Prediction of purchasing behavior using Deep learning techniques

Kodjo Agbemadon

UBFC
UNIVERSITÉ
BOURGOGNE FRANCHE-COMTÉ

UFC
UNIVERSITÉ
DE FRANCHE-COMTÉ

THÈSE DE DOCTORAT

DE L'ÉTABLISSEMENT UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

PRÉPARÉE À L'UNIVERSITÉ DE FRANCHE-COMTÉ

École doctorale n°37

Sciences Pour l'Ingénieur et Microtechniques

Doctorat d'Informatique

par

KODJO AGBEMADON

Prediction of purchasing behavior using Deep Learning techniques

Prédictions de comportements d'achats par techniques de Deep Learning

Thèse présentée et soutenue à Belfort, le 15 Décembre 2022

Composition du Jury :

PROF CHRÉTIEN STÉPHANE	Université Lumière lyon 2	Rapporteur
MCF HDR VERNIER FLAVIEN	Universite Savoie Mont Blanc	Rapporteur
PROF SPITÉRI PIERRE	ENSEEIHT	Examinateur
PROF COUTURIER RAPHAËL	Université Bourgogne Franche-Comté	Directeur de thèse
MCF HDR LAIYMANI DAVID	Université Bourgogne Franche-Comté	Codirecteur de thèse

N° | X | X | X |

ABSTRACT

Prediction of purchasing behavior using Deep learning techniques

Kodjo Agbemadon
University of Bourgogne Franche Comté, 2022

Supervisors: Raphaël Couturier and David Laiymani

Recently, large retail companies often have the same suppliers, although some focus more on their own private label. The competitiveness of the retail sector relies, in part, on anticipating consumer reaction and optimising the supply chain. The receipts generated during purchases produce a large volume of data. The processing of these data can allow numerous analyses and predictions, particularly on purchasing behavior and supply chain improvement. The objective of this thesis is to predict purchasing behavior. To achieve this, we propose machine learning methods to detect the factors that lead to a change in purchasing behavior. More precisely, we have worked on the following challenges:

1. The application of machine learning models on sales dataset to identify customers at risk of no longer buying in stores. This is known as ”churn”. This study uses linear regression as a baseline to compare machine learning models such as gradient boosting, MLP (Multilayer perceptron) and LSTM (Long Short Term Memory). LSTM outperformed the other approaches due to the fact that they were designed to be able to learn order dependence.

2. The application of machine learning models to predict overstocking and wastage of short-life products, including fresh and dairy products. A total of 5 machine learning models were compared in this study, including gradient boosting, LSTM, Transformer, Informer, and AutoFormer. The latter was able to give the best results thanks to the time series decomposition into trend and seasonality components.

3. We have also worked on feeding different machine learning models (Gradient Boosting, LSTM and Autoformer) with priori-known future weather data, in order to improve demand forecasting. In each case tested, the modified algorithms always

perform better than their original versions.

The work presented in this thesis was the result of a collaboration between the retail company Colruyt France and the department of computer science of complex systems (DISC) of the FEMTO-ST laboratory under a CIFRE contract.

KEYWORDS: Retail industry, Purchasing behavior prediction, Churn prediction, Overstock prediction, Time series classification, Deep learning

RÉSUMÉ

Prédictions de comportements d'achats par techniques de Deep learning

Kodjo Agbemadon
Université de Bourgogne Franche Comté, 2020

Encadrants: Raphaël Couturier et David Laiyman

De nos jours, toutes les grandes entreprises de vente au détail ont bien souvent les mêmes fournisseurs, même si certains se concentrent davantage sur leur propre marque de distributeur. La compétitivité du secteur de la vente au détail repose, entre autres, sur l'anticipation de la réaction des consommateurs et l'optimisation de la chaîne d'approvisionnement. Les tickets de caisse générés lors des achats produisent un grand volume de données, dont le traitement peut permettre de nombreuses analyses et prédictions notamment sur le comportement des clients et l'amélioration de la chaîne d'approvisionnement. L'objectif de cette thèse est de prédire le comportement d'achats des consommateurs. Pour y parvenir nous proposons des méthodes d'apprentissage automatique permettant de détecter les facteurs engendrant un changement de comportement client. Plus précisément, nous avons travaillé sur les problématiques suivantes :

1. L'application de modèles d'apprentissage automatique sur les données de vente afin d'identifier les clients risquant de ne plus acheter dans les magasins. On parle de clients "abandonnistes". Cette étude utilise la régression linéaire comme base de référence pour comparer les modèles d'apprentissage automatique comme le gradient boosting, le MLP (Multilayer perceptron) et le LSTM Long Short Term Memory). Les LSTM ont surpassé les autres approches du fait qu'ils ont été conçus pour être capables d'apprendre la dépendance d'ordre.

2. L'application de modèles d'apprentissage automatique pour prédire le sur-stockage et le gaspillage des produits à date limite de consommation réduite, notamment les produits frais et les produits laitiers. Au total, 5 modèles d'apprentissage automatique ont été comparés dans cette étude, dont le gradient boosting, le LSTM, le Transformer, l'Informer et l'Autoformer. Ce dernier a su donner les meilleurs

résultats grâce à son concept de décomposition des composantes tendancielles et saisonnières des séries temporelles.

