primary importance to solve the valuation problem, the other groups work hand in hand to provide better valuation predictions.

3.3 Assessing the Determinants of Start-up Valuation through Prediction and Causal Discovery.

In this section, we aim to identify which variables make the best predictors of start-up valuation and discover direct and indirect causal determinants of valuation with respect to the observable variables. Previous research indicates that investors rely on multiple criteria to evaluate early-stage companies and their potential for success (Franke, Gruber, Harhoff, and Henkel, 2008; Tumasjan, Braun, and Stolz, 2021; Zheng, Liu, and George, 2010). They include start-up, founder, and team attributes, financial information, intellectual property and alliances, and market factors, among others characteristics (Köhn, 2018). However, previous studies have focused on a restricted set of criteria, varying from one study to the other, from which it is difficult to draw general conclusions on the importance of each criterion. Moreover, valuation is subject to human biases in the context of cognitive limitations and incomplete and uncertain information processed by investors (Harrison, Mason, and Smith, 2015; Maxwell, Jeffrey, and Lévesque, 2011). As a result, different evaluating agents and methods may yield different value estimations of the same company.

The rise of artificial intelligence (AI) and machine learning (ML) has enabled the processing of large amounts of information to facilitate decision-making processes (Obschonka and Audretsch, 2019). Data science methods have been recently called for in entrepreneurship research (Lévesque, Obschonka,

and Nambisan, 2020; Schwab and Zhang, 2019). Algorithms can quickly process large quantities of unstructured and rapidly changing data from many sources, which is essential given the growing number of ventures in the modern world. They afford opportunities to explore patterns in the data (George, Haas, and Pentland, 2014) and provide insights into which variables, among a wide range of different candidates, are the best predictors of events.

This section addresses two research questions: (1) Which variables are the best start-up valuation predictors that allow distinguishing between more and less valuable start-ups? and (2) Which variables are the observable causal determinants of start-up valuation that influence the value of a start-up? To answer these questions, we have built a comprehensive dataset and applied ML and causal discovery methods to analyze start-up valuation factors. This study adopts a quantitative exploratory data-driven approach (Coad and Srhoj, 2020; Schwab and Zhang, 2019) focused on discovering key predictor variables and causal determinants of valuation of start-ups. In this study, we refer to predictors as variables allowing forecasting start-up valuation and causal determinants as variables influencing start-up valuation. The data-driven investigations are informed by prior research on the investment and valuation of early-stage companies. Specifically, we explore the contributions of factors related to financial capital, human capital, industry and market timing, and online legitimacy as these factors were shown to be important considerations in investment decision-making (Block, Fisch, Vismara, and Andres, 2019; Knight, 1994; Tumasjan, Braun, and Stolz, 2021). We analyze in this study 2366 valuations of start-ups in the United Kingdom collected from Crunchbase, the largest start-up database, and Companies House, the UK's registrar of companies, complemented with Twitter and Google Search data sources. These databases contain rich information on different aspects of start-ups, including non-financial variables pertaining to human capital, industry, and online legitimacy. The start-up's value is determined in this study by a transaction price of shares allotted by a company when a fundraising event happens. Using ML and causal discovery methods, we rank start-up valuation factors by their predictive power and identify direct and indirect causes of valuation.

3.3.1 Variables

Our study applies a quantitative exploratory approach (Schwab and Zhang, 2019) to discover which variables can be used to predict start-up valuation with ML and which variables are causal determinants of valuation. It utilizes a comprehensive set of variables to account for a wide range of factors related to the valuation of new ventures as informed by prior research in the field. The variables are obtained from four main sources: Crunchbase, Companies House, Twitter API, and Google Search API as described in Section 3.1.

The dependent variable start-up *valuation* is extracted from Crunchbase funding round records and Companies House registrar entries. The valuation data have been extracted from the following funding rounds: Angel, Pre-Seed, Seed, Equity Crowdfunding, Venture, Series A-F, Corporate round, and Private equity. Other types of funding rounds, as convertible note and debt financing, are out of the scope of this study. If a funding round in Crunchbase contains information about valuation, we collect it. In the opposite case, we look for the valuation in documents from Companies House registrar.

The predictor variables are divided into four groups.

- *financial capital* group includes the variables of past funding amount and funding rounds as well as the number of past crowdfunding campaigns and the amount of money collected in them. The data are retrieved from Crunchbase.

- *Human capital* variables are related to team size and roles, team experience, and team nationality and diversity. The measures were obtained from Crunchbase and Companies House.

- *industry and market timing* group includes variables such as start-up age, number of start-ups founded in the same industry, and industry costliness. In preliminary experiments, we also extracted additional information in the form of categorical variables. However, adding these variables did not improve the accuracy of the predictive model. For this reason, these variables were not included in our study. The measures were obtained from Crunchbase.

Table 3.5: Variables Used in the Analysis.

Variable	Measurement	Data Source
	Valuation	
Start-up valuation	Pre-money valuation of a company corresponding to funding round in Crunchbase (log). If not available, calculated as the total number of shares multiplied by the amount paid on each share, reported in SH01 form of Companies House, minus the fundraising amount reported in Crunchbase (log).	Crunchbase, Companies House
	Financial Capital	
Past funding	Total amount of obtained funding (log) in previous fundraisings.	Crunchbase
Funding rounds	No. of previously obtained funding rounds.	Crunchbase
Past crowdfunding	Total amount of obtained crowdfunding (log) in previous crowdfunding campaigns.	Crunchbase
Crowdfunding campaigns	No. of previously finished crowdfunding campaigns.	Crunchbase
	Human Capital	
	Team Size and Roles	
Board and advisors	No. of people listed as board members and advisors.	Crunchbase
Current team	No. of people who started to work and did not quit before the valuation date.	Crunchbase
Assigned officers	No. of officers appointed before the valuation date. A sum of active and resigned officers.	Companies House
Occupation (5 variables)	Proportion of assigned officers with a given occupation in a company for each of the following occupations: *CEO, Engineer, Consultant, Investment, Finance*.	Companies House
	Team Experience	
Officer age	Mean age of company's officers.	Companies House
Past appointments	Mean no. of previous (i.e. terminated before the valuation date) appointments of company's officers.	Companies House
Money raised by previous companies	Mean of mean amount of funding (log) obtained by previous companies of officers.	Crunchbase, Companies House
PhD degree	No. of officers with PhD degree (i.e., having "PhD" in name on Companies House).	Companies House
Occupation experience (4 variables)	Proportion of assigned officers with experience in a given occupation in other companies for each of the following occupations: *Director, Consultant, Investor, Manager*.	Companies House
Managerial experience	Proportion of assigned officers with experience in managerial position (such as CEO, CTO, Director etc.) in other companies.	Companies House

Continued on next page

Table 3.5. Continued

Variable	Measurement	Data Source
	Team Nationality and Diversity	
Nationality (5 variables)	Proportion of officers with particular nationality for the following nations: *British, American, German, French, Irish*.	Companies House
Different nations	No. of different assigned officers' nationalities.	Companies House
Foreign officers	Proportion of officers with nationality other than UK.	Companies House
Female officers	Proportion of female assigned officers.	Companies House
Age diversity	Standard Deviation (SD) of company's officers age.	Companies House
Time in start-up diversity	SD of time (in years) active officers have worked in the company at the valuation date.	Companies House
	Industry and Market Timing	
Start-up age	Age in years of the company at the moment of valuation (no. of days between the company foundation and valuation, divided by 365).	Companies House
Start-ups founded in industry last year	Mean no. of companies founded in the start-up's industry(ies) during a calendar year preceding the valuation year.	Crunchbase
Funding raised in industry last year	Total amount of funding (log) obtained by companies in the start-up's industry(ies) during a calendar year preceding the valuation year.	Crunchbase
Industry costliness	Mean of the average amount (log) of funding rounds obtained by companies in the start-up's industry(ies) during a calendar year preceding the valuation year.	Crunchbase
	Online Legitimacy	
	News Coverage	
News	Number of news related to the company.	Crunchbase
	Social Media	
Tweets	No. of tweets posted by the start-up.	Twitter API
Twitter replies	No. of replies posted by the start-up	Twitter API.
Twitter likes	Mean no. of likes of tweets posted by the start-up.	Twitter API
Twitter user mentions	Mean no. of user mentions in tweets posted by the start-up.	Twitter API
Twitter medias	Mean no. of medias in tweets posted by the start-up.	Twitter API
Twitter unique hashtags	No. of unique hashtags used by the start-up.	Twitter API
Twitter users replied	No. of unique users to which the start-up replied on Twitter.	Twitter API
Tweets length	Average length of tweets	Twitter API

Continued on next page

Table 3.5. Continued

Variable	Measurement	Data Source
	Web Visibility	
Relevant search results	No. of relevant, i.e. containing the exact company name, search results among the top 10 results in Google.	Google Search API
Results from own domain	No. of links to the company's own domain among top 10 results in Google.	Google Search API
Results from other domains (11 variables)	No. of links to each of the following domains among top 10 results in Google: *techcrunch.com, tech.eu, eu-startups.com, uktech.news, facebook.com, angel.co, crowdfundinsider.com, linkedin.com, twitter.com, crunchbase.com, medium.com*.	Google Search API

- *Online legitimacy* is measured by the variables related to news coverage, social media, and web visibility, obtained using Twitter API and Google Search API.

Having collected a large set of 205 variables, we first aim to filter out the ones that are unlikely to contribute to solving valuation prediction or causal discovery tasks. To this end we remove variables with low variance, i.e., those with the same value in more than 95% of cases. Second, we remove instances with missing data. Finally, we filter out variables that are not correlated with valuation according to the Spearman test (p-value <0.01). Overall, the final dataset used in the analysis includes 2366 instances and 58 variables, summarized in Table 3. Note, that for the team roles variables and team experience variables we initially extracted proportion of officers with (experience in) a given occupation for 50 most common occupations, but only the occupations and experience variables that remained in the analysis after the initial filtering are shown in the table. The same applies to proportion of officers with certain nationality and search results from certain domain variables.

3.3.2 Machine Learning Model

This study relies on a *gradient boosting* ML model (Friedman, 2001) to predict start-up valuation, using regression trees as weak regression models. This model was shown to outperform other models on benchmark tasks

(Caruana and Niculescu-Mizil, 2006) and was used to address various real-world problems, including credit risk assessments (Chang, Chang, and Wu, 2018), bank failure (Carmona, Climent, and Momparler, 2019; Climent, Momparler, and Carmona, 2019) or, in entrepreneurship research, new venture survival prediction (Antretter, Blohm, Grichnik, and Wincent, 2019; Blohm, Antretter, Sirén, Grichnik, and Wincent, 2020). Given its success, many computationally efficient variants of boosting techniques have been developed in the ML community: XGBoost (Chen and Guestrin, 2016a), LightGBM (Ke, Meng, Finley, Wang, Chen, Ma, Ye, and Liu, 2017), and CatBoost (Prokhorenkova, Gusev, Vorobev, Dorogush, and Gulin, 2018). The particular *CatBoost* implementation that we use in our study has been shown to perform well on a task closely related to ours, namely predicting the fundraising events for start-ups (Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke, 2018).

ML models with a large number of parameters, such as gradient boosting approaches, are able to effectively "memorize" the training dataset. Therefore, it is mandatory to estimate the model's performance on the test set, i.e., part of the data not seen by the model during training.

At the same time the realistic scenario of ML method application is to train a model on historical valuation data and then predict valuations in real-time. Therefore, we use 1855 valuations dated before 01-01-2019 for training and 511 later valuations for testing. The experiments are conducted with CatBoost Python package.

Model Interpretation

In recent years with the rise of ML applications in various domains, the explanation of complex nonlinear models' decisions became essential. An extensive research effort has been recently put into the field of *explainable ML* (Covert, Lundberg, and Lee, 2020a; Mathews, 2019; Molnar, 2019; Roscher, Bohn, Duarte, and Garcke, 2020). The core concept of explainable ML is *feature importance* which is the measure of the predictive power that a feature can provide to the model. Let us consider a supervised learning

task where a model f predicts the label Y given the input X of individual features $(X_1, X_2, ..., X_d)$ minimizing the loss function l.

Correlations and Univariate Predictors. One way to determine whether a particular variable might help predict the target variable is by computing the correlation between the predictor and the target variable. While correlation is closely connected to the performance of a univariate predictor, i.e., a model that contains just one predictor variable, one can also build an ML model on the variables of a particular group. We will refer to such models as *group models*. The performance of such a model gives insight into the aggregated usefulness of the feature group taking into account within-group features interaction. The drawback of this approach is that it does not account for feature interactions that a complex ML model can capture. In this study, we use the Spearman correlation coefficient to assess individual variables, and we calculate the coefficient of determination of the group models.

Features Ablation. The second technique widely used to assess the usefulness of a feature for the prediction is feature ablation approach that we have introduced in Section 3.2.3. The feature ablation approach consists in comparing the performance of a model f trained on the full feature set and the models $f_1, f_2, ..., f_d$ trained on the feature sets that exclude the variables 1,2, ...,d. The importance values $\phi_1, ..., \phi_d$ are then assigned according to the formula:

$$\phi_i = \mathbb{E}\left[l\left(f_i(X_{D\setminus\{i\}}), Y\right)\right] - \mathbb{E}\left[l\left(f(X), Y\right)\right] \tag{3.17}$$

the method can naturally be extended to measure a group S ablation importance measure ϕ_S:

$$\phi_S = \mathbb{E}\left[l\left(f_i(X_{D\setminus S}), Y\right)\right] - \mathbb{E}\left[l\left(f(X), Y\right)\right] \tag{3.18}$$

This technique also does not allow to capture the features interaction. For example, two strongly correlated features with considerable predictive power would receive low importance scores in a feature ablation study.

Permutation Importance. Another commonly used approach in ML studies is the permutation importance (Breiman, 2001). It involves random shuffling of the studied features's values across the dataset and measuring the drop in prediction accuracy on the contaminated dataset compared to the original one:

$$\phi_i = \mathbb{E}\left[l\left(f(X_i'), Y\right)\right] - \mathbb{E}\left[l\left(f(X), Y\right)\right] \tag{3.19}$$

where X_i' is input in which the $i-th$ feature values were randomly perturbated. Recently, a variant of this method for feature groups was proposed in Gregorutti, Michel, and Saint-Pierre (2015). However, this approach does not account for features' correlations and may create unrealistic or even impossible data instances (Hooker and Mentch, 2019). For example, using the permutation test for the start-up age variable in our dataset, we can obtain an instance with start-up age = 1 year and the number of funding rounds=7. The existence of such a start-up in real life is unlikely, and the model's prediction for such an instance is not very meaningful.

Shapley Additive Global importancE: SAGE. Classical Shapley value is a game theory concept that fairly allocates the game gain to each player in a grand coalition (Shapley, 1953). Shapley values have been applied to analyze ML models (Lipovetsky and Conklin, 2001; Štrumbelj and Kononenko, 2014). Lundberg and Lee, 2017 proposed SHAP (SHapley Additive exPlanations) – a technique for ML model explanation based on the Shapley values. The idea behind this method is to consider the prediction as a game and the features as players that can form a coalition. For a cooperative game $w : \wp(D) \to \mathbb{R}$, Shapley value of i-th feature is given by the following formula:

$$\phi_i(w) = \frac{1}{d} \sum_{S \subseteq D \setminus \{i\}} \binom{d-1}{|S|}^{-1} \left(w(S \cup \{i\}) - w(S)\right) \tag{3.20}$$

such value can be used as a feature importance measure satisfying properties that are considered useful for the model explanation task, such as additivity. A computation-efficient method was proposed to approximate the Shapley values for the tree-based models such as random forest and gradient boosting machines (Lundberg, Erion, Chen, DeGrave, Prutkin, Nair, Katz, Himmelfarb, Bansal, and Lee, 2020).

While SHAP values show how much the model's prediction on a particular instance relies on the features' values, Shapley Additive Global importancE (SAGE) (Covert, Lundberg, and Lee, 2020b) is a way to estimate the usefulness of a feature for the model's accuracy on the whole dataset. This approach averages over all possible subsets of features the gain in the accuracy brought by the feature (or feature group) of interest. A computationally efficient approximation technique was proposed by Covert, Lundberg, and Lee (2020b). This method has been shown to satisfy the following useful properties:

- SAGE values sum to the input's total predictive power $\sum_{i=1}^{d} \phi(v_f) = v_f(D)$,
- Features with deterministic relationship always have equal importance,
- $\phi_i(v_f) = 0$ if X_i is conditionally independent of $f(X)$ given all possible subsets of features X_S.

In practice, SAGE has been shown to outperform other feature importance attribution methods. For these reasons, in our experiments we use SAGE to estimate the usefulness of the feature groups and individual features for the start-up valuation prediction, while comparing it with the other approaches.

3.3.3 Causal Discovery

The ideal way to discover causal relations is to use interventions or randomized controlled trials. However, these approaches are in many cases costly, time-consuming, unethical, or just impossible to deploy in practice, especially in social sciences. Therefore, the traditional pipeline of discovering causal relationships between variables in social sciences is to build a theory about the possible causal relation between several variables and then validate it empirically using observational data (Hofman, Watts, Athey, Garip, Griffiths, Kleinberg, Margetts, Mullainathan, Salganik, Vazire, Vespignani, and Yarkoni, 2021). The common approach is to use regression analysis to test hypotheses based on the statistical significance of the coefficients (Shmueli, 2010). In our exploratory data-driven study with numerous predictors, it is infeasible

to hypothesize and then test different theoretical frameworks connecting the different variables. Therefore we aim to apply causal discovery algorithm on our task.