3. Nous avons également travaillé à l'intégration de données météorologiques futures connues à l'avance à différents modèles d'apprentissage automatique (Gradient Boosting, LSTM et Autoformer) afin d'améliorer la prédiction de consommation. Dans chaque cas de figure testé, les algorithmes modifiés donnent toujours de meilleurs résultats que leurs versions originales.

Les travaux présentés dans cette thèse ont été le fruit d'une collaboration entre la société de la grande distribution Colruyt France et le département d'informatique des systèmes complexes (DISC) du laboratoire FEMTO-ST dans le cadre d'un contrat CIFRE.

KEYWORDS: Commerce de détail, Prédiction du comportement d'achat, Prédiction des abandonnistes, Prédiction du surstock, Classification des séries temporelles, Deep learning

ACKNOWLEDGEMENTS

This work was done as a part of a CIFRE (N 2019/1139) project with Colruyt France, funded by the Ministry of Higher Education and Research of France, managed by the Association Nationale de la Recherche et de la Technologie (ANRT) and was partially supported by the EIPHI Graduate School (contract ”ANR-17-EURE-0002”).

CONTENTS

List of abbreviations

1-NN . . . . . 1 Nearest Neighbor

1-NN-DTW . . . 1-NN with DTW distance

AI . . . . . Artificial Intelligence

ANN . . . . . Artificial Neural Network

AR . . . . . Auto Regressive

ARIMA . . . . . Auto Regressive Integrated Moving Average

ARMA . . . . . Auto Regressive Moving Average

CNN . . . . . Convolutional Neural Network

DL . . . . . Deep Learning

DNN . . . . . Deep Neural Network

DT . . . . . Decision Tree

DTW . . . . . Dynamic Time Warping

DWT . . . . . Discrete Wavelet Transform

k-NN . . . . . k-Nearest Neighbors

LR . . . . . Logistic Regression

LSTM . . . . . Long Short-Term Memory

ML . . . . . Machine Learning

MLP . . . . . Multi Layer Perceptron

MTS . . . . . Multivariate Time Series

NN . . . . . Neural Network

RC . . . . . Reservoir Computing

RF . . . . . Random Forest

RFM . . . . . Recency, Frequency, Money

RMSE . . . . . Root Mean Square Error

RNN . . . . . Recurrent Neural Network

SVM . . . . . Support Vector Machine

I

DISSERTATION INTRODUCTION

1

INTRODUCTION

This thesis focuses on advancing the state-of-the-art of purchasing behavior prediction, with a particular emphasis on time series prediction and classification. It will leverage Deep Learning methods in an attempt to predict purchasing behavior. The presented research was conducted between the company Colruyt France and the Department of Computer Science and Complex Systems (DISC) of the laboratory FEMTO-ST in the framework of a CIFRE contract. In the rest of this thesis, the author will be referred to as "we", rather than "I". This introductory chapter discusses the general context and the considered use cases for our works, then briefly presents the contributions of this thesis.

1.1/ INTRODUCTION TO PURCHASING BEHAVIOR PREDICTION

Buying behavior in the retail sector, also known as selling behavior, is basically the decision-making process and actions of potential customers involved in the purchase and use of products. In other words, the fact that customers decide whether or not to purchase products from a retail company. Buying behavior deserves to be studied in detail as it is the main profit generating factor of a retail business. Understanding it means figuring out how a store's sales will evolve, which means buying from suppliers the products that consumers are really interested in and optimizing the supply chain. When these are properly arranged, it leads to the creation of profit for the company. In today's retail industry, even from a purely legal and fiscal point of view, every sales transaction must be traceable and therefore must generate data. Retail companies have a legal obligation to record all sales transactions and to be able to provide the integrity of these data during a tax audit. These transactions only concern the information on the receipt, not the consumer's personal information. There is also the data from the loyalty cards. These loyalty cards make it possible to associate customers with their purchases if the customers do not object beforehand.

With the advent of machine learning tools, these data have become extremely valuable.

The loyalty card, which originally was not considered for this purpose of associating data with customers, has seen its role evolve over time. Loyalty programs and cards date back to 1793, when American retailers began handing out copper coins to each customer after their purchase. The customer could collect these coins and redeem them on future purchases from the same retailer. This naturally drives the customer to return to the same retailer. Today, the loyalty card is mainly used to analyze the purchasing behavior per household or per customer.

Given the amount of data generated and with the possibility of crossing external data, it becomes essential to think about approaches based on machine learning or Deep Learning. Buying behavior, as seen above, is based on the decision of humans to buy a product. The prediction of buying behavior is no exception among other attempts to predict human behavior which is quite challenging. There will always be things that will be very difficult or impossible to predict, and others that Deep Learning models will manage to capture. As an illustration, here is a non-exhaustive list of examples of black swans, i.e. events that are difficult to predict:

- A client who goes on vacation due to an obligation or personal decision during an off-season. Especially the first time it happens.
- A health crisis that will disrupt all consumption habits. The Covid 19 health crisis is a good example, since it was able to change in very short period the consumption habits of a large number of consumers around the world.
- A customer who moves because of a major change in his life, for work, family. A customer can move for many reasons and end up in a place where he can not easily find a store of the same brand.
- An ageing loyal customer who dies, even if we do not want to, in the data we find this kind of pattern where a loyal customer suddenly stops buying from the retail company.

And here is a list of some predictable cases:

- A single customer who becomes a couple and starts having children (their consumption will still increase over time)
- A loyal customer who gradually starts to change his consumption habits, whether it is positive or negative. An important aspect is whether the transition is progressive.

The idea is to focus on everything that is predictable, especially since in datasets, there are more predictable cases than cases that are difficult to predict. This thesis has been accomplished with the aim of exploring machine learning and Deep Learning approaches

with the same objective of being able to predict purchasing behaviors. All the contributions are essentially focused on this aspect.

1.2/ Main Contributions of this Dissertation

The main contributions of this thesis focus on the cases of purchasing behavior prediction, namely churn detection, which means detecting customers who are likely to stop buying from the company, and demand prediction. These main contributions can be summarized as follows:

1. First, we propose a churn detection model based on Deep Learning. This study demonstrates how transactional data and machine learning can be utilized to forecast churn in the retail industry. A sample of 5,115,472 consumers loyalty cards records was used to train machine learning models. The outcomes firstly provide figures on the gap between the classical approaches and the machine learning models. Then, by comparing the machine learning models, it can be seen that some were more suitable for the type of problem than others. The Long Short Term Memory (LSTM) [11] for example was the most accurate in this contribution.

2. Secondly, we proposed a Deep Learning model allowing the prediction of overstock. As success in supply-chain relies on good stock management, we train the machine learning models to predict overstock. The term ”overstock” refers to too much stock in a store that has not been sold. It depends on both supply and consumer demand. If retailers can know the quantity that customers actually need, there will be no overstock. Therefore, in this contribution, we try to predict this quantity in order to provide it to the service in charge of the supply which will then step in, and will order exactly the necessary quantity. The results revealed that the attention based models called Autoformer [108] manage to learn the long-term trend and the seasonality effect to make better predictions. Thanks to a decomposition architecture that integrates the series decomposition block as an internal operator, the Autoformer manages to progressively aggregate the long-term trend part from the intermediate forecasts. By adding the learned seasonality, it manages to give the best results in this contribution.

3. Finally, we focused on feeding different machine learning models (Gradient Boosting, LSTM, Autoformer,...) with priori-known future weather data to improve the demand forecasting addressed in the previous contribution. The data used here for training and testing the models include product families demand is highly influenced by weather, such as barbecue meat, some beverages, appetizers... In this contribution, the previously mentioned models were compared. These models are gathered

in a group called “default models”, then a modification is applied to all these models that will allow them to consider the future values of the weather during the training and testing. The models resulting from this operation are called “modified models”. We compared the 2 groups, and the results showed that the modified models always gave the best results.

1.3/ DISSERTATION OUTLINE

The rest of this dissertation is organized as follow: Chapter 2 introduces the retail sector in general, the data generated during sales transactions and the regulations that govern the processing of these data. Chapter 3 discusses the scientific background around time series, their classifications, predictions and also some clustering algorithms. Chapter 4 presents the state-of-the-art on sales forecasting in the retail sector. Chapter 5 addresses the prediction of customers who are likely to stop buying in stores. It is known as ”churn” detection. Chapter 6 discusses the application of machine learning models to predict overstocking and waste of short-life products, including fresh and dairy products. Chapter 7 cover the integration of known future weather data with different machine learning models (Gradient Boosting, LSTM, Autoformer, Transformer and Informer) to improve consumption prediction.

II

Purchasing behavior prediction: From data collection to Artificial Intelligence

2

THE RETAIL INDUSTRY

Supermarkets are always trying to offer something that suits their customers, but it is clear that every human has specific needs. Offers/promotions will not necessarily be suitable for all customers. Some will prefer an offer and others who will not appreciate the same offer. This is why it is important to better analyze the buying behavior of each customer in order to make the most appropriate proposals for each of them. This same analysis allows to reduce the uncertainty of sales. The more the retail company knows about its customers, the more it can offer them a range of products that meet their needs and less and less products will end up unsold.

The retail sector includes all businesses that sell goods and services to consumers. Various retail sales and store types can be found all throughout the world, including grocery, convenience, discount, independent, department, DIY Do-it-yourself, electrical, and specialty shops. The retail sector expands consistently year after year and employs a significant workforce globally, especially with the rise in popularity of online shopping.

These days, all major retail competitors have roughly the same suppliers, although some are more focused on their own private label. The competitiveness of the retail sector is more about anticipating consumer reaction, supply chain optimization, forecasting and analysis.

In this chapter, we will present retail in general and the common challenges encountered, and especially those that the contributions in this thesis will address.

2.1/ CUSTOMER LOYALTY

Retail sector is a field that closely follows changes in society. It can be seen that in France, this sector has developed remarkably well over the last forty years, preceding consumer trends or adapting to societal evolutions. The constant increase of the sales surfaces, the natural growth of the market and an innovative expertise have allowed a regular progression in the retail sector. According to the Mercator [72], loyalty or “Customer

Loyalty management" is an action and policy linked to the product, price, communication, promotion or a specific program, designed to strengthen customer loyalty to a brand by reducing attrition and increasing customer share. A loyalty policy act on three components of loyalty: affective (emotional closeness), cognitive (preference) and conative [8] (buying behavior). As opposed to cognitive or affective, *conative* refers to intentional action, but not necessarily rational.