The goal of causal discovery is to restore causal Directed acyclic graph from observational data. In such a graph an edge represents a relation from a cause to its effect while the absence of an edge between two variables means that they are conditionally independent. Directed acyclic graph is a way to represent the probability distribution factorization in which the probability of a variable is conditioned on its parents. For example:

$$\text{Ⓧ} \leftarrow \text{Ⓩ} \rightarrow \text{Ⓨ} \equiv p(x, y, z) = p(x|z)p(y|z)p(z) \tag{3.21}$$

Causal discovery consists in first determining all dependence and conditional independence relations between variables and then in constructing a graph compatible with these relations. A theorem known as *Markov Condition* (Pearl et al., 2000) states a necessary and sufficient condition for a directed acyclic graph and a probability distribution to be compatible: "*A necessary and sufficient condition for a probability distribution to be compatible with a directed acyclic graph G is that every variable be independent of all its nondescendants in G, conditional on its parents.*"

An undirected graph corresponding to a causal directed acyclic graph is called skeleton. Figure 3.7 shows four different directed acyclic graphs corresponding to one skeleton. Only the v-structure sometimes referred as collider can be unambiguously determined from conditional dependencies between variables. Indeed, v-structure corresponds to the relations $\{x \perp\!\!\!\perp y$, $x \not\perp\!\!\!\perp z$, $y \not\perp\!\!\!\perp z\}$. The other three structures correspond to $\{x \not\perp\!\!\!\perp y$, $x \perp\!\!\!\perp y|z$, $x \not\perp\!\!\!\perp z$, $y \not\perp\!\!\!\perp z\}$. Because different directed acyclic graphs correspond to the same dependencies in data, most causal discovery methods find not fully directed acyclic graph but a *Markov equivalence class*. Markov equivalence class includes all the directed acyclic graphs that have the same skeletons and v-structures (Verma and Pearl, 1990). Thus in practice when applying a causal discovery algorithm we usually obtain a graph with some undirected edges which means that for some related variables we do not know the "cause-effect" direction.

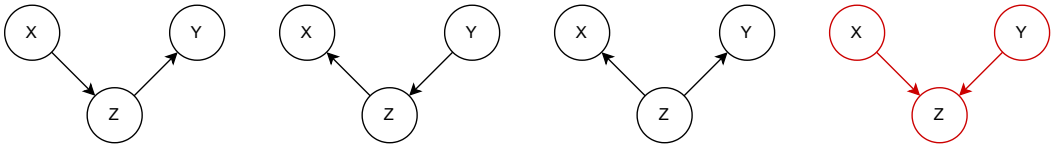


Figure 3.7: Directed acyclic graphs corresponding to the same skeleton. Only the v-structure (in red) can be oriented without ambiguity from observational data.

Many classes of methods have been developed to address the problem of causal discovery. One popular class of methods is the score-based class. Methods belonging to this class start with an empty graph, add currently needed edges, and then eliminate unnecessary edges in a pattern. One of the most known score-based methods is GES (Chickering, 2003) which, is guaranteed to converge to the Markov equivalent class of an acyclic true graph. However, GES is not as easy to illustrate because its trajectory depends on the relative strengths of the associations and conditional associations of the variables (Glymour, Zhang, and Spirtes, 2019). Recently, a new score-based method called NOTEARS was introduced which claimed to achieve state-of-the-art results (Zheng, Aragam, Ravikumar, and Xing, 2018). However, it was however later proven to be an unsuitable method for causal discovery (Kaiser and Sipos, 2021).

More recently, it has been shown that algorithms based on noise are able to distinguish between different directed acyclic graphs in the same Markov equivalence class. However, these kinds of methods usually require additional assumptions about the generative process: for example, LiNGAM (Shimizu, Hoyer, Hyvärinen, and Kerminen, 2006) requires linearity and non-Gaussian noise whereas ANM requires additive noise. Furthermore, such approaches do not scale as well as constraint-based methods (Glymour, Zhang, and Spirtes, 2019).

The most well-known class of causal discovery algorithms is the so-called constraint-based class which relies on conditional independence tests and a small set of orientation rules to identify causal decisions. Within this class, the most popular algorithm is Peter-Clark algorithm (PC), originally introduced in (Spirtes, Glymour, and Scheines, 2000).

The PC algorithm aims at optimizing the number of computations necessary to assess whether two variables are conditionally independent or not by considering conditioning variables that are likely to be parents of the two variables. Even if it grows exponentially with the maximal degree of the graph, large sparse graphs can be easily inferred using the PC algorithm. Starting with a complete undirected graph G, the algorithm checks the dependency for all pairs of vertices and removes or keeps links according to whether or not the two vertices are considered to be independent. Then it checks the conditional independencies between dependent vertices by first computing it for each adjacent pair x and y in G and for each vertex z (other than x) adjacent to y in G. If z is able to remove the dependency between x and y then the algorithm removes the edge between them and adds z to their separation set Sepset(x, y). Then, it gradually increases the number of variables to condition on, and proceeds as above till a conditional independence is found or all sets of vertices adjacent to y have been considered for the conditioning. Once the skeleton has been constructed, the algorithm applies series of rules (Spirtes, Glymour, and Scheines, 2000; Colombo and Maathuis, 2014), starting by identifying v-structures using the so-called origin of causality.

PC-Rule 0 (Origin of Causality): *For every triplet $x - z - y$ such that x and y are not adjacent and $z \notin$ Sepset(x, y), orient the triple as $x \to z \leftarrow y$.*

When all v-structures have been identified using the above rule, the PC algorithm orients as many of the remaining undirected edges as possible, by repeating the following rules until no other changes can be made.

PC-Rule 1: *In a triple $x \to y - z$ such that x and z are not adjacent, orient $y - z$ as $y \to z$*

PC-Rule 2: *If there exists a direct path from x to y and an edge between x and y orient this edge as $x \to y$.*

PC-Rule 3: *Orient $x - y$ as $x \to y$ whenever there are two paths $x - z \to y$ and $x - t \to y$.*

A different orientation in PC-Rule 1 would lead to new v-structures, which is not possible as the origin of causality should identify all v-structures. A

different orientation in PC-Rule 2 would lead to a cycle, whereas a different orientation in PC-Rule 3 would lead to either a cycle or a new v-structure when orienting the remaining undirected edges. From a theoretical viewpoint, the above procedure is sound and complete (Meek, 1995; Andersson, Madigan, and Perlman, 1997) in the set of Markov equivalence graphs, where "sound" means that all causal relations detected by the rules are correct, and "complete" that all possible causal relations in the Markov equivalence class are detected by the algorithm.

Another popular algorithm from constraint-based class is Fast Causal Inference(FCI) algorithm (Zhang, 2008), which unlike PC does not require unconfoundedness assumption. In our preliminary experiments we applied this algorithm on our data. However, we found that FCI results are much harder to interpret than PC results on a dataset with more than 50 variables. On the other hand, by including so many variables in our analysis we hope to minimize the chance of unconfoundedness assumption violation.

In this thesis, we chose the order-independent version of PC because, regardless of the achievements of other methods, PC (without further assumptions) still stands as one of the state-of-art methods: it has the advantage to be generally applicable, is based on a fully developed theory, is nonparametric, and is easily interpretable (Glymour, Zhang, and Spirtes, 2019). We are interested in giving a causal explanation, via the inferred causal graph, to variables that are also predictive with respect to a target variable. To a certain extent, PC can be considered as a bridge between the causal world and the predictive world since the presence of an edge between two variables in the graph inferred by PC necessarily represents a correlation between these variables, whereas the absence of an edge between two variables necessarily represents independence or a correlation that vanishes if conditioned on a subset of other variables; this is not necessarily true for all other methods (Aliferis, Statnikov, Tsamardinos, Mani, and Koutsoukos, 2010).

Using the PC algorithm, we aim to provide valuable insights into the causal relations between various start-up properties and the valuation of the start-up.

In this study, we use the shrinkage estimator for the mutual information (Ledoit and Wolf, 2003) with significance level 0.01 as a statistical inde-

Table 3.6: Top 5 Hyper-Parameters Sorted by Mean R^2 on 5-Fold Cross-Validation on Train Data.

	Iteration	L_2 regularization	Learning rate	Cross-validation mean R^2	Cross-validation standard deviation
I	2000	10	0.05	0.6991	0.0593
II	1000	1	0.05	0.6973	0.0622
III	2000	0	0.01	0.6955	0.0612
IV	1000	0	0.05	0.6939	0.0601
V	2000	1	0.05	0.6938	0.0639

pendence test for the PC algorithm. As this algorithm allows incorporating prior knowledge by forbidding certain edges, we forbid directed edges from valuation to other variables, since these variables are measured before the valuation date, and from all other variables to start-up age. In the experiments, we use the R software package *bnlearn*[8] introduced in Scutari, 2010.

3.3.4 Experiments

Machine Learning Model's Hyperparameter Choice

Our study adopts the standard approach to hyperparameter choice – cross-validation on train data. We use 5-fold cross-validation to find the best set of parameters in the following grid: {'iterations': [500, 1000, 2000], 'learning_rate': [0.005, 0.01, 0.05, 0.1], 'l2_leaf_reg': [0, 1, 10]}. The details about these hyper-parameters can be found on the CatBoost package official website (https://catboost.ai/en/docs/concepts/parameter-tuning). In Table 3.6 we present the top five sets of hyper-parameters. The best set of hyper-parameters (I) was used in the study.

Start-up Valuation Predictors

The overall accuracy of the model learned using all variables amounts to 0.644 for the coefficient of determination R^2 and to 0.443 for the Root Mean Square Error (RMSE). Table 3.7 provides feature groups importance rankings with different measures described in Section 3.3.2. We refer to the model built

[8]More information is available at: https://www.bnlearn.com/

Table 3.7: Feature Group Ranking Analysis

Feature group	SAGE	Group R^2	Group Ablation ΔR^2	Permutation Importance ΔR^2
Financial Capital	0.153	0.466	0.044	0.171
Human Capital	0.084	0.451	0.039	0.074
Online Legitimacy	0.078	0.332	0.028	0.074
Industry and Market Timing	0.044	0.178	0.023	0.043

Notes. R^2 of a model built on the entire set of features is 0.644.

on the entire set of features as full. As described in the model interpretation section, Group Model shows how predictive is a group isolated from other groups, Feature Group Ablation value shows how much worse is the model built in the absence of the feature group. Permutation Importance shows how much the model built on full data relies on the feature group, and finally, SAGE characterizes expectation of feature group usefulness across hypothetical feature groups coalitions.

The first observation to be made is that all measures rank the different feature groups in the same order of importance: Financial Capital is the most important group, followed by Human Capital and Online Legitimacy, relatively close to each other, and then by Industry and Market Timing of lesser importance according to each measure. The second observation is that, according to the feature group ablation tests, removing any single feature group does not strongly affect the model's performance. In addition, the difference between the first two groups, Financial Capital and Human Capital, seems low according the group ablation scores and more important according to the permutation importance scores and SAGE. This means that the model built on all features relies more on Financial Capital than on Human Capital, whereas trained in the absence of each group, the model achieves almost the same accuracy.

SAGE assigns a higher importance to financial capital (SAGE $= 0.153$) compared to other groups of variables for predicting start-up valuation. Human capital and online legitimacy are similar in their predictive power (SAGE $= 0.084$ and SAGE $= 0.078$, respectively), whereas industry and market timing has a lower rank (SAGE $= 0.044$). The R^2 score ranks the groups of variables in the same order; however, the gap between financial capital and human capital becomes smaller ($R^2 = 0.466$ for the former and 0.451

for the latter). This means that, when taken in isolation, financial capital and human capital have a similar predictive power, but when combined with other variables, financial capital becomes a more important group for prediction.

Table 3.8 provides the ranking (third column) of the variables in the four groups analyzed based on their SAGE value (fourth column). It also provides the Spearman correlation coefficient (fifth column) of each variable with the valuation as well as the R^2 scores and SAGE values (first column) of models built with each group of variables. It is interesting to note that the top ten variables, in terms of importance as measured by SAGE, belong to all four groups. Out of these top ten variables, the five most predictive are past funding, start-up age, current team, assigned officers, and search from Techcrunch.com. Lastly, a negative SAGE value suggests that the variable is not interesting for prediction purposes when combined with the other variables. Indeed, training a new model excluding the 11 variables with negative SAGE values resulted in a model which is slightly more accurate according to the R^2 score, which now amounts to 0.657 on the test set, and similar according to the RMSE, which amounts to 0.435 on the test set.

Observing SAGE values and Spearman correlation coefficients, one can see that being strongly correlated with the valuation does not necessarily imply being a good predictor of the valuation. For example, the two most correlated with the valuation variables, namely past funding and start-up age, are also the best predictors according to SAGE. This is not the case for the variable funding rounds, which is well correlated with valuation but has a negative SAGE value meaning that it may be preferable to dispense with this variable when predicting with all other variables. A possible explanation for the observed difference between some correlation and importance scores is the distribution shift between the train and test datasets: the average value of the funding rounds is 1.4 in the train set and 1.7 in the test set. The nature of this difference is hard to identify: on the one hand, the fundraising process for start-ups may be changing over time; on the other hand, the reason might be the incompleteness of the information about old funding rounds in Crunchbase.

Table 3.8: Variables' Predictive Power Ranking and Causal Relations to Startup Valuation.

Variable group	Variable (Ranked by SAGE)	Predictive power			Causal relation			
		SAGE rank	SAGE	Spearman ρ	Direct causes	Indirect causes	Siblings	Other
Financial capital (SAGE = 153, R^2 = 0.466)	**Past funding**	1	155.33	0.56	X			
	Past crowdfunding	36	0.43	0.18			X	
	Crowdfunding campaigns	50	-0.09	0.18		X		
	Funding rounds	57	-5.05	0.48	X			
Human capital (SAGE = 84, R^2 = 0.451)	**Current team**	3	21.91	0.35	X			
	Assigned officers	4	19.32	0.51	X			
	Money raised by previous companies	6	10.33	0.38		X		
	Officer age	10	6.65	0.40		X		
	Past appointments	12	5.85	0.39		X		
	Foreign officers	14	4.75	0.15		X		
	Different nations	17	3.76	0.34	X			
	American officers	18	3.36	0.26	X			
	British officers	19	2.66	-0.08		X		
	Female officers	20	2.00	0.10				X
	PhD degree	21	1.78	0.15			X	
	Age diversity	22	1.62	0.21		X		
	Managerial experience	24	1.49	0.23		X		
	Officers w/ director experience	25	1.39	0.23		X		
	Officers w/ investor experience	28	1.17	0.20	X			
	French officers	32	0.79	0.13				X
	Officers w/ CEO occupation	33	0.78	0.11				X
	Officers w/ investment occupation	37	0.36	0.23		X		
	Officers w/ finance occupation	39	0.28	0.15		X	X	
	Time in startup diversity	40	0.28	0.28		X		

Continued on next page

Table 3.8 Continued

Variable group	Variable (Ranked by SAGE)	Predictive power			Causal relation			
		SAGE rank	SAGE	Spearman ρ	Direct causes	Indirect causes	Siblings	Other
	Irish officers	43	0.24	0.09				X
	German officers	45	0.08	0.15		X		
	Officers w/ consultant occupation	47	-0.01	0.08				X
	Officers w/ manager experience	48	-0.05	0.07				X
	Officers w/ consultant experience	51	-0.10	0.21				X
	Officers w/ engineer occupation	53	-0.28	0.08				X
	Board and advisors	54	-0.54	0.24		X		
Online legitimacy (SAGE = 78, $R^2 = 0.332$)	**Search results from techcrunch.com**	5	10.60	0.21	X			
	Twitter users replied	7	10.26	0.17		X		
	Twitter likes	8	10.04	0.21	X			
	Relevant search results	9	8.59	0.23		X		
	Search results from own domain	11	6.41	0.28	X			
	News	13	5.66	0.39			X	
	Twitter user mentions	15	4.70	0.14		X		
	Twitter replies	16	4.63	0.17		X		
	Search results from facebook.com	23	1.55	-0.10				X
	Search results from angel.co	26	1.25	-0.09		X		
	Search results from twitter.com	27	1.17	-0.08			X	
	Search results from tech.eu	30	1.05	0.15		X		
	Tweets	31	0.93	0.16		X		
	Search results from eu-startups.com	34	0.72	0.14		X		
	Twitter medias	35	0.51	0.10				X

Continued on next page

Table 3.8 Continued

Variable group	Variable (Ranked by SAGE)	Predictive power			Causal relation			
		SAGE rank	SAGE	Spearman ρ	Direct causes	Indirect causes	Siblings	Other
	Search results from crunchbase.com	38	0.33	-0.07				X
	Tweets length	41	0.27	0.20			X	
	Search results from linkedin.com	42	0.24	-0.08			X	
	Search results from medium.com	44	0.16	-0.06		X		
	Search results from crowdfundinsider.com	46	0.08	0.08				X
	Search results from uktech.news	49	-0.08	0.11		X		
	Twitter unique hashtags	52	-0.23	0.13				X
	Startup age	2	38.73	0.56	X			
Industry and market timing (SAGE = 44, R^2 = 0.178)	Industry costliness	29	1.16	0.38	X			
	Startups founded in industry last year	55	-2.17	-0.16				X
	Funding raised in industry last year	56	-2.26	0.20			X	

Notes. SAGE - Shapley Additive Global importancE, multiplied by 1000 for readability. Spearman ρ - Spearman rank correlation coefficient significant at p-value < 0.01. We cannot determine whether the variable officers w/ finance occupation is an indirect cause or a sibling, because the direction of the edge connecting it to a direct cause is unknown. Top-10 variables by their predictive power are highlighted in bold.

Causal Determinants of Start-up Valuation

To discover the observable causes of start-up valuation, we build the causal structure of the variables using the PC algorithm. Running the PC algorithm on all 58 variables yielded a causal graph with 98 directed edges, corresponding to cause-to-effect relations, and 10 undirected edges which correspond to edges PC was not able to orient[9]. To further describe the resulting graph structure, we adopt the following terms: 1) X is a *direct cause* of Y if there is a directed edge $X \to Y$ in the graph, 2) X is a *sibling* of Y if there exists a node Z such that directed edges $Z \to X$ and $Z \to Y$ are both present

in the graph and 3) V_1 is an *indirect cause* of V_k if there is a directed path from V_1 to V_k: $V_1 \rightarrow V_2 \rightarrow ... \rightarrow V_k$. We choose to highlight these types of relations to the valuation for the following reasons. First, given how the PC algorithm works, we know that direct causes are significantly correlated with the valuation even given all other variables. It is of course possible that, for one of the direct causes and the valuation, there is an omitted common cause or intermediate cause that we could not measure. However, the number and diversity of the variables included in this study helps alleviate this concern. Further, siblings and indirect causes are independent from valuation given (a subset of) direct causes. The difference between siblings and indirect causes is that, assuming that our graph is correct, changing an indirect cause would affect valuation, whereas changing a sibling would not.

Table 3.8 shows the causal relations between each variable and the valuation of a start-up, as derived by the PC. Figure 3.8 depicts the valuation, its direct causes and first indirect causes, i.e., the direct causes of the direct causes. Figure 3.9 shows the valuation, its direct causes and siblings. As can be seen in Table 3.8, PC identifies only 12 variables as direct causes of valuation, 24 variables as indirect causes, and eight variables are siblings. In the financial capital group, past funding and funding rounds directly affect the valuation. The direct causes from the human capital group include current team, assigned officers, different nations, American officers, and officers with investor experience. At the same time, some highly correlated with start-up valuation variables of the group such as mean officer age, officer past appointments, and money raised by previous companies do not directly affect valuation in our model. The online legitimacy variables that directly affect valuation are search results from Techcrunch.com, Twitter likes, and search results from own domain. Finally, two variables of the industry and market timing group, namely start-up age and industry costliness, are identified as direct causes of valuation.

[9]The PC algorithm outputs the completed partially directed acyclic graph which represents the Markov equivalence class of the true causal graph. As such, some edges may not be oriented.

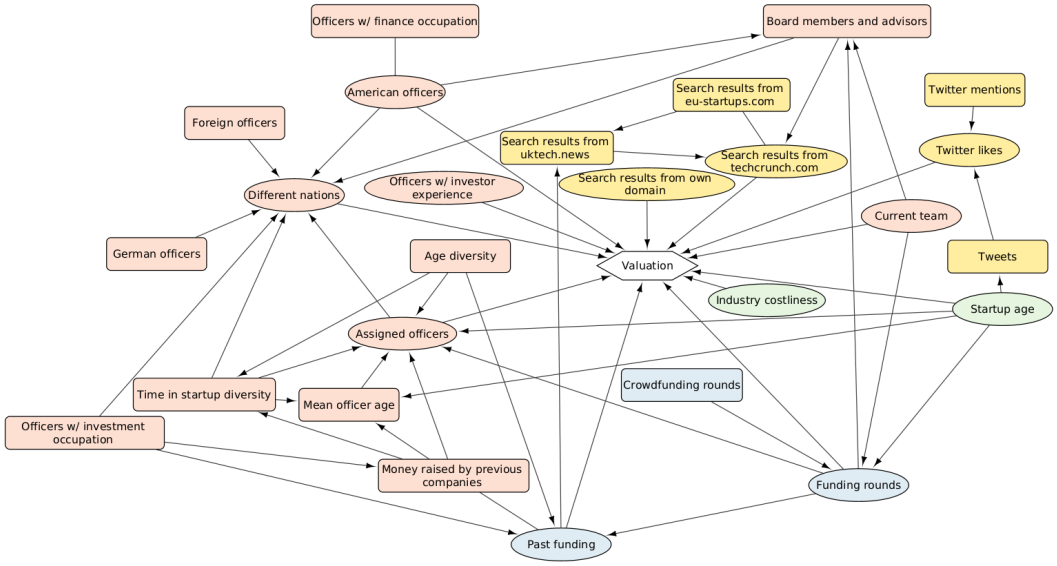


Figure 3.8: Causal graph computed by PC algorithm: direct and first indirect causes of valuation.

Notes. Ellipse shape nodes denote direct causes and rectangle shape nodes denote first indirect causes. Color indicates the variable group: blue - financial capital, yellow - online legitimacy, red - human capital, green - industry and market timing. Edges without arrows indicate possible indirect causes.

Comparison of predictive and causal discovery analyses

Table 3.8 allows comparing the predictors with the causal determinants of start-up valuation. The median SAGE value for the variables identified as direct causes is 8.2, for indirect causes the median SAGE value is 1.25, for siblings 0.80, and for other variables 0.16. These results demonstrate that direct causes of the valuation also make the best predictors. We will further refer to a predictor as good if its SAGE value is above or equal to 1.16, which corresponds to the median SAGE value over all variables. Figure 3.10 presents the distribution of good and bad predictors among each type of causal relations. As can be seen, all direct causes except funding rounds are good predictors (11 out of 12 or 92%) and more than half of the indirect causes (13 out of 24 or 54%) are also good predictors. About 38% of siblings (3 out of 8) and only two variables out of 14 (14%) from the other group are good predictors. This comparison confirms that the variables directly influencing start-up valuation also tend to be good valuation predictors. While changes in direct or indirect causes can affect start-up valuation,

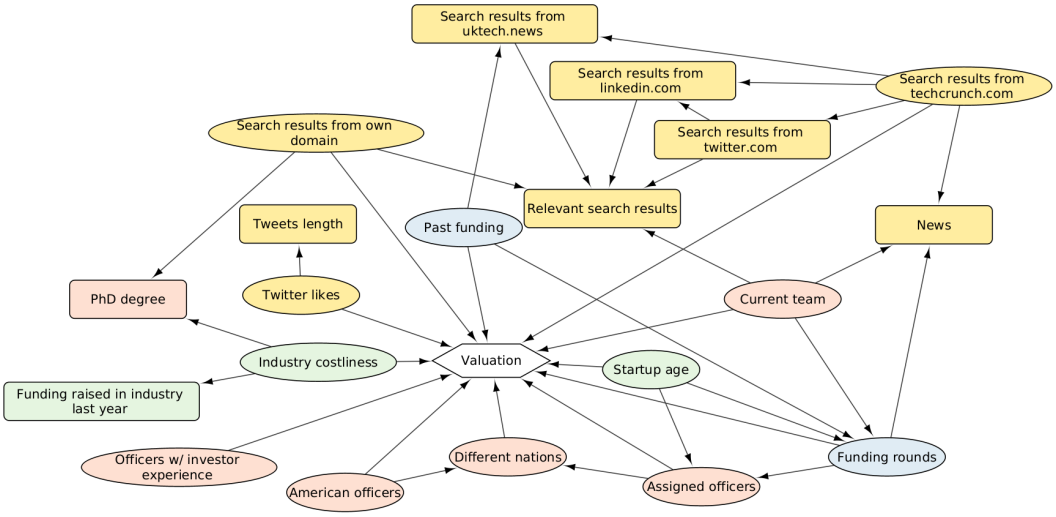


Figure 3.9: Causal graph computed by PC: direct causes of valuation and their siblings.

Notes. Ellipse shape nodes denote direct causes and rectangle shape nodes denote their siblings. Color indicates the different groups of variables: blue - financial capital, yellow - online legitimacy, red - human capital, green - industry and market timing.

changes in siblings or other variables should not. Overall, 83% of good predictors are also causal determinants (direct or indirect) that influence start-up valuation. Inversely, 67% of direct or indirect causal determinants are also good predictors which allow distinguishing between higher and lower value of a start-up.

The above results show that there is indeed a strong relation between causal determinants and good predictors of valuation. However, variables that may not be direct or indirect causes of valuation can still be useful for prediction as they may contain useful correlation, possibly specific to the dataset studied. This explains why some variables obtain positive and relatively high SAGE values and Spearman coefficients while not having a direct or indirect effect on the valuation.

Robustness of Empirical Findings

In order to estimate the robustness of our empirical findings, we performed a series of experiments which aim to estimate the stability of (a) the SAGE values and rankings of variables, with respect to the hyper-parameters of

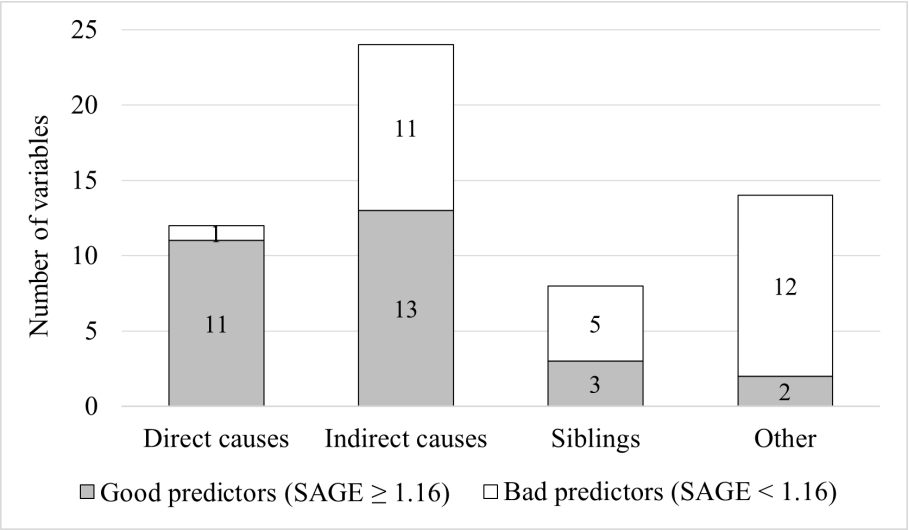


Figure 3.10: Comparison of predictors and causal determinants of start-up valuation.

Table 3.9: SAGE Values Obtained from ML Models with Different Hyperparameters.

SAGE ranking	Feature	SAGE	SAGE variance	SAGE	SAGE
1	Past funding	155.33	3.93	155.21	166.93
2	Startup age	38.73	1.24	39.60	36.28
3	Current team	21.91	0.78	21.51	20.92
4	Assigned officers	19.32	0.69	21.96	18.58
5	Search results from techcrunch.com	10.60	0.58	10.29	10.54
6	Money raised by previous companies	10.33	0.38	9.66	8.52
7	Twitter users replied	10.26	0.66	11.43	9.90
8	Twitter likes	10.04	0.62	10.11	11.55
9	Relevant search results	8.59	0.38	9.62	7.90
10	Mean officer age	6.65	0.37	5.42	8.16
11	Search results from own domain	6.41	0.30	6.06	6.52
12	Mean officer past appointments	5.85	0.25	7.25	3.13
13	News	5.66	0.22	5.42	5.57
14	Foreign officers	4.75	0.29	4.84	3.37
15	Twitter mentions	4.70	0.40	5.04	4.15
16	Twitter replies	4.63	0.43	4.60	3.99
17	Different nations	3.76	0.16	3.60	3.16
18	American officers	3.36	0.20	3.75	5.49
19	British officers	2.66	0.23	2.50	3.86
20	Female officers	2.00	0.15	1.95	1.66
21	PhD officers	1.78	0.22	2.41	1.84

Continued on next page

Table 3.9 Continued

SAGE ranking	Feature	SAGE	SAGE variance	SAGE	SAGE
22	RSD officer age	1.62	0.18	1.31	2.22
23	Search results from facebook.com	1.55	0.16	1.64	1.22
24	Management experience	1.49	0.10	1.81	1.52
25	Officers w/ director experience	1.39	0.11	1.28	1.18
26	Search results from angel.co	1.25	0.14	1.52	1.02
27	Search results from twitter.com	1.17	0.15	1.26	1.75
28	Officers w/ investor experience	1.17	0.11	1.21	1.10
29	Mean industry costliness	1.16	0.27	-0.14	3.01
30	Search results from tech.eu	1.05	0.10	0.87	1.01
31	Tweets	0.93	0.30	0.21	0.82
32	French officers	0.79	0.06	0.83	0.80
33	Officers with CEO occupation	0.78	0.10	0.62	0.58
34	Search results from eu startups.com	0.72	0.08	0.69	0.78
35	Twitter medias	0.51	0.35	0.58	0.67
36	Past crowdfunding	0.43	0.05	0.11	0.54
37	Officers w/ investment occupation	0.36	0.08	0.41	0.27
38	Search results from crunchbase.com	0.33	0.06	0.36	-0.02
39	RSD time in company	0.28	0.24	0.07	2.11
40	Officers w/ finance occupation	0.28	0.04	0.19	0.22
41	Tweets length	0.27	0.38	0.34	0.56
42	Search results from linkedin.com	0.24	0.10	0.41	0.74
43	Irish officers	0.24	0.07	0.25	0.12
44	Search results from medium.com	0.16	0.07	0.10	0.19
45	Search results from crowdfundinsider.com	0.08	0.03	0.05	0.07
46	German officers	0.08	0.02	-0.63	0.20
47	Officers w/ consultant occupation	-0.01	0.02	-0.00	-0.03
48	Officers w/ manager experience	-0.05	0.08	0.17	0.18
49	Search results from uktech.news	-0.08	0.02	0.02	-0.06
50	Crowdfunding campaigns	-0.09	0.02	-0.07	0.04
51	Consulting experience	-0.10	0.09	0.05	0.00
52	Twitter unique hashtags	-0.23	0.51	0.12	0.28
53	Officers w/ engineer occupation	-0.28	0.06	-0.08	-0.09
54	Board members and advisors	-1.54	0.12	-1.72	-0.31
55	Mean startups founded in industry LY	-2.17	0.24	-2.32	-0.30
56	Mean money invested in industry LY	-2.26	0.37	-2.15	-1.77
57	Funding rounds	-5.05	0.31	-4.29	-3.78

the predictive model and (b) of the causal relations obtained. For this latter analysis, we study how the relations change if one slightly changes the set of

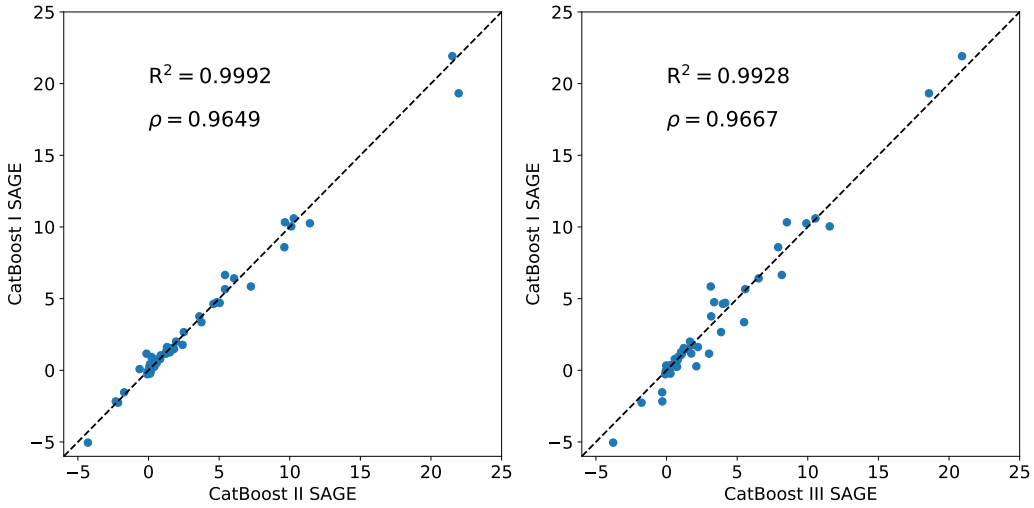


Figure 3.11: SAGE values obtained from ML models with different hyperparameters. Coefficients of determination R^2 and Spearman rank correlation values are given.

variables or modifies the threshold used in the independence tests at the basis of the PC algorithm. This stability analysis reveals that the SAGE rankings are stable across the three best models. Similarly, the causal relations are also relatively stable to the thresholds used to select the variables and to assess their (in)dependence.

SAGE Feature Ranks

We study here whether the rankings provided by SAGE are stable across different hyperparameter values. To do so, we focus on the three best models (CatBoost I , CatBoost II, CatBoost III) the hyperparameters of which are given in Table 3.6. As shown in Table 3.9, the difference between the best model on the cross-validation set and the second and the third best models is relatively small (less than 0.004 on the R2 score). In addition, the R2 score on the test set is stable and amounts to 0.649, 0.651, and 0.645 for CatBoost I, CatBoost II, and CatBoost III, respectively. The similarity, in terms of performance, of the different models may be surprising since the hyperparameters they rely on differ (CatBoost II and CatBoost III differ from CatBoost I on two out of three hyperparameters and don't have any hyperparameter in common).

Table 3.10: Stability of PC results with respect to the set of variables considered: confusion matrix for causal relations obtained on Set 0 (main paper) and Set II (with Spearman p-value set to 0.05).

	Direct cause (p=0.05)	Indirect Cause (p=0.05)	Sibling (p=0.05)	Other (p=0.05)
Direct cause (p=0.01)	11	1	0	0
Indirect cause (p=0.01)	0	21	1	1
Sibling (p=0.01)	1	0	6	1
Other (p=0.01)	0	3	0	11

We display in Table 3.9 the SAGE values for each feature and each model as well as the variance estimates of the SAGE values (see page 6 of Covert et al. (2020)) for CatBoost I.

As can be seen from Table 3.9, the SAGE values obtained by the models CatBoost I, CatBoost II, and CatBoost III are similar, at least for the most important features. Only for 26% and 35% of the variables, respectively CatBoost II and CatBoost III SAGE values differ from CatBoost I SAGE values by more than one third. Among the variables, with SAGE values above-median the stability is higher and the percentage of variables that change their SAGE value by more than one third is 7% and 21% for models CatBoost II and CatBoost III respectively.