Customer loyalty can be:

1. Active: loyalty that results from an attachment or preference, of a rational or affective nature, for a brand or a company.

2. Passive: loyalty that results from personal factors (routine, laziness, perceived risk to change) or external factors, which make it difficult or impossible for a customer to change brand or supplier.

With these definitions, it is clear that loyalty is the result of a long-term relationship with the customer. This will allow to develop the sales towards the regular customers but also towards new customers who will have been conquered. Some customers will be considered as promoters of the brand by their commitment to make the brand known to them, others will be neutral and finally some will be detractors.

To build customer loyalty, "companies must know their customers and identify the criteria for choice and appreciation that they consider essential in order to adapt their offer to their needs. Not all customers have the same buying motivations and this can be complex"— [62]. A customer is unique, but companies cannot easily meet a single need multiplied by several thousand or even several tens of thousands of customers.

However, it is known that most consumers are not totally loyal but often buy in other stores than their favorite one, hence the notion of "customer volatility". There are 3 types of stores [30]:

- A primary store is the store that the consumer visits first and spends the most money in;

- A secondary store is a store that is regularly visited by the consumer and that immediately follows the main store in terms of spending;

- An occasional store that is visited occasionally, it can be considered as a convenience store to supplement the purchase of products.

2.2/ FREQUENTLY USED CUSTOMER LOYALTY ASSESSMENT TOOLS AND REGULATIONS

In retail industry, companies have different tools to measure customer loyalty. They operate through a personalized marketing, they must address the customer in a unique way, i.e. what they buy, what they consume. To address customers, companies use mainly digital through personalized mail, promotions or questionnaires. For the analysis and evaluation of customer loyalty, companies often use the tools presented below such as RFM (Recency, Frequency, Money), Loyalty Programs, Website, Newsletter. We will also briefly talk about the General Data Protection Regulation (GDPR) which governs the processing of private customer data.

RFM Recency, Frequency, Money (RFM) analysis is also used by many companies but each one interprets it in its own way. The RFM model assigns a score from 1 to 5 (worst to best) to customers in each of the three categories. This tool allows to:

- Identify the firm's best customers based on the nature of their spending habits.
- Evaluates customers by scoring them in three categories: how recently they've made a purchase, how often they buy, and the size of their purchases.
- the share of revenue from new customers (versus repeat customers).

This type of model can be crossed with other information such as the purchase rate versus the age of the customers for example.

Loyalty programs Loyalty programs are often materialized by the issuance of loyalty cards [28], which are widely used by retailers. Indeed, most supermarkets and hypermarkets have millions of "magic cards" in circulation. Thus, the card carrying customer loyalty programs becomes an essential element, not only to build customer loyalty, but also to avoid being disqualified from competitors. Loyalty cards are real trackers that allow to know precisely the purchasing behavior of customers, the total of purchases, discounts, the stores that customers frequent, their location... [30]. In a database dedicated to customer analysis, all this information is cross-referenced to determine typical typologies per customer (case of Colruyt's RFM model). However, some authors, Dowling and Uncles [10] are concerned about the effectiveness of these loyalty cards and, above all, about the psychological reaction of consumers that these loyalty programs can cause.

Website In an increasingly digitalized life, websites are essential for the good of the company. They allow to get in touch with consumers through communications, because

they can be consulted at any time of the day or night, they allow the company to expand its customer base and reach more prospects. Unlike other media, the Web is accessible from anywhere in the world at the same time as many other users. The website is a real investment in the short and medium term because it brings a gain in notoriety and time freed up. Indeed, unlike a physical store, it is not closed and the Web can answer questions asked by consumers in an autonomous way through the FAQ (Frequently Asked Questions) or ChatBot (interactive conversational robots which discuss with customers and answer some basic requests), these tools allow not to ask for an advisor for low added value requests. Moreover, with specialized software such as SimilarWeb, companies have access to reports on the usage of the website, for example, to know the global/national ranking of the website, the number of visits per week/month/year, the bounce rate, the average number of pages per visit, the average time spent on the website, the audience's interests, the marketing channels, the searched keywords... Websites are real sources of information that allow to have a lot of knowledge about their customers.

Newsletter The Newsletter is a form of mailing that consists in sending messages to an acquired and predefined database on a periodic basis. It allows to keep in touch with the customers and fulfills several objectives such as:

- The development of customer loyalty.
- Prospecting for new customers.
- The revival of inactive customers.
- Generating traffic on the website.
- Improving the brand image of the company.
- Increase sales and ROI (Return On Investment).

Newsletters reduce communication costs for the company, but can still be easily unread or deleted by the customer. However, even printed newsletters are not spared from being unread or thrown in the trash.

General Data Protection Regulation The digital economy, at the heart of business growth and competitiveness, relies heavily on the trust of customers and citizens. This trust can only be granted or maintained if companies and administrations behave in a fair and transparent manner when processing personal data. The General Data Protection Regulation (GDPR) provides a structure for building this trust [78]. This European directive was created to protect users from web giants such as GAFA (Google, Apple, Facebook, Amazon). In France, GDPR is relatively recent, the law has been applicable

since March 25, 2018, GDPR takes up in principle the text of the Data Protection Act of January 6, 1978. In general, this legal text ensures the protection of individuals in the processing of personal data (which fall under the privacy and freedoms of individuals), the affirmation of rights for the protected person, all under the control of a new administrative and independent authority, the CNIL (Commission nationale de l'informatique et des libertés). The GDPR allows customers to exercise several rights regarding their data. Customers have the right to ask, free of charge, what personal data a retail company such as Colruyt holds (right of access) and, if necessary, to have it rectified (right of rectification), transferred (right to data portability) or deleted free of charge (right of erasure). The customer may also request to limit the processing of his or her personal data (right to limitation of processing) or to object to such processing (right to object). With regard to the newsletters sent by the company, the customer may at any time object to the processing of his data for marketing purposes directly by e-mail by clicking on the unsubscribe button under each e-mail. According to the conditions provided by the CNIL, the customer can also at any time ask the company to access his personal data. In processing customer data such as loyalty card data, the company must be in compliance with the GDPR because it risks fines of up to 20 million euros or 4% of the group's global turnover, but also a bad reputation.

2.3/ COLRUYT GROUP & COLRUYT FRANCE

Colruyt France is a retail company with a chain of supermarkets mainly located in the Franche-Comté region, France. It is the partner company of the CIFRE project that supported the work done during this thesis. Colruyt France belongs to a larger group called Colruyt Group.

2.3.1/ COLRUYT GROUP

Colruyt Group is present in Belgium (number one Belgian food retailer [29] as of August 2022), France, Luxembourg and India. In 2022, the group had a total of 744 stores with 32,996 employees. It includes other retail subsidiaries such as Collect&Go, DreamLand, Okay, BIO-planet, ColliShop, CRU, Colex, Dreambaby and Fiets. Its wholesale and foodservice subsidiaries include Spar, Alvo, Mini market, CocciMarket, Panier Sympa and Solucious. The group has subsidiaries in other activities such as Dats24, Symeta and Eoly. Between 2021 and 2022, the company has invested 488 million euros, and has obtained a turnover of 10.049 billion. This turnover is divided between Retail (8,165 millions €), Wholesale and Foodservice (1,065 millions €), Other activities (819 millions €) The details on the turnover distribution between the different types of services can be

found on the Figure 2.1.

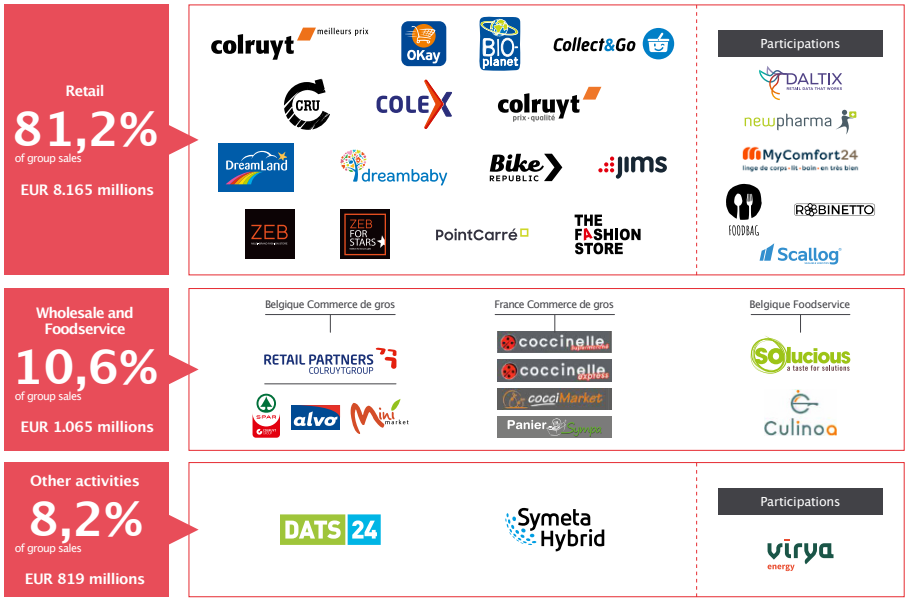


Figure 2.1: The different subsidiaries of Colruyt Group with the distribution of the turnover recorded in 2022 per type of service

2.3.2/ COLRUYT FRANCE

The group made its first move into France in 1996 with the acquisition of Ripotot distribution group that was soon renamed to Codifrance (Colruyt Distribution France). Its first store was opened in Pontarlier (25) and in 2014 Colruyt France has launched a “Click & Collect” service named Collect&Go. This service has undergone an impressive evolution especially due to the Covid 19 health crisis. In 2022, Colruyt France includes 91 stores (from 680 to 1,200m^2) with 2,587 employees. Its stores are designed to be local stores but still offer a large choice of more than 9,000 product references mainly in fresh and dry food. The shelves of the supermarket contain the essential non-food items. Each store is equipped with a meat-cutting laboratory, so the company is able to offer preparations, always made on the spot, often based on regional specialities.

Colruytplus card It was created in 2014 with the aim of bringing immediate discounts to customers. Indeed customers benefit from discounts on some promotional products but also -5% permanent on all Colruyt own brands (Boni, Belle France, Délice de Belle France). Customers who do not have the card will not be able to benefit from the discounts granted. The card allows the customer to be rewarded immediately. Thereafter, he can benefit from a rewarded loyalty if he meets the criteria of purchases, frequency, amount, ... defined by the company. At the very beginning, this loyalty card was con-

ceived and used with the strict purpose of offering discounts to all customers who carry it, and with the idea of pushing them to be loyal to the company.