To further illustrate the relations between SAGE values obtained with different models, we plot them against each other in Figure 3.11. We also calculate coefficients of determination R2 and Spearman rank correlation values. Our results show a high similarity of SAGE values and rankings obtained with different models. Indeed, the coefficient of determination and Spearman rank correlation coefficient values in Figure 3.11 are high (above 0.99 for R^2 and above 0.96 for the Spearman rank correlation coefficient), which further shows that both the variables' SAGE values and SAGE rankings are consistent across the models. We can thus conclude that the empirical findings on the variables' predictive power are stable across the top models.

PC Results Stability

We perform additional analyses to estimate the stability of our causal discovery results. In the first analysis, we test how much the PC algorithm results change if the set of potential causal determinants is slightly modified. To this

end, we modified the variable filtering process by changing the threshold on the p-value of the Spearman correlation coefficient to the target variable from 0.01 (original set, denoted Set 0, used in the main study) to 0.025 and 0.05. This procedure yielded two additional sets of variables, referred to hereafter as Set I and Set II, of 58 and 60 variables additionally containing Search results from forbes.com, for both sets, and Twitter retweets, and Officers with director occupation variables, for Set II. We then ran the PC algorithm on these augmented sets and compared the resulting graphs. We found that the sets of direct causes remain almost intact for both models, with only the variable Twitter likes being replaced by the variable News in both graphs . Compared to the original graph, 89% of the variables of the graph obtained with Set I belong to the same causal group: direct causes, indirect causes, siblings, other. This proportion drops to 78% with Set II, which is still important.

Table 3.10 displays the confusion matrix of the causal relations obtained with Set 0 and Set II. In this table, the numbers on the diagonal correspond to the number of variables which have the same causal relation to the valuation in both models. The numbers outside the diagonal correspond to the number of variables that have different causal relations to the valuation in the two models. For example, 11 variables are direct causes in both models, and one variable that is a direct cause in Set 0 is an indirect cause in Set II.

The second analysis that we run concerns the stability of the PC algorithm results with respect to the parameter α, which corresponds to the threshold on the p-value of the statistical test used in PC to assess whether variables are independent or not.

Figure 3.12 displays the Hamming distance between the causal graph presented in the main study, obtained with $\alpha = 0.01$, and causal graphs obtained with. Higher values of α (0.02, 0.03, 0.04 and 0.05). The Hamming distance between two graphs with the same set of nodes is equal to the number of addition/deletion operations required to turn one graph into another. For example, a Hamming distance of 27 between two graphs can mean that we need to delete 10 edges and add 17 to go from one graph to another or remove 1 edge and add 26, etc. As one can note, in the worst case, around three-quarters of the edges remain the same, which correspond to roughly

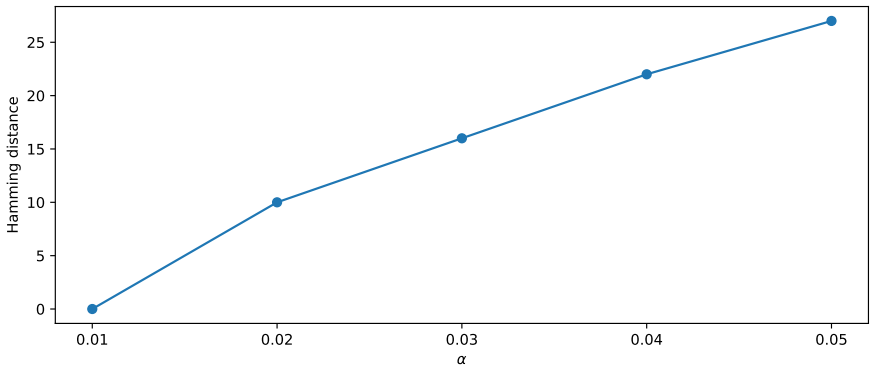


Figure 3.12: PC results with respect to the p-value used by PC algorithm (hyperparameter α).

81 edges out of 108. It is worth noticing that α values outside the (0.01, 0.05) range are not typically used in the literature and that, as the Hamming distance is defined only between graphs with the same set of nodes, so we did not use it in our previous analysis.

To further look into the difference between the graph models obtained by the PC algorithm with different values for α, we display in Table 3.3 the confusion matrix of the causal relations obtained with $\alpha = 0.01$ and $\alpha = 0.05$. As one can note, the most important group for our analysis, namely direct causes, remains almost unchanged: indeed, all the 12 variables that were identified as direct causes by the PC algorithm with $\alpha = 0.01$ are identified as direct causes with $\alpha = 0.05$ as well. The majority (74%) of the indirect causes identified with $\alpha = 0.01$ are still indirect causes with with $\alpha = 0.05$. This percentage however drops to 50% for siblings and other causal relations.

As can be seen from Figure 3.2 and Table 3.3, changes in values of the parameter α of PC lead to almost no changes in the direct causes identified and to moderate changes for indirect causes. Siblings and other relations are definitely more impacted by a change in the value α. This said, these latter relations are less interesting for determining the actual causes of start-up valuation.

Table 3.11: Stability of PC results with respect to the p-value used by PC (hyperparameter α) : confusion matrix for causal relations with $\alpha = 0.01$ (main paper) and $\alpha = 0.05$.

	Direct cause ($\alpha = 0.05$)	Indirect Cause ($\alpha = 0.05$)	Sibling ($\alpha = 0.05$)	Other ($\alpha = 0.05$)
Direct cause ($\alpha = 0.01$)	12	0	0	0
Indirect cause ($\alpha = 0.01$)	0	17	5	1
Sibling ($\alpha = 0.01$)	2	4	1	1
Other ($\alpha = 0.01$)	0	5	2	7

3.3.5 Discussion

The abundance of available data and extensive developments in data science have made AI methods increasingly applied in entrepreneurship research (Obschonka and Audretsch, 2019). This study continues the emerging research strand on using data science methods to approach the start-up valuation problem. The novelty of this study is the comparison of the predictive power ranking and the causal discovery of the factors of start-up valuation.

The data science methods in entrepreneurship have raised an interest towards quantitative inductive studies (Schwab and Zhang, 2019). The outputs of these studies can serve as inductive inputs for subsequent deductive tests to validate and refine the exploratory findings (Kolkman and Witteloostuijn, 2019). This study has applied ML algorithms and causal discovery for predicting and exploring the valuation of start-ups based on a set of characteristics informed by previous studies. Our findings quantitatively produced a list of factors which are the most relevant for valuation prediction and determined complex causal relationships between them. For researchers, it offers rich data-driven information that can be taken as the elements for further theorizing. For example, Lévesque, Obschonka, and Nambisan (2020) proposed a framework for integrating AI with theory testing and theory building. A design thinking perspective and shared problem solving was recently proposed in entrepreneurship research (Hyytinen, 2021) focusing on problems in which both practitioners and scholars are interested (Kleinberg, Ludwig, Mullainathan, and Obermeyer, 2015). Future studies are suggested to consider alternating between inductive investigations and deductive tests on different datasets to produce explanation next to prediction and develop

more powerful predictive models and theories (Hofman, Watts, Athey, Garip, Griffiths, Kleinberg, Margetts, Mullainathan, Salganik, Vazire, Vespignani, and Yarkoni, 2021; Schwab and Zhang, 2019; Shmueli, 2010). This study offers a good starting point to extend the understanding of predictors and causal determinants of start-up valuation.

Combining Predictive Power and Causal Relations of Start-up Valuation Factors

By combining the predictive and causal discovery analyses of start-up valuation factors, this study enriches the literature on investment and valuation of new firms in several ways (e.g., Block, Fisch, Vismara, and Andres, 2019; Mason and Stark, 2004; Miloud, Aspelund, and Cabrol, 2012). First, by leveraging the ML algorithms, it empirically shows which financial capital, human capital, online legitimacy, and industry and market timing variables are the most predictive for the valuation of start-ups. Second, while prior research in the field has been generally focused on the correlations between factors and the start-up valuation (e.g., Tumasjan, Braun, and Stolz, 2021), our study explores the causal structure of valuation factors driven by observed (or observational) data. To the best of our knowledge, causal discovery analysis has not been used before in the context of start-up valuation. Third, this study compares the results from the two analyses and provides insights into the similarities and differences between predictors and causal determinants of start-up valuation, which is a largely underexplored field in the start-up valuation literature. In the following, we discuss the different sets of variables identified by the analyses and outline the potential avenues for further investigations.

Financial capital. This group was shown to have the highest predictive power for the start-up valuation. Past funding obtained by a start-up as well as the number of previously secured funding rounds directly affect valuation in our analysis. These observations are in accord with the literature emphasizing financial criteria in investment decision-making (e.g., Block, Fisch, Vismara, and Andres, 2019). Financial information is important for investors to assess the potential level of profitability, the obtained financial resources, and the required amount of funding in the future (Mason and Stark, 2004).

Interestingly, although past funding has the highest predictive power among all variables, the number of previously secured funding rounds does not contribute to valuation prediction accuracy in most variables' ensembles. This finding illustrates that even if the variable is both strongly correlated with the target and directly affects it, it might not help solving the prediction task in a realistic setting. It is possible that the amount of money raised in a funding round might be more important than the number of funding rounds, and start-ups with only a few funding rounds might achieve high valuations if the amounts of these funding rounds are large. Similarly, the number of crowdfunding campaigns led by the start-up was not found to be a good valuation predictor but appeared to indirectly influence valuation by affecting the number of previously secured funding rounds. Future research is suggested to further investigate the role of fundraising and crowdfunding for predicting and explaining start-up valuation.

Human Capital. Our study addresses the call for including a richer set of human capital characteristics, particularly related to team heterogeneity, as the factors of start-up valuation (Köhn, 2018). In our analysis human capital group is the second most predictive after financial capital, and several variables from this group are also the causal determinants of the valuation.

Specifically, the number of different nationalities in a start-up is identified as a direct cause of its valuation. Different nations and closely related foreign officers variables are also found to be better than average predictors of start-up valuation. The national diversity of top management teams is an under-researched topic in the literature (Nielsen, 2010) with only a few studies attempting to establish the link between the team national diversity and venture success. For instance, Steffens, Terjesen, and Davidsson (2012) showed that a management team's national diversity might be beneficial only under certain conditions, whereas Vogel, Puhan, Shehu, Kliger, and Beese (2014) found a positive association between a management team's national diversity and investors' assessment of a start-up. Hart (2014) suggested that nationally diverse teams might be more effective because of the broader professional network but quantitatively found a modest impact of national diversity on firm performance and highlighted the need for a larger dataset. Our results extend prior studies by empirically specifying the role of national diversity as both a good predictor and direct cause of valuation on a

large dataset and in the presence of all other variables used in the analysis. Moreover, our findings identified the presence of American officers in the team as a good predictor and a direct cause of start-up valuation. The role of connections with the US start-up ecosystem may serve as an explanation and a potential avenue for further investigation.

Furthermore, the investor experience of officers was found to be a direct cause to start-up valuation while also being a better than average predictor of valuation. Some start-ups list their investors among officers; in our sample, 7% of start-ups have at least one officer with a venture capitalist occupation. The majority of start-ups' officers with investor experience are investors in these start-ups. While empirically validating the results of prior research that experienced investors tend to invest in more successful start-ups (Blohm, Antretter, Sirén, Grichnik, and Wincent, 2020) that achieve higher valuations, our findings extend the valuation literature by showing that investor experience of officers helps predict start-up valuation as well as influences it.

Another observation from our results is that the variables of team experience (officer age, past appointments, and money raised by previous companies) are among the best predictors of valuation but not direct causes of it. This result extends previous findings on a positive relationship between team start-up experience and venture performance (Steffens, Terjesen, and Davidsson, 2012) by specifying the nature of this relationship. Similarly, the role of age heterogeneity has been specified by showing that, although being a good predictor, it is not directly linked to valuation, but rather affects start-up's past funding that, in turn, affects valuation.

Online legitimacy. Previously, start-up's web presence was studied in the context of ML prediction of start-up success (Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke, 2018; Xiang, Zheng, Wen, Hong, Rose, and Liu, 2012). In our study, we go one step further and identify the causal effect of various web visibility variables on start-up valuation while also assessing their usefulness for start-up valuation prediction. Our findings show that the search results from the start-up's own domain directly affect valuation in our model. This variable characterizes the web visibility of a start-up. Indeed, if one searches for a start-up's name but does not find its website among the

first ten search results provided by Google, the start-up is not visible enough on the web. This variable also contributes to accurate valuation prediction, ranking 11th out of 57 variables.

Furthermore, the search results from Techcrunch.com, a specialized American newspaper, is another important predictor and a direct cause of start-up valuation with respect to observed variables. Although a start-up's appearance in the news can affect its visibility and success, the opposite is also possible: successful start-ups might have more chances to get into the news. Intriguingly, our results reveal that the number of start-up appearances in any news source is a good valuation predictor, however it does not directly affect valuation but instead has a common cause with it. Hence, it is possible that not any news appearances may influence success, but only the ones from specialized reputable newspapers. Similarly, other web visibility variables including the number of relevant search results and search results from particular websites such as facebook.com, angel.co, and twitter.com are found to be good predictors of start-up valuation but not the direct determinants of it. We believe that web visibility of start-up is an avenue worth investigating in future research.

Our findings also contribute to the research on the role of social media for start-up valuation. Antretter, Blohm, Grichnik, and Wincent (2019) and Blohm, Antretter, Sirén, Grichnik, and Wincent (2020) have shown that various characteristics of a start-up's activity in Twitter can be used to predict new venture survival. In accordance with their results, we find that two Twitter-related variables, namely Twitter users replied and Twitter likes, are among the top 10 best valuation predictors; also, Twitter user mentions and Twitter replies are good predictors. Our causal discovery analysis offers insights into the causal relations between Twitter variables and the valuation. In our model, Twitter likes is the only direct cause of valuation among other Twitter variables. Most other Twitter-related variables are indirect causes of valuation, affecting it through Twitter likes or other direct causes in our graph model. This suggests that valuation may depend more on the social appreciation than on an entrepreneur's effort to build online legitimacy. However, the number of Twitter likes, in turn, depends on the number of Tweets and Twitter mentions of other users by a start-up, suggesting that Twitter activity is indirectly related to valuation through the

social appreciation. Recognizing the complex relationships between social media variables helps advance the conversation on start-up valuation factors. To this end, future studies are suggested to analyze and compare the role of different social media networks for start-up valuation.

Industry and Market Timing. This variable group is ranked the last in our predictive power analysis. Low predictive power of the variable group can be partially explained by a smaller number of variables compared to other groups. Two variables from this variable group, namely start-up age and industry costliness, are found to directly affect valuation. Moreover, both these variables are also good valuation predictors with start-up age being the second most predictive variable after past funding. If the data permit, future empirical research is encouraged to explore a richer set of industry-related variables and their predictive power and causal relations with the start-up valuation.

Implications for Investors and Entrepreneurs

Assessment of a start-up's value is a critical task when making investment decisions. Investors analyze multiple criteria related to financial, organizational, and market characteristics (Block, Fisch, Vismara, and Andres, 2019; Festel, Würmseher, and Cattaneo, 2013). Data-driven approaches can help investors to scout investment opportunities and limit biases (Blohm, Antretter, Sirén, Grichnik, and Wincent, 2020). By analyzing the relative importance of various factors for predicting start-up valuation, we inform investors about the most relevant characteristics when evaluating and selecting early-stage companies. Our findings guide investors on the data collection priorities. Indeed, collecting data about the enterprises is a tedious task, and one might need to decide what type of information is the most useful. Although financial capital is the most important player, not all financial factors were found useful for the valuation prediction. Moreover, early-stage start-ups often lack financial data (Baum and Silverman, 2004) or its collection may not be possible.

In light of these constraints, our study shows that a good valuation prediction can be achieved by combining the other important factors reflecting human

capital, online legitimacy, and industry and market timing. Past funding, start-up age, team size (current team and assigned officers), team experience (officer age and money raised by previous companies), web visibility (search results from techcrunch.com and relevant search results), Twitter activity and appreciation (Twitter users replied and Twitter likes) were identified among the best predictors of start-up valuation. Furthermore, the findings inform investors about which factors are the causal determinants that allow influencing the start-up valuation. This is particularly relevant for mentoring start-ups in the investor portfolio and helping them grow into high-value companies. The direct and indirect causes of valuation provide an overview of how the valuation factors are interconnected. For example, results show that the Twitter activity of entrepreneurs does not help to improve start-up valuation if there is no social appreciation of the venture expressed by Twitter likes.

Understanding which factors are the attributes of high-valued start-ups and which are the drivers of stat-up valuation is also beneficial for entrepreneurs, since it helps them to set up a direction for the efforts and resources they put into their business. For example, they might consider increasing the national diversity of their team to benefit from international connections or increase the web visibility of their venture by improving their website or social media postings, to raise the interest of their target audience and gain more appreciation.

3.4 Conclusion

In this chapter, we studied startup valuation problem from different perspectives. We explored the discrepancy between the distribution properties of the funding rounds in which the startup valuation is given to the public with the ones in which it is left private. To tackle the difficulties that this difference imposes on machine learning model training, we applied different domain adaptation methods.

In the second part of our research, we studied how well a startup valuation can be predicted without the funding amount information and studied which

variables make the best startup valuation predictors. Furthermore, to the best of our knowledge, we were the first to apply causal discovery methods to the startup valuation problem. This analysis allowed us to gain valuable insights into the factors that directly and indirectly affect startup valuation.

4

Startup Fundraising

As discussed in Chapter 2, most studies of startup success rely on manually curated features available in proprietary databases such as Crunchbase[1] and describing, *e.g.*, the previous funding events of a given startup, with the amount raised, their type (as seed, angel or venture), the number of rounds without valuations, etc. Maintaining such a database is nevertheless a costly, time consuming task that is furthermore likely to be incomplete in the sense that it is very difficult to be exhaustive in the startups covered. For these reasons, we investigate here the possibility to predict funding events of startups by only resorting to features that can automatically be extracted from freely, publicly available sources of information.