A typical day in a Colruyt supermarket can generate hundreds of thousands till receipts, which also means hundreds of thousands registrations in databases. Out of this batch of records, on average 75% are related to loyalty cards and the rest are not. Non-loyalty card data are as important as loyalty card data. Loyalty card data can be used to make predictions at the customer level, and non-loyalty card data can be used to make global predictions of demand. A good example of this is trying to predict the perfect quantity of a product family to order from suppliers to avoid unsold products. This kind of prediction can very well be done by ignoring customer identification and just focusing on sales data and external factors. Until before this thesis, the machine learning model training was only done on Belgian data because all data science profiles were located in Belgium.

2.4/ CONCLUSION

We have just seen a presentation of customer loyalty, then the tools often used to evaluate it and the regulations that govern analysis and treatment of this type of data. We have seen that the retail industry started doing shopping behavior analytics and even finding techniques to evaluate and leverage the results of these shopping behavior analytics long before the advent of machine learning. This underscores the importance of customer analytics in the retail industry. We have talked about the classic and widely used analytics technique called RFM, which consists of segmenting customers based on recency, frequency and monetary criteria. We also presented how newsletters and websites are used in retail to maintain customer relationships and the General Data Protection Regulation (GDPR) with its role as guardian and defender of the customer's personal data rights. Finally, we presented the retail company behind this study, Colruyt France, with the corporate group Colruyt Group, to which it belongs, and more specifically its loyalty card system under the name Colruytplus.

3

Machine Learning for Time series

In this chapter, we will see in detail what time series represent, as well as the classification and regression approaches on time series data. Throughout this thesis, the term forecasting can be used as a substitute for regression. However the term prediction combines classification and regression, we will refer to classification when we have to predict the belonging to a class, and regression when we have to predict a value. We will also see what role Machine Learning plays in time series prediction approaches.

3.1/ Time series

Time series [31] are a set of observations arranged chronologically. These measurements can be taken continuously over time, or at a series of discrete time points. Conventionally, these two types of time series are respectively known as continuous and discrete time series. For instance, the time axis is discrete for discrete time series. Typically, for a continuous time series, the observed variable is a continuous variable recorded continuously on a graph [14]. In computer science, the representation of time series is discrete, but by misuse of language, a discrete time series, whose recorded values are relatively very close one to the other, can be considered as a continuous time series.

A time series can be univariate, meaning that only one variable varies over time. For example, data collected from a sensor that measures the temperature of a room every second. Therefore, every second, a value representing the temperature is generated. A time series can also be multivariate, meaning that several variables vary over time. For example, a combination of 3 weather indices, including temperature, wind and humidity. There are three values, one for each axis (x, y, z) and they vary simultaneously in time.

Time series are observed in relation to various events, and by a large variety of analysts or researchers, such as the economist observing yearly wheat prices, the meteorologist studying daily rainfall in a given city, the physicist studying the ambient noise level at a given point in the ocean, the electronic engineer studying the internal noise of a radio

receiver, the marketing analyst studying the impact of a product over period [2], or a data scientist trying to predict consumer purchasing behavior. A basic example of time series is plotted in Figure 3.1.

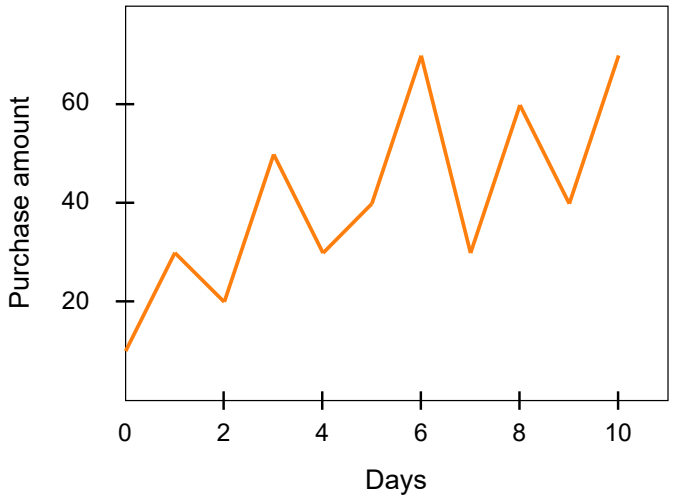


Figure 3.1: A graph showing the sales of one product at one store during 6 days.

The focus of this thesis will be on discrete time series. These can occur in three distinct ways, namely:

- By being sampled from a continuous series (an example in the weather field is an hourly measurements of temperature).
- By being aggregated over a period of time (for instance, total weekly sales).
- As an intrinsically discrete sequence (e.g. the number of participants in a monthly event over several months).

Time series are essentially a way to model sequential data to facilitate its analysis. This analysis is a “time series forecasting” problem when it is about predicting values. Or a “classification” problem when it comes to predicting which class an input belongs to, by showing the model many examples with labels. It can also be a clustering problem which is similar to classification except that there are no labels.

Classical time series analysis seeks to decompose the variation in a time series into its component parts:

- **Seasonal variation.** When similar patterns of behavior are observed at specific times of the year, this type of variation is typically annual in period and occurs for numerous series, whether measured weekly, monthly, or quarterly. As an illustration, ice cream sales are always highest in the summer. Notably, seasonal variation

cannot be determined if a time series is only measured annually (i.e., once per year).

- **Trend.** This type of variation is present when a series exhibits a steady upward growth or a steady downward decline over at least several consecutive time intervals. As an example, consider the world population growth. It has increased annually for many years. Trend may be loosely defined as ”long-term change in the mean level” however, there is no mathematically appropriate definition. The length of the observed series influence the perception of trend.
- **Irregular fluctuations.** The term ”irregular fluctuations” is frequently used to describe any variation that remains after trend, seasonality, and other systematic effects have been eliminated. Consequently, they may be entirely random, in which case they cannot be predicted. Nonetheless, they may exhibit short-term correlation (see below) or contain one-off discontinuities.

When the variation is dominated by a regular linear trend and/or regular seasonality, classical methods perform quite well. However, they perform poorly when the trend and/or seasonal effects vary over time or when the values of irregular fluctuations are correlated. In general, autocorrelation refers to the correlation between successive values of the same time series. It is frequently observed that successive residuals from a trend-and-seasonal model exhibit short-term (auto)correlation when separated by a brief time interval. To improve forecasts, a more sophisticated modeling approach may be necessary.

The following sections will focus on the two different types of predictions we make on time series throughout this thesis, namely classification and regression.

3.1.1/ CLASSIFICATION

Time series classification is a technique that uses supervised Machine Learning to cluster similar samples for classification. In this context, time series constitute these samples.

Time series classification is one of the supervised learning [27] methods widely discussed in Data mining literature over the last two decades [23, 40]. Supervised learning involves learning a mapping between a set of input variables X and an output variable Y and then applying this mapping to predict the outputs for unobserved data [27]. This is very similar to the goal of the first contribution 5 of this thesis. To predict whether a customer will not buy from the company in the future or not. To put it differently, to know if the customer will in the future belong to the churners class or not.

Since 2015, hundreds of time series classification algorithms have been proposed as temporal data availability has increased [90, 107]. Time series data are present in al-

most every task that requires some sort of human cognitive process due to their natural temporal ordering [53]. In fact, any classification problem that uses registered data and takes some notion of ordering into account can be qualified as a time series classification problem [74].

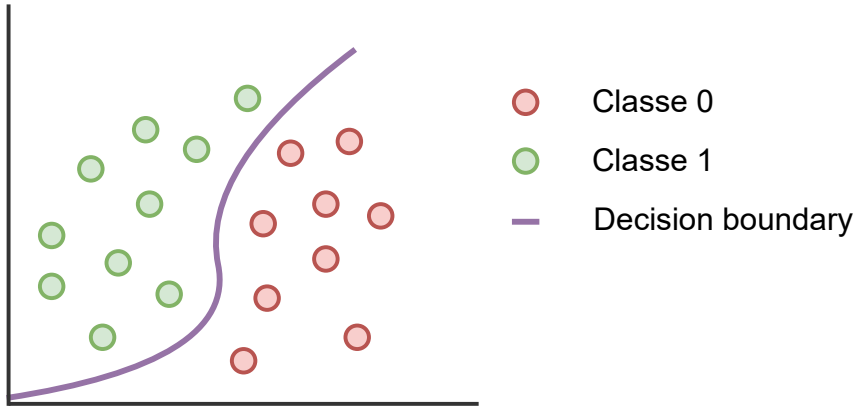


Figure 3.2: Here is a representation of a classification. The decision boundary could be linear or non-linear, depending on the chosen model.

On Figure 3.2, class 0 and 1 are the different possible classes. For example, one class can represent customers who are loyal and another class can represent customers who are not. Here, there are only 2 classes, so it is a binary classification, but in a classification problem, there can be more than 2 classes. This thesis focuses on binary classification.

Time series classification attempts to divide time series into distinct categorical classes and is used in a variety of applications. Pattern recognition, biometrics, sequences, signal recognition, sound, trajectories, churn prediction, and other applications are examples [43]. Time series classification, like any other classification problem, requires the use of a classification algorithm or procedure to identify distinct classes. The algorithm that will be used is determined by both the type of data available and the specific purpose and application. Time series data can be classified according to whether they are real value or discrete value, uniformly or non-uniformly sampled, and whether the data series are of equal or unequal length. Before classification operations can be performed, non-uniformly sampled data must be converted to uniformed data.

3.1.2/ FORECASTING

Time series forecasting is basically a technique that consists in predicting events through a temporal sequence using various algorithms. This technique makes it possible to predict future events by analyzing past trends, assuming that future trends depend on historical trends.

Time series forecasting models predict the future values of a target y_t for an entity at time t. That entity can be observed at the same time and represents a logical grouping of temporal information — such as purchase amounts, weather station measurements in climatology, wheat prices or rainfall in a given city. One-step forecast models have the following structure in the simplest scenario.

$$\hat{y}_{t+1} = f(y_{t-k:t}, x_{t-k:t}, s_i) \tag{3.1}$$

where $\hat{y}_{t+1}$ is the predicted value from the model, $y_{t-k:t} = y_{t-k}, \ldots, y_t$, $x_{t-k:t} = x_{t-k}, \ldots, x_t$ are observations of the exogenous and target inputs across a look-back window of k, s_i is static metadata related to the entity (for example, the customer's seniority, gender...), and $f()$ is the prediction function that the model has learned. Although this thesis concentrates on 1-D targets for univariate forecasting, the same components may be applied to multivariate models without losing generality [96].