More precisely we want to predict whether a startup will secure funding round in a given amount of time (*horizon*) given its feature vector, which corresponds to a traditional binary classification task. Our contribution on this is twofold:

1. First, we show how one can extract a rich set of features describing startups from freely, publicly available sources of information as startup websites, social media and company registries;

2. Second, we show that, by using state-of-the-art ML methods with these features, one can obtain prediction results that rival the ones obtained with manually curated features.

[1] https://www.crunchbase.com/

4.1 Data Collection

In this section, we describe in detail the process of collecting the data required to solve the prediction task. The first step in this process is to build a list of the startups for analysis, along with the links to their websites. To do that, one can use any sufficiently large list of startups available on the web. In our case, we gathered 22K startups from multiple sources, such as hubs, investors and conferences, across the world. Once having collected the list of the startups websites, we extract information from the following sources:

- Startup's own website,
- Twitter API[2],
- Google search API[3],
- Country-specific registration data on companies (*e.g.* Infogreffe [4]) containing information about firms, such as the office locations and number of employees.[5]

Distinguishing feature of our dataset is its geographical variety of the companies. While most of the previous works focus on the startups from the USA (Giardino, Wang, and Abrahamsson, 2014; Zhang, Ye, Essaidi, Agarwal, Liu, and Loo, 2017), we analyze the startups all across the world with a slight focus on Europe. Figure 4.1 illustrates the distribution of top-10 countries in our dataset.

4.1.1 Features

Once the data from the above sources has been collected, one need to extract a proper set of features in order to define a space where the prediction task can be done efficiently. Below, we describe four categories of features that

[2] `https://developer.twitter.com`
[3] `https://developers.google.com/`
[4] `https://www.infogreffe.com`
[5] The complete list can be found in the Appendix 6.1.

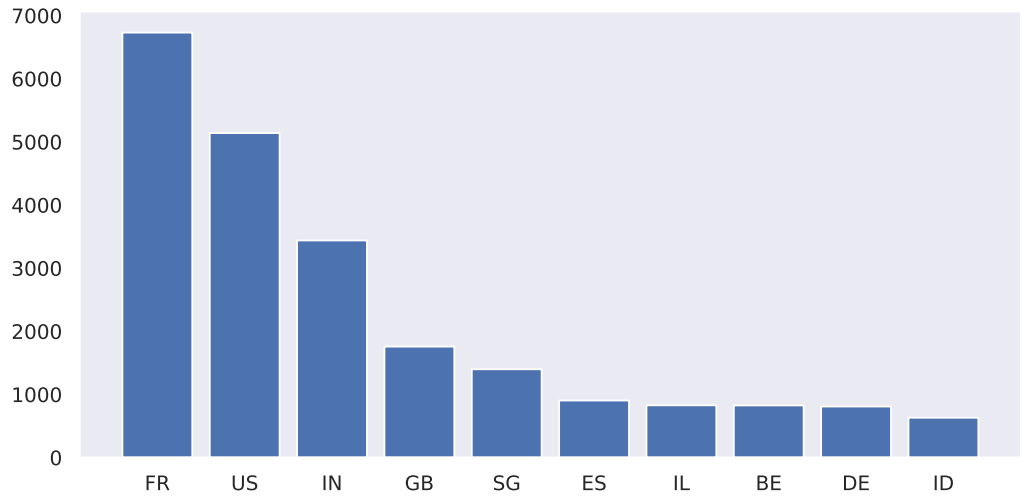


Figure 4.1: Geographical distribution of the top-10 countries in our dataset.

we extracted, along with the intuitions behind, from the web for the purpose of startup success prediction.

General features

The following features are considered as the core information about startups:

- Country of origin,
- Age,
- Number of employees,
- Number of offices,
- Number of people featured on Team page of startup's website.

Importance of these features for the task of fundraising prediction is quite obvious: venture's evolution in different countries varies. Age as well as the number of employees and offices characterize different stages of startup evolution and the properties of fundraising process strongly depend on the stage of the venture.

The country of the startup's origin is extracted from the address pages on the company website. We use statistical methods to infer which country is the most likely the country of the company. To do that, we employ regular expressions to extract phone numbers (via country codes), then simply look around the phone number, in a fixed window size, to find the country. The country with the most occurrence is then taken as the country of origin and (in case of ties no country is selected).

To infer the age of the startup, we simply use its creation date. Most countries give public access to a registry of all companies, in which one can usually find the creation date. Another heuristic we use is to infer the creation date from the dates of the creation of different media from the company: website and the social media. In case of different createion dates identified by the two previously mentioned sources, the older date is taken as the creation date. According to our observations, in the context of startups, this is in most cases a very good approximation of creation date: in 28% of cases, it return the correct creation date and in 72% of cases, it returns a creation date with maximum of two years shift.

The number of people featured on Team page of startup's website is extracted as follows: usually the team page follows a repeating template containing information about every person (name, role, social media links, picture, etc). We find and extract this repeating template and then use statistical methods to verify that it corresponds to people names, job functions, etc. Finally, information about the number of employees and offices is extracted from the country-specific databases.[5]

Financial features

History of startup's previous funding rounds is evidently an important factor for predicting future fundraising as different fundraising rounds happen usually with similar patents w.r.t. the previous rounds secured by the startup. The process of detecting funding events for startups is described in Section 4.1.2. In this work, we propose to extract the following features to summarize the financial history of a startup:

Table 4.1: Startup features used in this study.

Group	Name	Description
General	Country	Country of a startup's origin
	Age	Age of a startup
	Number of employees	Official number of employees
	Number of offices	Official number of offices
	People on team page	Number of people featured on Team page of a startup's website
Financial	N previous rounds	Total number of previously secured funding rounds
	Last fundraising amount	The amount of money secured in the last funding round
	Time since fundraising	Days since the last detected funding round
	Mean fundraising amount	Mean size of previously secured funding rounds
	Maximal fundraising amount	Size of the biggest previous funding round
Social Networks	Social media accounts	Does a startup have an account in Facebook/Linkedin/Instagram/Youtube or blog on its own website?
	Twitter statistics	Number of tweets and mean/max likes and retweets obtained for each month during the last year
	Twitter lingual statistics	Modal language of user tweets in each month during the last year
	Twitter hashtags	How many times a startup used each of the most popular 500 hashtags among startups during the last year
Web presence	Number of relevant results	Number of pages relevant to the startup in the first 10 result from Google search
	Total results	Total number of results reported by Google
	Domains	Number of results from each of 500 most popular domains (only top 10 Google search results are analyzed)

- Total number of previously secured funding rounds,
- Last fundraising amount,
- Mean and maximal amount of previously secured rounds,
- Time since the last secured round.

Google search results features

In (Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke, 2018), the authors have shown that a highly useful set of features for the task of the startup success prediction can be extracted from crawling of the observable web for the startup presence. For the purpose of extraction of these features, the authors analyze the data from Yandex[6], a major Russian search engine.

[6] `https://yandex.ru`

For each startup, they count the number of references to the startup's website on the webpages from different domains. This data, however, is not easily accessible by ordinary web users.

Accordingly, in the present study, similar information using more widely available tools has been extracted, in particular Google search API. For each startup, search results with a date within a year preceding the start of *prediction period* have been analyzed. Given a startup name and a date range, a query to Google API is made and irrelevant results were filtered in order to perform the analysis of domains frequencies similar to (Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke, 2018). In order to exclude irrelevant results, we check whether the snippet of a search result contains the startup's name or not. Since the purpose of this work is to build models with entirely free tools, we constrain ourselves to the amount of queries available with the Google Cloud Platform Free Tier [7]. Therefore, we obtain only top 10 results for a given startup name and a date range. The following statistics are then extracted form these search results:

- Number of relevant results: we assume that result is related to startup only if a snippet of result contains the startup's name,
- Total number of results as reported by Google,
- Number of search results from each of 500 domains *popular* domains.

For the latter, we simply sort all the domains appearing in the results based on the number of times they contain the name of startups under investigation. We then take the top 500 domains. The intuition is to take the domains which are more likely to talk about startups and, as a result, reduce the amount of noise in our feature space.

Social networks presence features

Over the last two decades, the impact of social networks on different social, economic and political processes became remarkable. Given the fact that for a startup it is crucial to reach the potential audience via social media, this

[7] `https://cloud.google.com/free/docs/gcp-free-tier`

category of features can heavily impact the prediction performance. One can note such impacts in the investigations done in the literature such as (Zhang, Ye, Essaidi, Agarwal, Liu, and Loo, 2017) which highlighted the importance of social media presence for a crowdfunding success of a startup. We use here, first, a set of social media features that are binary and indicate whether a startup has an account in several popular social media:

- Facebook,
- Instagram,
- Linkedin,
- YouTube,
- Twitter,
- Blog on the startup own website.

This information is extracted with a simple script that searches for social network buttons on a website of a startup. The second set of features extracted from social media corresponds to statistical information of a startup's website:

- Number of people that give reference to their Linkedin account on the team page of the startup,
- Number of entries in blog during the last year.

These two features indicate the willingness of the startup to appear in the social media and to be visible and followed by others.

Because of the important presence of startups on Twitter and since information from Twitter is readily available (contrary to other social media like Facebook and LinkeIn), we also extract features that describe the activity of each startup on this particular social media in the year the precedes the year for which funding events are predicted:

- Aggregated monthly startup's Twitter account statistics for the last year including: number of posted tweets, mean/max number of likes and retweets of user's tweets; modal language of user tweets,
- total number of different users that mention startup's account in their tweets during the last year,
- Information about *hashtags* used by the startup: *a)* for all startups, we collect their last 2300 tweets from which we establish a list of the 500 most frequently used hastags; each startup is then represented as a 500-dimensional vector the dimensions of which correspond to the number of times the startup used the hashtag during the last year.

Table 4.1 summarizes the features explained above. The type of each feature (categorical or numerical) as well as its nature (sparse or dense) are also illustrated in the two rightmost columns.

4.1.2 Data labeling

Another challenge in solving the task of predicting startup success from open sources is labeling the data. While commercial databases often contain dates of funding events where amounts are usually extracted manually by human experts, we aim to automatically detect startups fundraising from news and Twitter. For this purpose, we subscribed to RSS feeds for a large set of news websites focused on startups.[8]

For each sentence in tweets and news headlines, we proceed as follows:

1. identify money amounts using regular expressions (ex: $5M).
2. identify mentions of known startup using either startup names or Twitter screen names.

[8]Some examples are given in Appendix 6.2. The exhaustive list countains several thousands of entries.

Figure 4.2: Illustration of the funding extraction algorithm: detected money amount and startup name are shown in boxes; fundraising verb is underlined.

3. retain as a candidate funding event the startup mention and money amount if they are separate by a fundraising verb[9].

4. merge candidate funding events if they occur within a three-month period as given by the tweet or news date.

An example of this approach is illustrated in Figure 4.2 where the verb is underlined in purple, and startup name and the amount are shown in orange and green boxes respectively.

This algorithm, despite its simplicity, is able to extract a surprisingly large amount of funding events. Indeed, we were able to detect 9139 funding events that took place during the last 5 years corresponding to an average of about 7% startups securing a funding event each year.

This said, the above algorithm may falsely assign funding events to startups. For example, in particular when there are several startups with similar names. In order to estimate the error rate in the detected funding events, we sampled 200 funding events detected by our algorithm and manually checked if they were correct or not. We found 17 false funding events, *i.e.* meaning that the rate of false positives amounts to 8.5%.

[9]The complete list: raise, take, get, grab, score, secure, receive, close, announce, complete

Another problem is that not all funding events are identified as funding events can be reported in different ways. In addition, it is worth noticing that information about some of the funding rounds is held private and, therefore, cannot be extracted. In order to estimate the number of funding events not identified by our algorithm, we randomly sampled 200 startups for which we did not find a funding event in a chosen one-year time period and found that 12 of them (*i.e.* 6%) actually raised money.

4.2 Prediction models

In this section, we describe the machine learning frameworks we explored to predict new funding events.

4.2.1 Positive-Unlabeled setting

The procedure of data labeling that we proposed in 4.1.2 does not include assigning negative labels to the startups. Thus, it might be natural to consider our problem in the context of Positive-Unlabeled (PU) learningBekker and Davis, 2018.

PU learning is a field of machine learning that studies the principles of training a binary classifier in a case when some positive objects are labeled as positive while the majority of objects does not have any label and consists of objects of both classes. The PU learning methods can generally be divided into two categories. The first one aims to find true negatives in the data and then perform normal binary classification (Liu, Lee, Yu, and Li, 2002) (Li and Liu, 2003). The second category aims to modify the loss function to account for the PU setting. (Elkan and Noto, 2008) first proposed to treat unlabeled data as weighted positive and negative data mixture. (Du Plessis, Niu, and Sugiyama, 2014) introduced the approach described below.

Let $X \in \mathbb{R}^d$ and $Y \in \{\pm 1\} (d \in \mathbb{N})$ denote the input and output random variables, $p_p(x) = p(x|Y = +1)$ and $p_n(x) = p(x|Y = -1)$ be the positive and

negative marginals, $\pi_p = p(Y = +1)$ and $\pi_n = p(Y = -1) = 1 - \pi_p$ stand for class-prior probabilities.

For decision function $g : \mathbb{R}^d \to \mathbb{R}$ the empirical risk $\hat{R}$ is minimized, which can be decomposed into empirical risk on positive and negative examples:

$$\hat{R} = \pi_n \hat{R}_n^-(g) + \pi_p \hat{R}_p^+(g) \tag{4.1}$$

Given the loss function $l : \mathbb{R} \times \{\pm 1\} \to \mathbb{R}$ one can estimate:

$$\hat{R}_p^+ = (1/n_p) \sum_{i=0}^{n_p} l(g(x_i^p), +1) \tag{4.2}$$

The difficulty of PU learning is the lack of data sampled from $p_n(x) = p(x|Y = -1)$, therefore it is not possible to use estimate $\hat{R}_n^- = (1/n_n) \sum_{i=0}^{n_n} l(g(x_i^n), -1)$. However, the following approximation can be used:

$$\begin{aligned} p(x) &= \pi_n p_n(x) + \pi_p p_p(x) && (4.3) \\ \pi_n p_n(x) &= p(x) - \pi_p p_p(x) && (4.4) \end{aligned}$$

$$\pi_n \hat{R}_n^- = \hat{R}_U^- - \pi_p \hat{R}_p^- = (1/n_u) \sum_{i=1}^{n_u} l(g(x_i^u), -1) - \pi_p (1/n_p) \sum_{i=1}^{n_p} l(g(x_i^p), -1) \tag{4.5}$$

$$\begin{aligned} \hat{R}(g) = \pi_p \hat{R}_p^+ + \hat{R}_U^- - \pi_p \hat{R}_p^- = \pi_p (1/n_p) \sum_{i=1}^{n_p} l(g(x_i^p), +1) + \\ (1/n_u) \sum_{i=1}^{n_u} l(g(x_i^u), -1) - \pi_p (1/n_p) \sum_{i=1}^{n_p} l(g(x_i^p), -1) \end{aligned} \tag{4.6}$$

In (Kiryo, Niu, Plessis, and Sugiyama, 2017) the authors propose to modify the loss to make it more suitable to neural networks. They note that the theoretical guaranties obtained for the method in (Du Plessis, Niu, and Sugiyama, 2014) need an assumption on the Rademacher complexity (Mohri, Rostamizadeh, and Talwalkar, 2012), which does not hold for the neural networks. At the same time they show that in practice neural networks overfit while trying to make the term $\pi_p(1/n_p)\sum_{i=1}^{n_p} l(g(x_i^p), -1)$ as large as possible to get negative loss. They propose simple modification of loss to mitigate the problem:

$$\hat{R}(g) = \pi_p \hat{R}_p^+ + \max(\hat{R}_U^- - \pi_p \hat{R}_p^-, 0) \tag{4.7}$$

In our work we apply this modification called Positive-Unlabeled Learning with Non-Negative Risk Estimator(nnPU) in order to see whether it is beneficial compared to the standard binary classification. To this end we construct a multilayer perceptron and train it on our dataset *a)* with the standard sigmoid loss and *b)* with the loss proposed in (Kiryo, Niu, Plessis, and Sugiyama, 2017).

Another approach to PU learning that we applied to our data is Positive Unlabeled AUC Optimization (Sakai, Niu, and Sugiyama, 2017). In the standard binary classification AUC for a classifier g is defined as:

$$AUC(g) = \mathbb{E}_p[\mathbb{E}_n[I(g(x^P) \geq g(x^N))]] \tag{4.8}$$

where $\mathbb{E}_p$ and $\mathbb{E}_n$ are the expectations over $p_p(x)$ and $p_n(x)$, respectively. $I(\mathring{u})$ stands for the indicator function.

In practice a composite classifier $f(x, x') = g(x) - g(x')$ can be trained by minimizing the empirical AUC risk (Herschtal and Raskutti, 2004) (Davis and Goadrich, 2006) defined as:

$$\hat{R}(f) = \frac{1}{n_P n_N} \sum_{i=1}^{n_P} \sum_{j=1}^{n_N} l(f(x_i^P, x_j^N)) \tag{4.9}$$

where $l(m)$ is a surrogate loss.

Using formula 4.3 in Sakai, Niu, and Sugiyama, 2017 the authors propose the following expression for AUC risk in PU setting:

$$\hat{R}_{PU}(f) = \frac{1}{\pi_n n_P n_U} \sum_{i=1}^{n_P} \sum_{j=1}^{n_U} l(f(x_i^P, x_j^U)) - \frac{1}{\pi_n n_P (n_P - 1)} \sum_{i=1}^{n_P} \sum_{i'=1}^{n_P} l(f(x_i^P, x_{i'}^P)) + \frac{\pi_p}{\pi_n (n_p - 1)} \quad (4.10)$$

and further construct a kernel based method that efficiently optimizes the given PU risk. In our work we use the authors original implementation of this method in python available on GitHub [10] to solve the task of startup success prediction.

4.2.2 Positive-Negative setting

Despite the arguments in favor of PU setting for our task, there are also several factors that incline us to give preference to the traditional binary classification setup, referred to as PN for Positive-Negative(PN), which consists in treating the startups for which we could not detect funding events as negative examples. The most important one is that currently the PU setting is rather restrictive in terms of available algorithms. The proposed modifications of the loss function make the loss either non-convex (Du Plessis, Niu, and Sugiyama, 2014) (Kiryo, Niu, Plessis, and Sugiyama, 2017) or non smooth (Du Plessis, Niu, and Sugiyama, 2015) which impedes their use with the algorithms relying on the second order optimization techniques. Another point is that the theoretical results for the PU learning are obtained with assumption that the objects labeled as positives are always positive. At the same time in our dataset, as discussed in 4.1.2 the share of falsely assigned funding events is estimated to be around 9%. The last consideration is that due to the class imbalance in our dataset, treating all the unlabeled objects as negative might be less harmful than in class-balanced PU setting.

[10] `https://github.com/t-sakai-kure/pywsl`

For the given reasons, we performed a series of experiments in traditional binary classification settings. We tested the performance of the most widely used machine learning models such as Logistic Regression, Random Forest, and a recent gradient boosting algorithm with support of categorical variables called CatBoost (Prokhorenkova, Gusev, Vorobev, Dorogush, and Gulin, 2018). In the preliminary experiments, we also compared different popular gradient boosting algorithms CatBoost (Prokhorenkova, Gusev, Vorobev, Dorogush, and Gulin, 2018), XGBoost (Chen and Guestrin, 2016b) and LightGBM (Ke, Meng, Finley, Wang, Chen, Ma, Ye, and Liu, 2017) and found that they yield similar performance on our dataset. Therefore, we only report the results obtained with CatBoost. As discussed in 4.2.1 in order to study the impact of nnPU loss modification we also trained a neural network in traditional binary classification setting. We also did our best to reproduce the approach WBSSP introduced in Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke, 2018. Since this algorithm is very specific to the dataset for which it was developed *i.e.* it specifies which exact feature should go to which logistic regression group, it is impossible to exactly reproduce it for a dataset with a very different set of features. However, we followed the general idea and built logistic regression models on semantic groups of features, and then built a CatBoost model using logistic regressions as features in addition to non-sparse initial features of the dataset. The details for about semantic groups of features as well as the information about sparsity can be found in table 4.1.

4.3 Evaluation

In this section, we detail the experiments conducted to assess the effectiveness of the proposed model. We also investigate the effectiveness of the PU framework (see Section 4.2), in addition to an extensive discussion on the importance of (set of) features.

4.3.1 Data split and metrics

Our algorithm of populating train and test sets is identical to the one described in Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke, 2018: we design a function that, given a name of a startup and a date d, extracts the feature vector X of the startup using only the information available before the date d (*e.g.* previous funding events, earlier activity on social networks etc.). Another function, given a name of a startup and a date d, returns the binary label – whether a funding event was detected for the startup in a year since the date d. For $d \in$ {01-09-2014, 01-09-2015, 01-09-2016, 01-09-2017 }, we extract (X, y) pairs for each startup in the list and populate train set. For $d =$ 01-09-2018 the extracted pairs go to test set.

To evaluate our results, we use the area under the ROC curve (AUC) (Fawcett, 2006), which illustrates the behavior of the prediction w.r.t. True Positive Rate (TPR) and False Positive Rate (FPR) at different points, and has been used in a vast variety of tasks to assess the classification performance (Hand and Till, 2001; Korolev, Safiullin, Belyaev, and Dodonova, 2017). It furthermore can properly assess the effectiveness of classification models in the presence of noisy labels as well as in the case of imbalanced classes (Zhang, Wu, and Shengs, 2014; He and Garcia, 2009). On top of that, we adopt the same strategy as Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke, 2018 and used the F-score with $\beta = 0.1$ in order to have a more significant impact of precision and, more importantly, to be able to compare with the results presented in Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke, 2018. We also assess the performance of the models via precision on the top 100 ($P@100$) and on the top 200 ($P@200$) results.

As discussed in Section 4.1.2, label extracted by our method are sometimes incorrect. In (Jain, White, and Radivojac, 2017), the authors proposed a way to estimate binary classifier performance in PU setting. They theoretically show, that if f_1 and f_2 are the distributions of the positive and negative objects, the unlabeled examples are drawn from the distribution $f(x) = \alpha f_1(x) + (1 - \alpha) f_0(x)$ and the labeled examples come from the distribution $g(x) = \beta f_1(x) + (1 - \beta) f_0(x)$ the true AUC of a classifier can be recovered from the obtained on the noisy labels AUC^{pu} with the following formula:

$$\text{AUC} = \frac{\text{AUC}^{pu} - \frac{1-(\beta-\alpha)}{2}}{\beta - \alpha} \tag{4.11}$$

where α and β are related to the amount of noise in the labels. As described in Section 4.1.2, for our dataset, we experimentally estimate $\alpha = 0.06$, since 6% of the startups, for which we did not identify funding events, *i.e.* unlabeled startups actually received money, and $\beta = 1 - 0.085 = 0.915$, since there exists 8.5% of falsely detected funding events. Therefore, for all our experiments, we report both raw value of AUC and the AUC value corrected according to the Equation (4.11).

4.3.2 Positive-Unlabeled results

As discussed in Section 4.2.1, we conducted some experiments in order to find out whether the PU learning can be beneficial compared to the traditional binary classification setting or not. To this end, we constructed a multilayer perceptron (MLP) model with two hidden layers of size 100 and had it trained with learning rate 0.0001 and batch size of 1000 for 5 epochs. As mentioned in Section 4.2.1, we first run the network with a standard sigmoid (σ) loss function, and then with the PU modified loss function (Non-Negative Risk Estimator) that has been presented in Kiryo, Niu, Plessis, and Sugiyama, 2017, a method we refer to as nnPU. Finally we investigated the kernel-based approach explained in Sakai, Niu, and Sugiyama, 2017, referred to as PU-AUC hereafter.

Table 4.2 illustrates the results of these experiments. The upper part shows the two neural network methods while the last line is separated from the rest as the objective function is different. The first observation one can make is that among the neural network methods, the one based on PU yields better results. Indeed, according to a Wilcoxon rank test, it is significantly better with $p < 0.05$ for $P@200$ and $F@200$, and with $p < 0.01$ for $ROC - AUC$. In addition, the best results are obtained with PU-AUC that significantly outperforms all the other methods on all metrics at $p < 0.01$. It is however hard to say whether the difference should be attributed to the modification of the loss function for imbalanced dataset or to the different nature of the

Table 4.2: PU setting. The upper part of the table illustrated the neural network approaches and the lower part shows the direct optimization of AUC from (Sakai, Niu, and Sugiyama, 2017). Neural networks are trained with 10 random seeds, mean and std. reported.

Loss function	P@100	$F_{0.1}$@100	P@200	$F_{0.1}$@200	ROC-AUC raw/corrected
Standard σ	0.26(0.02)	0.22(0.01)	0.24(0.01)	0.22(0.01)	0.78(0.01) / 0.83(0.01)
nnPU	0.27(0.01)	0.22(0.01)	0.25(0.01)	0.23(0.01)	0.79(0.01) / 0.84(0.01)
PU-AUC	**0.45**	**0.38**	**0.43**	**0.39**	**0.82/0.87**

base algorithms. All in all, these results show that PU based approaches yield better results than a simple PN method as MLP (or even Logistic Regression as illustrated below).

4.3.3 Positive-Negative results

For traditional binary classification setting, we use implementations of Logistic Regression(LR) and Random Forest(RF) from Scikit-learn (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot, and Duchesnay, 2011) and the recent gradient boosted tree method CatBoost (Prokhorenkova, Gusev, Vorobev, Dorogush, and Gulin, 2018). For all the methods, we perform 5-fold cross validation of hyperparameters on the train set. For the CatBoost, we found that using rather a high value of the coefficient at the L_2-regularization term of the cost function 100 helps to mitigate the overfitting problem.

To implement our version of WBSSP (Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke, 2018), we used the same Logistic Regression and CatBoost implementations. For each group illustrated in Table 4.1, we built a Logistic Regression; the training set is split into 5 folds, and out-of-fold predictions are used for the downstream classifier. At the same time, features that do not have sparse flag were directly fed into the final CatBoost classifier.

Table 4.3 shows the results of the three employed ensemble methods as well as the WBSSP-based approach, *i.e.* the one explained above. As mentioned in Section 4.2.2, this approach is particularly designed for the dataset presented

Table 4.3: PN setting models comparison. The best values are shown in bold. Values shown in parenthesis and marked with * are calculated on the labels corrected by the human experts for the top 200 companies.

Model	P@100	$F_{0.1}$@100	P@200	$F_{0.1}$@200	ROC-AUC raw	ROC-AUC corrected
LR	0.380	0.317	0.345	0.315	0.774	0.821
RF	**0.580**	**0.483**	0.470	0.429	0.796	0.847
CatBoost	0.53 (0.640*)	0.442(0.531*)	**0.480 (0.580*)**	**0.439(0.528*)**	**0.834**	**0.890**
WBSSP-based	0.52	0.433	0.470	0.429	0.825	0.881

in (Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke, 2018) and, as a result, it cannot be compared directly to our approach. However, it is the closest pipeline to that of (Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke, 2018) which adapts to our dataset.

As it can be seen in Table 4.3, Logistic Regression totally fails to provide good results with respect to the two other ensemble methods, *i.e.* RF and CatBoost. That can be simply explained by the fact that Logistic Regression is not able to predict the loss from the designed features which are difficult to be separated linearly. This basically comes form the complexity and the heterogeneous nature of features explained in Section 4.1.

Ensemble methods, however, are able to overcome this complexity and stress on important features in order to linearly separate subregions of the space and combine them, via weak learners, in order to perform better predictions. If we compare however RF and CatBoost, we can see that differences between top-100/200 precision and $F_{0.1}$ scores are small. CatBoost is nevertheless significantly better than RF in terms of ROC-AUC. This can be explained by the fact that CatBoost benefits from the gradient boosting framework and is able to perform optimization in functional space. On top of that, compared to the RF, it behaves in a more robust way in dealing with categorical and heterogeneous features, as it benefits from history-based ordered target statistics Prokhorenkova, Gusev, Vorobev, Dorogush, and Gulin, 2018. All in all, CatBoost is the best performing model over all PN and PN methods that we investigated.

After selecting the best model, based on its significantly better ROC-AUC, we set out to obtain the estimate of its performance not contaminated by the

Table 4.4: Comparison to the results reported in (Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke, 2018) on the dataset presented therein with WBSSP pipeline. Values marked with * are calculated on the labels corrected by the human experts for the top 200 companies.

Model	P@100	$F_{0.1}$@100	P@200	$F_{0.1}$@200	ROC-AUC
Our best model	0.640*	0.531*	0.580*	0.528*	0.890 (corrected)
WBSSP	0.626	0.383	0.535	0.439	0.854

noise in our test set labels. To this end, we took the list of top-200 companies according to our model and asked a human expert to manually check label for each startup. The metrics calculated with the corrected labels are shown in parentheses in Table 4.3.

As our last analysis on these experiments, we illustrate in Table 4.4 the best results of Table 4.3, *i.e.* CatBoost, along with those reported in Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke, 2018. Although there is no direct way to compare metrics obtained on two completely different datasets, the categories of features are rather similar and, accordingly, can provide a good insight into these two approaches. In terms of size and class balance the datasets are also comparable: 15128 objects with 8% of positives in (Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke, 2018) dataset vs. 33165 objects with 6.3% of positives in our dataset. As it can be seen from the table, the results reported by the same metrics illustrate that our classifier scores are consistently higher than the ones reported in (Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke, 2018).

4.3.4 Ablation Analysis

In studies like the present one, the diversity of possible numerical and categorical feature makes it sometimes difficult to have deep insights on the prediction models. Accordingly, a feature importance analysis is usually crucial in order to better understand and analyze the model. For this reason, we performed an ablation analysis aiming at assessing the importance of the different feature groups.

Table 4.5: Ablation Analysis

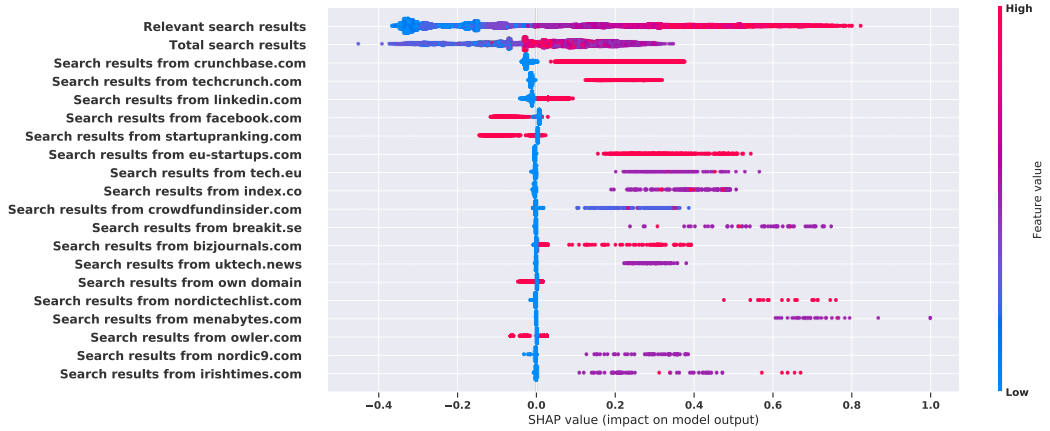
Features	P@100	$F_{0.1}$@100	P@200	$F_{0.1}$@200	ROC-AUC raw/corrected
All features	0.530	0.442	0.480	0.439	0.834/0.890
No Financial	0.530	0.442	0.440	0.402	0.819/0.873
No Social Net.	0.480	0.400	0.405	0.370	0.820/0.873
No Web presence	0.500	0.417	0.475	0.434	0.808/0.861

To this end, we repetitively exclude one semantic groups of features, presented in Table 4.1 in Section 4.1, and measure how it affects the model performance. Evidently, we keep always the General category as it contains the core information needed for the model. Table 4.5 presents the results of this analysis. The fist observation one can make from this table is that including all the features provides the best performance regardless of the metric. This is an important point as it indicates that all the categories presented in Section 4.1 are involved in boosting the performance of the prediction task. The second point is that social network information plays an important role in boosting the performance as removing it brings the most important deterioration to all metrics but ROC-AUC. When it comes to overall performance, measured by ROC-AUC, the removal of each semantic group negatively impacts the performance, again indicating that all features are important to predict funding events.

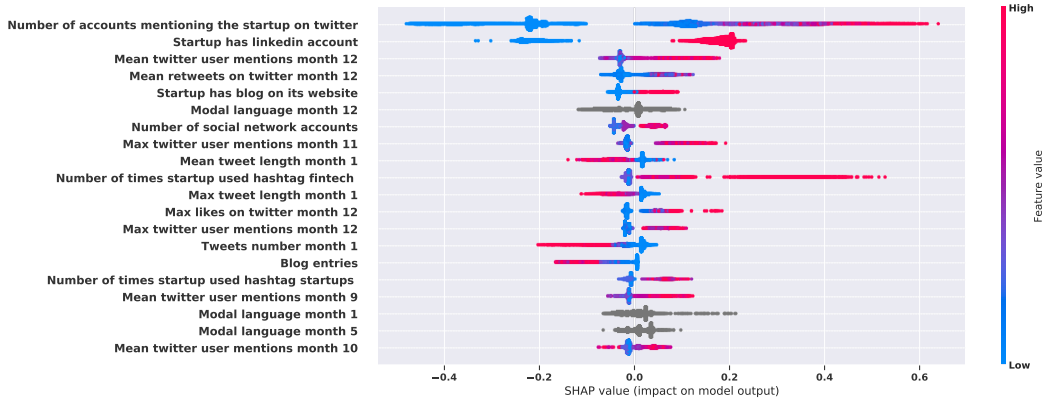
4.3.5 Feature Importance

To complete the analysis of the features retained, we make use SHAP values that were discussed in Section 3.3.2. SHAP provides, for each feature and each example, a measure of the impact of the feature on the decision that a model makes on the example. The calculation of this impact is based on the comparison of classifier's output on a full feature vector and the expectation of the classifier's output over feature vectors with the studied feature value replaced by all the possible values of the feature. SHAP plots are then constructed, where the x-axis corresponds to SHAP impact values and the y-axis to the different features. A dot on the figure finally corresponds to an example for which the corresponding feature (y-axis) has the SHAP impact value given in the x-axis. Note that on the y-axis, features are sorted

(a) Financial features



(b) Web presence



(c) Social media activity

Figure 4.3: Feature importance analysis using SHAP-values for the different semantic groups of features (a) financial, (b) the web presence and (c) social media activities. Features are stacked vertically based on their importance from top to bottom of each figure. Each dot represent an instance in the dataset with its corresponding SHAP value on X-axis.

according to their importance, the topmost feature being the most important one. The importance of a feature is measured by given by the sum, over all examples, of the absolute values of the SHAP impact scores.

Figure 4.3 displays the SHAP plots we obtained for the different semantic groups. Note that all the features are displayed for the financial group, whereas only the 20 most important features are displayed for the web-

presence and the social network groups. Several conclusions can be drawn from Figure 4.3. We present here the most obvious ones. First, the amount of the pages mentioning the startup among the first 10 results provided by a search engine for a query that contains the startup's name is an important feature for predicting the future fundraisings. The same can be said about the number of different people that mention the startup on Twitter. Second, if the pages on LinkedIn or Crunchbase returned by search engine are a positive indicator for future fundings, the pages returned from Facebook or startupranking.com are a negative indicator. Third, not only users mentioning a startup on twitter are a positive sign for the future funding rounds, but also a startup mentioning other users is.

4.4 Conclusion

We have studied in this chapter the problem of predicting funding events for startups. To do so, and contrary to previous studies that have used information from commercial databases, we have solely relied on information that can be extracted from freely, publicly available sources as startup websites, social media and company registries. The features we rely on can be easily obtained from these sources. Furthermore, the prediction models we use are simple and wide-spread; ensemble methods in a standard positive-negative setting indeed yield the best prediction results. Despite these constraints (easily obtained features, simple prediction models), that guarantee that our methods can be re-implemented, the results we obtain are better than the ones obtained with more information sources and more complex models.

Several aspects of our work can nevertheless be improved. One possibility for the future research is to further explore the importance of each feature and explain why certain features (as the presence on specific social networks) are negative indicators for predicting funding events. Second direction of research would be to use parsing techniques to limit the amount of false positives when detecting funding events in tweets and news headlines. Also using word embeddings to identify additional *fundraising* verbs, and more generally operators, in order to increase the recall (yet, the most important problem to solve is the one related to false positives) is an interesting di-

rection. In addition, several sources of information, also readily available, could be envisaged to complement the features we have considered. Patent databases could for example be mined in order to get indicators of the invention portfolio of a startup, an element that is taken into account by many investors. Publicly available information from investment companies, from which technological domain and market information can be inferred, could also be used to further predict which investor is likely to be interested in which startup. The space of potential use cases is large and we hope that the current work will pave the way to new studies on startup analysis.

5 Conclusion

This dissertation has studied the problem of startup valuation and fundraising. We mine different data sources, including the ones novel to the startup success research. Further, we apply various artificial intelligence methods such as machine learning, domain adaptation, and causal discovery. Our insights into the factors that affect startup valuation and the features that allow predicting it provide a valuable perspective for researchers and practitioners.

5.1 Summary of contributions

In Chapter 3 we study startup valuation from several complementary angles. First, we investigate whether the publicly available data about startup valuations provided by the fundraising participants are representative with respect to the general population of startup valuations in funding rounds. Having identified a significant distribution shift, we proposed various domain adaptation techniques to help the generalization of machine learning models. Second, we collected a rich dataset for the startup valuation prediction task and studied with various explainable artificial intelligence and causal discovery methods which variables allow predicting valuation and which variables affect valuation. We also provide a comparison of the two sets of variables. To the best of our knowledge, such an approach is unique in literature. The methods to overcome the reported-unreported valuations distributional discrepancy, as well as insights into the startup valuation factors and predictors, are among our contributions.

In Chapter 4 we explore the possibility of startup success prediction in the situation when proprietary databases are not available. Collecting information about the startups and their funding rounds from the web is the first problem that we tackle. The second problem that we address is that it is impossible

to collect the information about all the funding round. To resolve this issue, we propose to learn a model in a positive-unlabeled rather than a traditional binary classification setting. Our best-performing model shows performance on par with the models trained on large proprietary datasets proposed in the literature.

5.2 Future Work

Further directions for the startup valuation prediction studies could be the following. First, our approach for predicting startup valuation in a funding round via domain adaptation is limited to the European region due to data availability. Extracting startup valuation in funding round information for the startups from other regions, e.g., USA or China, would allow one to propose a universal valuation prediction model. Further, developing domain adaptation methods suitable for gradient-boosted trees regression models would also be valuable. Finally, a study of startup valuation change in time would be of great interest.

In the context of startup valuation predictors and factors, an exciting continuation of our research would be exploring the boundary conditions, such as pre-revenue and post-revenue startups. Also, replication studies are encouraged to test the model in other regions and countries, such as the United States or China, since the fundraising patterns in other regions may not follow those of the UK or Europe. Exploration of other causal discovery methods, in particular, the ones that do not rely on the unconfoundedness assumption, such as Fast Causal Inference (Spirtes, Glymour, and Scheines, 2000; Zhang, 2008) is also left for the future research.

An interesting development of the fundraising prediction studies would be the development of a positive-unlabeled learning method compatible with gradient-boosted tree models. Indeed, in our experiments, we demonstrate that positive-unlabeled learning improves results for neural networks compared to the standard positive-negative setting. However, our best-performing gradient-boosted trees model cannot be trained in a positive-unlabeled setting in the same manner. Another interesting perspective would be incorporating

the information about the startup position in the business entities network and applying graph learning methods in combination with a rich feature set designed with domain expertise.

6 Appendix

6.1 Business register sources

- https://businessregister.kompany.com
- https://portal.kyckr.com/
- https://data.opendatasoft.com/explore/dataset/sirene%40public/api/
- https://www.societe.com/
- https://www.infogreffe.fr/
- https://developer.companieshouse.gov.uk/api/docs/
- https://beta.companieshouse.gov.uk/
- https://eng.kurzy.cz/prodej-dat/databaze-firmy.htm
- https://www.ytj.fi/en/index/whatisbis/opendata.html
- https://www.ytj.fi/en/
- http://avoindata.prh.fi/tr_en.html#/
- http://kbopub.economie.fgov.be/kbopub/zoeknaamfonetischform.html
- https://kbopub.economie.fgov.be/kbo-open-data/login?lang=fr
- https://search.cro.ie/company/CompanySearch.aspx

- https://services.cro.ie/overview.aspx
- https://datacvr.virk.dk
- https://www.brreg.no/home/
- https://www.unternehmensregister.de/ureg/?submitaction=language&language=en
- https://www.registroimprese.it/en/
- http://www.rmc.es/Sociedades.aspx
- http://www.infocif.es/
- http://www.fi.se/en/our-registers/company-register/
- https://www.zefix.ch/fr
- https://firmenbuch.at/
- https://www.rcsl.lu/mjrcs/
- https://companies-register.companiesoffice.govt.nz/
- https://ica.justice.gov.il/GenericCorporarionInfo/SearchCorporation?unit=8
- https://beta.registresdentreprisesaucanada.ca/chercher
- https://www.registreentreprises.gouv.qc.ca/RQAnonymeGR/GR/GR03/GR03A2_19A_PIU_RechEnt_PC/PageRechSimple.aspx
- https://www.sec.gov/edgar/searchedgar/companysearch.html
- http://developer.edgar-online.com/apps/mykeys
- https://icis.corp.delaware.gov/Ecorp/EntitySearch/NameSearch.aspx
- http://www.gsxt.gov.cn/index.html

- https://www.gov.ph/data/search/type/dataset
- https://data.gov.sg/
- http://www.ocr.gov.np/index.php/np/
- https://data.gov.in/catalog/company-master-data
- http://seninfogreffe.com/

6.2 Some examples of sources for data labeling

- https://500.co/feed/
- https://agfundernews.com/feed/
- http://www.arabianbusiness.com/feed/startup/feed.xml
- https://www.austrianstartups.com/feed/
- https://betakit.com/feed/
- https://bothsidesofthetable.com/feed
- https://www.businessweekly.co.uk/rss.xml
- https://www.cnbc.com/id/10001274/device/rss
- http://www.ecap-partner.com/news/feed/
- https://www.entrepreneur.com/latest.rss
- http://www.finsmes.com/feed
- https://www.frenchweb.fr/feed
- https://www.geekwire.com/startups/feed/

- https://iamanentrepreneur.in/feed/
- https://tech.economictimes.indiatimes.com/rss/startups
- http://knowstartup.com/feed/
- https://www.entrepreneur.com/latest.rss
- http://www.finsmes.com/feed
- https://www.frenchweb.fr/feed
- https://www.geekwire.com/startups/feed/
- https://iamanentrepreneur.in/feed/
- https://tech.economictimes.indiatimes.com/rss/startups
- http://knowstartup.com/feed/
- https://www.maddyness.com/feed
- http://feeds.mashable.com/Mashable
- https://medium.com/feed/startups-for-news

Bibliography

Aliferis, Constantin F., Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D. Koutsoukos (2010). "Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation". In: *J. Mach. Learn. Res.* 11, 171–234.

Amis, D. and H.H. Stevenson (2001). *Winning Angels: The Seven Fundamentals of Early-stage Investing*. Financial Times Prentice Hall. Financial Times Prentice Hall.

Andersson, Steen A., David Madigan, and Michael D. Perlman (1997). "A Characterization of Markov Equivalence Classes for Acyclic Digraphs". In: *The Annals of Statistics* 25.2, pp. 505–541.

Antretter, Torben, Ivo Blohm, Dietmar Grichnik, and Joakim Wincent (2019). "Predicting new venture survival: A Twitter-based machine learning approach to measuring online legitimacy". In: *Journal of Business Venturing Insights* 11, e00109.

Baum, Joel A.C. and Brian S. Silverman (2004). "Picking winners or building them? Alliance, intellectual, and human capital as selection criteria in venture financing and performance of biotechnology startups". In: *Journal of Business Venturing* 19.3, pp. 411–436.

Bekker, Jessa and Jesse Davis (2018). "Learning From Positive and Unlabeled Data: A Survey". In: *ArXiv* abs/1811.04820.

Beliaeva, Tatiana, Mariia Garkavenko, Hamid Mirisaee, Eric Gaussier, Agnès Guerraz, and Cédric Lagnier (2020). "Estimating startup valuation with AI: Evidence from green technology startups in Europe". In: *European Centre for Alternative Finance Research Conference*. Utrecht University.

Bengtson, Eric and Dan Roth (2008). "Understanding the value of features for coreference resolution". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 294–303.

Block, Joern, Christian Fisch, Silvio Vismara, and René Andres (2019). "Private equity investment criteria: An experimental conjoint analysis of venture capital, business angels, and family offices". In: *Journal of Corporate Finance* 58, pp. 329–352.

Blohm, Ivo, Torben Antretter, Charlotta Sirén, Dietmar Grichnik, and Joakim Wincent (2020). "It's a peoples game, isn't it?! A comparison between the investment returns of business angels and machine learning algorithms". In: *Entrepreneurship Theory and Practice*.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Carmona, Pedro, Francisco Climent, and Alexandre Momparler (2019). "Predicting failure in the US banking sector: An extreme gradient boosting approach". In: *International Review of Economics & Finance* 61, pp. 304–323.

Caruana, Rich and Alexandru Niculescu-Mizil (2006). "An empirical comparison of supervised learning algorithms". In: *Proceedings of the 23rd International Conference on Machine learning*, pp. 161–168.

Chang, Yung-Chia, Kuei-Hu Chang, and Guan-Jhih Wu (2018). "Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions". In: *Applied Soft Computing* 73, pp. 914–920.

Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer (2002). "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16, pp. 321–357.

Chen, Tianqi and Carlos Guestrin (2016a). "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

– (2016b). "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, pp. 785–794.

Chickering, David Maxwell (2003). "Optimal Structure Identification with Greedy Search". In: *J. Mach. Learn. Res.* 3.null, 507–554.

Climent, Francisco, Alexandre Momparler, and Pedro Carmona (2019). "Anticipating bank distress in the Eurozone: An extreme gradient boosting approach". In: *Journal of Business Research* 101, pp. 885–896.

Coad, Alex and Stjepan Srhoj (2020). "Catching Gazelles with a Lasso: Big data techniques for the prediction of high-growth firms". In: *Small Business Economics* 55.3, pp. 541–565.

Collis, Jill (2012). "Determinants of voluntary audit and voluntary full accounts in micro-and non-micro small companies in the UK". In: *Accounting and Business Research* 42.4, pp. 441–468.

Collis, Jill and Robin Jarvis (2002). “Financial information and the management of small private companies”. In: *Journal of Small Business and Enterprise Development* 9(2), pp. 100–110.

Colombo, Diego and Marloes H. Maathuis (2014). “Order-Independent Constraint-Based Causal Structure Learning”. In: *Journal of Machine Learning Research* 15.116, pp. 3921–3962.

Cooper, Arnold C., F. Javier Gimeno-Gascon, and Carolyn Y. Woo (1994). “Initial human and financial capital as predictors of new venture performance”. In: *Journal of Business Venturing* 9.5, pp. 371–395.

Covert, Ian, Scott Lundberg, and Su-In Lee (2020a). “Explaining by Removing: A Unified Framework for Model Explanation”. In: *arXiv preprint arXiv:2011.14878*.

Covert, Ian, Scott M Lundberg, and Su-In Lee (2020b). “Understanding global feature contributions with additive importance measures”. In: *Advances in Neural Information Processing Systems* 33, pp. 17212–17223.

Cox, D. R. (1972). “Regression Models and Life-Tables”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2, pp. 187–202.

Csaszar, Felipe, Miguel Nussbaum, and Marcos Sepulveda (2006). “Strategic and cognitive criteria for the selection of startups”. In: *Technovation* 26.2, pp. 151–161.

Damodaran, Aswath. (2001). *The dark side of valuation : Valuing old tech, new tech, and new economy companies*. eng. Upper Saddle River, NJ: Financial Times Prentice Hall.

Davis, Jesse and Mark Goadrich (2006). “The relationship between Precision-Recall and ROC curves”. In: *Proceedings of the 23rd international conference on Machine learning*. ACM, pp. 233–240.

Dhochak, Monika and Prince Doliya (2020). “Valuation of a startup: Moving towards strategic approaches”. In: *Journal of Multi-Criteria Decision Analysis* 27.1-2, pp. 39–49.

Du Plessis, Marthinus C, Gang Niu, and Masashi Sugiyama (2014). “Analysis of learning from positive and unlabeled data”. In: *Advances in neural information processing systems*, pp. 703–711.

Du Plessis, Marthinus Christoffel, Gang Niu, and Masashi Sugiyama (2015). “Convex Formulation for Learning from Positive and Unlabeled Data”. In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML’15. Lille, France: JMLR.org, 1386–1394.

Elkan, Charles (2001). “The foundations of cost-sensitive learning”. In: *International joint conference on artificial intelligence*. Vol. 17. 1. Lawrence Erlbaum Associates Ltd, pp. 973–978.

Elkan, Charles and Keith Noto (2008). “Learning classifiers from only positive and unlabeled data”. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 213–220.

Fawcett, Tom (2006). “An introduction to ROC analysis”. In: *Pattern recognition letters* 27.8, pp. 861–874.

Fernando, Basura, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars (2013). “Unsupervised Visual Domain Adaptation Using Subspace Alignment”. In: *2013 IEEE International Conference on Computer Vision*, pp. 2960–2967.

Ferrati, Francesco and Moreno Muffatto (2020). “Using Crunchbase for research in entrepreneurship: Data content and structure”. In: *Proceedings of the 20th European Conference on Research Methodology for Business and Management Studies: ECRM 2020*, p. 342.

Festel, G., Martin Würmseher, and G. Cattaneo (2013). “Valuation of Early Stage High-tech Start-up Companies”. In: *International journal of business* 18, pp. 216–231.

Florin, Juan, Michael Lubatkin, and William Schulze (2003). “A Social Capital Model of High-Growth Ventures”. In: *The Academy of Management Journal* 46.3, pp. 374–384.

Franke, Nikolaus, Marc Gruber, Dietmar Harhoff, and Joachim Henkel (2008). “Venture Capitalists’ Evaluations of Start–Up Teams: Trade–Offs, Knock–Out Criteria, and the Impact of VC Experience”. In: *Entrepreneurship Theory and Practice* 32.3, pp. 459–483.

Friedman, Jerome H (2001). “Greedy function approximation: A gradient boosting machine”. In: *Annals of statistics* 29.5, pp. 1189–1232.

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky (2016). “Domain-adversarial training of neural networks”. In: *The Journal of Machine Learning Research* 17.1, pp. 2096–2030.

Garkavenko, Mariia, Tatiana Beliaeva, Eric Gaussier, Hamid Mirisaee, Agnès Guerraz, and Cédric Lagnier (2022). “Assessing the Determinants of Start-up Valuation through Prediction and Causal Discovery”. In: *Entrepreneurship Theory and Practice*, in press.

Garkavenko, Mariia, Hamid Mirisaee, Eric Gaussier, Agnès Guerraz, and Cédric Lagnier (2021). "Valuation of Startups: A Machine Learning Perspective". In: *Proceedings of the 43rd European Conference on Information Retrieval*. Springer, pp. 176–189.

Garkavenko, Mariia, Hamid Mirisaee, Eric Gaussier, Agnès Guerraz, and Cédric Lagnier (2022). *Where Do You Want To Invest? Predicting Startup Funding From Freely, Publicly Available Web Information*. arXiv: 2204.06479.

Gastaud, Clement, Theophile Carniel, and Jean-Michel Dalle (2019). *The varying importance of extrinsic factors in the success of startup fundraising: competition at early-stage and networks at growth-stage*. arXiv: 1906.03210 [q-fin.GN].

George, Gerard, Martine R. Haas, and Alex Pentland (2014). "Big Data and Management". In: *Academy of Management Journal* 57.2, pp. 321–326.

Giardino, Carmine, Xiaofeng Wang, and Pekka Abrahamsson (2014). "Why early-stage software startups fail: a behavioral framework". In: *International conference of software business*. Springer, pp. 27–41.

Glymour, Clark, Kun Zhang, and Peter Spirtes (2019). "Review of Causal Discovery Methods Based on Graphical Models". In: *Frontiers in Genetics* 10.

Gregorutti, Baptiste, Bertrand Michel, and Philippe Saint-Pierre (2015). "Grouped variable importance with random forests and application to multiple functional data analysis". In: *Computational Statistics & Data Analysis* 90, pp. 15–35.

Hand, David J and Robert J Till (2001). "A simple generalisation of the area under the ROC curve for multiple class classification problems". In: *Machine learning* 45.2, pp. 171–186.

Harrison, Richard T., Colin Mason, and Donald Smith (2015). "Heuristics, learning and the business angel investment decision-making process". In: *Entrepreneurship & Regional Development* 27.9-10, pp. 527–554.

Hart, David M (2014). "Founder nativity, founding team formation, and firm performance in the US high-tech sector". In: *International Entrepreneurship and Management Journal* 10.1, pp. 1–22.

He, Haibo and Edwardo A Garcia (2009). "Learning from imbalanced data". In: *IEEE Transactions on knowledge and data engineering* 21.9, pp. 1263–1284.

Herschtal, Alan and Bhavani Raskutti (2004). "Optimising area under the ROC curve using gradient descent". In: *Proceedings of the twenty-first international conference on Machine learning*. ACM, p. 49.

Hofman, Jake M., Duncan J. Watts, Susan Athey, Filiz Garip, Thomas L. Griffiths, Jon M. Kleinberg, Helen Z. Margetts, Sendhil Mullainathan, Matthew J. Salganik, Simine Vazire, Alessandro Vespignani, and Tal Yarkoni (2021). "Integrating explanation and prediction in computational social science." In: *Nature* 595, 181–188.

Hooker, Giles and Lucas Mentch (2019). "Please stop permuting features: An explanation and alternatives". In: *arXiv preprint arXiv:1905.03151*.

Hyytinen, Ari (2021). "Shared problem solving and design thinking in entrepreneurship research". In: *Journal of Business Venturing Insights* 16, e00254.

Ioffe, Sergey and Christian Szegedy (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *arXiv preprint arXiv:1502.03167*.

Jacobusse, Gert and Cor Veenman (2016). "On Selection Bias with Imbalanced Classes". In: *Discovery Science*. Ed. by Toon Calders, Michelangelo Ceci, and Donato Malerba. Cham: Springer International Publishing, pp. 325–340.

Jain, Shantanu, Martha White, and Predrag Radivojac (2017). "Recovering true classifier performance in positive-unlabeled learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1, pp. 2066–2072.

Kaiser, Marcus and Maksim Sipos (2021). "Unsuitability of NOTEARS for Causal Graph Discovery". In: *ArXiv* abs/2104.05441.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu (2017). "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in Neural Information Processing Systems*, pp. 3146–3154.

Kim, Taekyung and Changick Kim (2020). "Attract, Perturb, and Explore: Learning a Feature Alignment Network for Semi-supervised Domain Adaptation". In: *arXiv preprint arXiv:2007.09375*.

Kiryo, Ryuichi, Gang Niu, Marthinus C du Plessis, and Masashi Sugiyama (2017). "Positive-Unlabeled Learning with Non-Negative Risk Estimator". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 1674–1684.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer (2015). "Prediction Policy Problems". In: *American Economic Review* 105.5, pp. 491–495.

Knight, Russell M. (1994). "Criteria Used by Venture Capitalists: A Cross Cultural Analysis". In: *International Small Business Journal* 13.1, pp. 26–37.

Köhn, Andreas (2018). "The determinants of startup valuation in the venture capital context: A systematic review and avenues for future research". In: *Management Review Quarterly* 68.1, pp. 3–36.

Kolkman, Daan and Arjen van Witteloostuijn (2019). *Data Science in Strategy: Machine learning and text analysis in the study of firm growth*. Tinbergen Institute Discussion Papers 19-066/VI. Tinbergen Institute.

Korolev, Sergey, Amir Safiullin, Mikhail Belyaev, and Yulia Dodonova (2017). "Residual and plain convolutional neural networks for 3d brain mri classification". In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, pp. 835–838.

Ledoit, Olivier and Michael Wolf (2003). "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection". In: *Journal of empirical finance* 10.5, pp. 603–621.

Lévesque, Moren, Martin Obschonka, and Satish Nambisan (2020). "Pursuing Impactful Entrepreneurship Research Using Artificial Intelligence". In: *Entrepreneurship Theory and Practice*.

Li, Xiaoli and Bing Liu (2003). "Learning to classify texts using positive and unlabeled data". In: *IJCAI*. Vol. 3. 2003. Citeseer, pp. 587–592.

Lipovetsky, Stan and Michael Conklin (2001). "Analysis of regression in game theory approach". In: *Applied Stochastic Models in Business and Industry* 17.4, pp. 319–330.

Liu, Bing, Wee Sun Lee, Philip S. Yu, and Xiaoli Li (2002). "Partially Supervised Classification of Text Documents". In: *Proceedings of the Nineteenth International Conference on Machine Learning*. ICML '02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 387–394.

Lundberg, Scott M, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee (2020). "From local explanations to global understanding with explainable AI for trees". In: *Nature Machine Intelligence* 2.1, pp. 56–67.

Lundberg, Scott M and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Advances in Neural Information Processing Systems*, pp. 4765–4774.

Macmillan, Ian C., Robin Siegel, and P. N. Subba Narasimha (1985). "Criteria used by venture capitalists to evaluate new venture proposals". In: *Journal of Business Venturing* 1.1, pp. 119–128.

Martin, Lynn M, Izzy Warren-Smith, Jonathan M Scott, and Stephen Roper (2008). "Boards of directors and gender diversity in UK companies". In: *Gender in Management: An International Journal* 23.3, pp. 194–208.

Mason, Colin and Matthew Stark (2004). "What do Investors Look for in a Business Plan?: A Comparison of the Investment Criteria of Bankers, Venture Capitalists and Business Angels". In: *International Small Business Journal* 22.3, pp. 227–248.

Mathews, Sherin Mary (2019). "Explainable artificial intelligence applications in NLP, biomedical, and malware classification: A literature review". In: *Intelligent Computing-Proceedings of the Computing Conference*. Springer, pp. 1269–1292.

Maxwell, Andrew L., Scott A. Jeffrey, and Moren Lévesque (2011). "Business angel early stage decision making". In: *Journal of Business Venturing* 26.2, pp. 212–225.

Meek, Christopher (1995). "Causal Inference and Causal Explanation with Background Knowledge". In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. UAI'95. Montréal, Qué, Canada: Morgan Kaufmann Publishers Inc., 403–410.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Distributed Representations of Words and Phrases and Their Compositionality". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'13. Lake Tahoe, Nevada: Curran Associates Inc., 3111–3119.

Miloud, Tarek, Arild Aspelund, and Mathieu Cabrol (2012). "Startup valuation by venture capitalists: An empirical study". In: *Venture Capital* 14, pp. 151–174.

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar (2012). *Foundations of Machine Learning*. The MIT Press.

Molnar, Christoph (2019). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/.

Moreno-Torres, Jose G., Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera (2012). "A unifying view on dataset shift in classification". In: *Pattern Recognition* 45.1, pp. 521–530.

Nair, Vinod and Geoffrey E. Hinton (2010). "Rectified Linear Units Improve Restricted Boltzmann Machines". In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML'10. Haifa, Israel: Omnipress, 807–814.

Nielsen, Sabina (2010). "Top management team diversity: A review of theories and methodologies". In: *International Journal of Management Reviews* 12.3, pp. 301–316.

Obschonka, Martin and David B Audretsch (2019). "Artificial intelligence and big data in entrepreneurship: A new era has begun". In: *Small Business Economics*, pp. 1–11.

OECD and European Commission (2021). *The Missing Entrepreneurs 2021*, p. 327.

Pearl, Judea et al. (2000). "Models, reasoning and inference". In: *Cambridge, UK: CambridgeUniversityPress* 19, p. 2.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Prokhorenkova, Liudmila, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin (2018). "CatBoost: Unbiased boosting with categorical features". In: *Advances in Neural Information Processing Systems*, pp. 6638–6648.

Quintero, Sebastián (2019). "An Empirical Perspective on Startup Valuations". In: Radicle Working Paper.

Roe, Byron P, Hai-Jun Yang, Ji Zhu, Yong Liu, Ion Stancu, and Gordon McGregor (2005). "Boosted decision trees as an alternative to artificial neural networks for particle identification". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 543.2-3, pp. 577–584.

Roscher, Ribana, Bastian Bohn, Marco F Duarte, and Jochen Garcke (2020). "Explainable machine learning for scientific insights and discoveries". In: *IEEE Access* 8, pp. 42200–42216.

Sahlman, W. A. and Daniel R Scherlis (1987). "A Method For Valuing High-Risk, Long-Term Investments: The "Venture Capital Method"". In: Harvard Business School.

Saito, Kuniaki, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko (2019). "Semi-supervised domain adaptation via minimax entropy". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8050–8058.

Sakai, Tomoya, Gang Niu, and Masashi Sugiyama (2017). "Semi-supervised AUC optimization based on positive-unlabeled learning". In: *Machine Learning* 107, pp. 767–794.

Schwab, A. and Zhu Zhang (2019). "A New Methodological Frontier in Entrepreneurship Research: Big Data Studies". In: *Entrepreneurship Theory and Practice* 43, pp. 843–854.

Scutari, Marco (2010). "Learning Bayesian Networks with the bnlearn R Package". In: *Journal of Statistical Software* 35.3, pp. 1–22.

Shapley, Lloyd S (1953). "A value for n-person games". In: *Contributions to the Theory of Games* 2.28, pp. 307–317.

Sharchilev, Boris, Michael Roizner, Andrey Rumyantsev, Denis Ozornin, Pavel Serdyukov, and Maarten de Rijke (2018). "Web-based startup success prediction". In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 2283–2291.

Shimizu, Shohei, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen (2006). "A Linear Non-Gaussian Acyclic Model for Causal Discovery". In: *J. Mach. Learn. Res.* 7, 2003–2030.

Shimodaira, Hidetoshi (2000). "Improving predictive inference under covariate shift by weighting the log-likelihood function". In: *Journal of Statistical Planning and Inference* 90.2, pp. 227–244.

Shmueli, Galit (2010). "To Explain or to Predict?" In: *Statistical Science* 25.3, pp. 289–310.

Smart, Geoffrey H. (1999). "Management Assessment Methods in Venture Capital: An Empirical Analysis of Human Capital Valuation". In: *The Journal of Private Equity* 2.3, pp. 29–45.

Soto-Simeone, Aracely, Charlotta Sirén, and Torben Antretter (2020). "New venture survival: A review and extension". In: *International Journal of Management Reviews* 22.4, pp. 378–407.

Spirtes, P., C. Glymour, and R. Scheines (2000). *Causation, Prediction, and Search*. 2nd. MIT press.

Steffens, Paul, Siri Terjesen, and Per Davidsson (2012). "Birds of a feather get lost together: New venture team composition and performance". In: *Small Business Economics* 39.3, pp. 727–743.

Štrumbelj, Erik and Igor Kononenko (2014). "Explaining prediction models and individual predictions with feature contributions". In: *Knowledge and information systems* 41.3, pp. 647–665.

Sugiyama, Masashi and Klaus-Robert Müller (2005). "Model Selection Under Covariate Shift". In: *Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005*. Ed. by Włodzisław Duch, Janusz Kacprzyk, Erkki Oja, and Sławomir Zadrożny. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 235–240.

Tumasjan, Andranik, Reiner Braun, and Barbara Stolz (2021). "Twitter sentiment as a weak signal in venture capital financing". In: *Journal of Business Venturing* 36.2, p. 106062.

Verma, Thomas and Judea Pearl (1990). "Equivalence and Synthesis of Causal Models". In: *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*. UAI '90. USA: Elsevier Science Inc., 255–270.

Vogel, Rick, Tatjana Xenia Puhan, Edlira Shehu, Doron Kliger, and Henning Beese (2014). “Funding decisions and entrepreneurial team diversity: A field study”. In: *Journal of Economic Behavior & Organization* 107, pp. 595–613.

Xiang, Guang, Zeyu Zheng, Miaomiao Wen, Jason Hong, Carolyn Rose, and Chao Liu (2012). “A supervised approach to predict company acquisition with factual and topic features using profiles and news articles on techcrunch”. In: 6.1, pp. 607–610.

Yin, Bangqi and Jianxi Luo (2018). “How Do Accelerators Select Startups? Shifting Decision Criteria Across Stages”. In: *IEEE Transactions on Engineering Management* 65, pp. 574–589.

Żbikowski, Kamil and Piotr Antosiuk (2021). “A machine learning, bias-free approach for predicting business success using Crunchbase data”. In: *Information Processing & Management* 58.4, p. 102555.

Zhang, Jiji (2008). “On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias”. In: *Artificial Intelligence* 172.16, pp. 1873–1896.

Zhang, Jing, Xindong Wu, and Victor S Shengs (2014). “Active learning with imbalanced multiple noisy labeling”. In: *IEEE transactions on cybernetics* 45.5, pp. 1095–1107.

Zhang, Qizhen, Tengyuan Ye, Meryem Essaidi, Shivani Agarwal, Vincent Liu, and Boon Thau Loo (2017). “Predicting startup crowdfunding success through longitudinal social engagement analysis”. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, pp. 1937–1946.

Zhang, Shengming, Hao Zhong, Zixuan Yuan, and Hui Xiong (2021). “Scalable Heterogeneous Graph Neural Networks for Predicting High-Potential Early-Stage Startups”. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD ’21. Virtual Event, Singapore: Association for Computing Machinery, 2202–2211.

Zhang, Yanru and Ali Haghani (2015). “A gradient boosting method to improve travel time prediction”. In: *Transportation Research Part C: Emerging Technologies* 58, pp. 308–324.

Zheng, Xun, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing (2018). “DAGs with NO TEARS: Continuous Optimization for Structure Learning”. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS’18. Montréal, Canada: Curran Associates Inc., 9492–9503.

Zheng, Yanfeng, Jing Liu, and Gerard George (2010). “The dynamic impact of innovative capability and inter-firm network on firm valuation: A longitudinal study of biotechnology start-ups”. In: *Journal of Business Venturing* 25.6, pp. 593–609.

List of Figures

1.1 (a) Funding round: pre-money and post-money valuation difference. (b) News about the funding round example: we learn that Hugging Face raised $ 40 million. 5

3.1 Example of the filed SH01 form pages containing the information required to infer the fundraising amount, the valuation of the company and the equity of the investor. 22

3.2 Comparison of funding amounts between `Source` and `Target`. . 26

3.3 EU vs. the UK data: (a) funding amounts, (b) valuations, and (c) investors' equities in the funding rounds with announced valuation. The D statistics of the Kolmogorov-Smirnov test is provided in each case. 28

3.4 Comparison of European funding rounds with announced valuation ($\texttt{Source}_{EU}$), unannounced valuation ($\texttt{Target}_{EU}$) and the set of funding rounds for which the valuation was extracted from Companies House ($\texttt{Target}_{LAB}$). (a) funding amounts (b) startups' valuations (c) obtained investors' equities. The D statistics of the Kolmogorov-Smirnov test is provided in each case. 30

3.5 Unsupervised, Semi-Supervised and Supervised Domain Adaptation. 31

3.6 DANN architecture for the task of startup valuation prediction. Label is startup valuation. Source domain S is the distribution of the funding rounds for which the startup valuation was announced and target domain T is the distribution of the funding rounds for which the startup valuation was not announced (`Source` and `Target` sets respectively, introduced in Section 3.2.1) 34

3.7 Directed acyclic graphs corresponding to the same skeleton. Only the v-structure (in red) can be oriented without ambiguity from observational data. 55

3.8 Causal graph computed by PC algorithm: direct and first indirect causes of valuation. 65

3.9 Causal graph computed by PC: direct causes of valuation and their siblings. 66

3.10 Comparison of predictors and causal determinants of start-up valuation. 67

3.11 SAGE values obtained from ML models with different hyper-parameters. Coefficients of determination R^2 and Spearman rank correlation values are given. 69

3.12 PC results with respect to the p-value used by PC algorithm (hyperparameter α). 72

4.1 Geographical distribution of the top-10 countries in our dataset. 83

4.2 Illustration of the funding extraction algorithm: detected money amount and startup name are shown in boxes; fundraising verb is underlined. 89

4.3 Feature importance analysis using SHAP-values for the different semantic groups of features (a) financial, (b) the web presence and (c) social media activities. Features are stacked vertically based on their importance from top to bottom of each figure. Each dot represent an instance in the dataset with its corresponding SHAP value on X-axis. 101

List of Tables

2.1 AI-based Studies on startups' Investment and Valuation. 11

2.2 Exemplary Studies of the Factors of Startups' Investment and Valuation. 15

3.1 Summary of the data. CB: Crunchbase, CH: Companies House. . 31

3.2 Startup features used in this study. 38

3.3 Experimental results on $\texttt{Target}_{LAB(test)}$set. Bold numbers are used for models statistically significantly better than the other models (Wilcoxon signed-rank test with $p < 0.001$). Baselines are separated with a vertical line. S, T and T_L denote `Source`, `Target` and $\texttt{Target}_{LAB}$respectively. 41

3.4 Contribution of different feature groups. Bold numbers are used for models statistically significantly better than the other models (Wilcoxon signed-rank test with $p < 0.05$). 43

3.5 Variables Used in the Analysis. 47

3.6 Top 5 Hyper-Parameters Sorted by Mean R^2 on 5-Fold Cross-Validation on Train Data. 58

3.7 Feature Group Ranking Analysis 59

3.8 Variables' Predictive Power Ranking and Causal Relations to Startup Valuation. 61

3.9 SAGE Values Obtained from ML Models with Different Hyperparameters. 67

3.10 Stability of PC results with respect to the set of variables considered: confusion matrix for causal relations obtained on Set 0 (main paper) and Set II (with Spearman p-value set to 0.05). . . 70

3.11 Stability of PC results with respect to the p-value used by PC (hyperparameter α) : confusion matrix for causal relations with $\alpha = 0.01$ (main paper) and $\alpha = 0.05$. 73

4.1 Startup features used in this study. 85

4.2 PU setting. The upper part of the table illustrated the neural network approaches and the lower part shows the direct optimization of AUC from (Sakai, Niu, and Sugiyama, 2017). Neural networks are trained with 10 random seeds, mean and std. reported. 97

4.3 PN setting models comparison. The best values are shown in bold. Values shown in parenthesis and marked with * are calculated on the labels corrected by the human experts for the top 200 companies. 98

4.4 Comparison to the results reported in (Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov, and Rijke, 2018) on the dataset presented therein with WBSSP pipeline. Values marked with * are calculated on the labels corrected by the human experts for the top 200 companies. 99

4.5 Ablation Analysis 100