3.2/ MACHINE LEARNING

Machine Learning is basically a subfield of Artificial Intelligence (AI) through which a system can be built by showing it examples and not by programming situations imagined in advance. The way humans solve problems has inspired the construction of Artificial Intelligence. Machine Learning has focused on creating algorithms that can learn from examples. The term ”Machine Learning” was coined by Arthur Samuel, an IBM computer scientist and a pioneer in AI and video games.

There are several forms of Machine Learning: supervised, unsupervised [84], semi-supervised and reinforcement learning. Each form of Machine Learning has different approaches, but they all share the same underlying theory. In this thesis we will focus on supervised learning [27]. Here are some terminologies often found in Machine Learning, with their meanings:

- data collection: Collecting the data from which the algorithm will learn
- data preparation: Formatting and engineering the data into the optimal format, extracting important features and reducing dimensionality
- training: Also known as “fitting”, this is where the Machine Learning algorithm learns when shown the cleaned data
- evaluation: This is when the model is tested to see if it performs well
- tuning: This is where the fine tuning of the model takes place to maximize its performance

Machine Learning has a sub-domain known as Deep Learning which will be of particular interest to us throughout this thesis.

3.2.1/ DEEP LEARNING

Deep Learning become an important part of many real-world applications. The explosive growth and availability of data, as well as the relentless advancement in the computational speed of hardware, has led to the emergence of new studies on Deep Learning. Deep Learning, which has its origins in classical Neural Networks, manages to significantly outperform its predecessors [87]. An illustration of the relation between Deep Learning and its predecessors can be seen on Figure 3.3. Many of the latest Deep Learning techniques have shown promising results in different types of applications such as visual data processing, natural language processing (NLP), speech and audio processing, time series forecasting, and many other applications[80, 63].

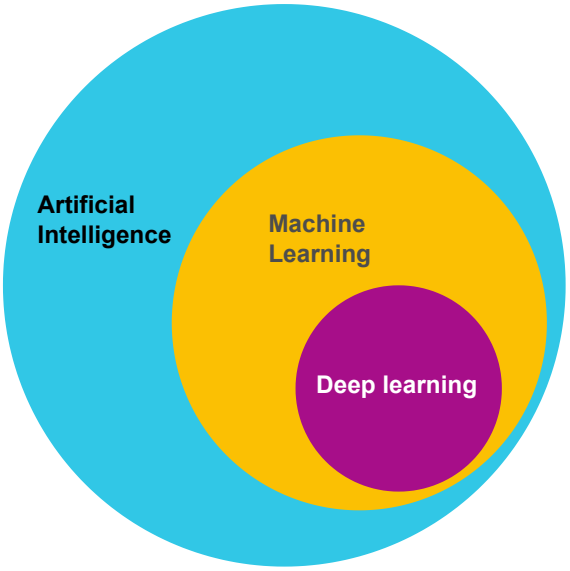


Figure 3.3: Illustration of relations between Deep Learning, Machine Learning and Artificial Intelligence.

Traditionally, the effectiveness of Machine Learning algorithms has depended heavily on the quality of the input data representation. Poor data representation often leads to a decrease in performance compared to good data representation [105]. Many studies on feature engineering have attempted to address this problem. These studies focus on the construction of features from raw data. In addition, feature engineering is often very domain-specific and requires significant human effort. In comparison, Deep Learning algorithms perform feature extraction in an automated manner, allowing researchers to extract defining features with minimal domain knowledge and human effort. These algorithms include a layered data representation architecture in which high-level features can

be extracted from the last layers of the networks and low-level features can be extracted from the lower layers. These architectures were initially inspired by Artificial Intelligence (AI), which simulated key sensory areas of the human brain. Our brain is capable of automatically extracting data representations from various scenes. The input is scene information from the eyes, and the output is the classified objects. This highlights the primary benefit of Deep Learning: it tries to mimic how the human brain works. Deep Learning has achieved remarkable success in many sectors and is now one of the most popular research areas in the machine-learning community.

A Deep Learning model essentially has an activation function, an input, an output, hidden layers, and a loss function. Any Deep Learning model tries to find a relationship between the input data and outputs using an algorithm and tries to make predictions on the unseen data. This algorithm tells the model how to find the value of the parameters (weights) that will minimize the error while matching the inputs and outputs. These algorithms are called optimizers. They significantly affect the accuracy of the Deep Learning model. They also affect the learning speed of the model.

We will see below the essential components of a Deep Learning model.

Neuron An artificial neuron is a mathematical function designed to roughly imitate a biological neuron. The artificial neuron is the elementary unit of an Artificial Neural Network. It receives one or more inputs representing excitatory postsynaptic potentials and inhibitory postsynaptic potentials at the level of neuronal dendrites, then adds them together to produce an output [6]. This output or activation represents the action potential of a biological neuron that is transmitted along its axon. Usually, each input is weighted separately, and the sum is passed through a nonlinear function called the activation function. Activation functions usually have a sigmoidal form, but they can also take the form of other non-linear functions such as ReLU. In Figure 3.4, we can see that a biological neuron is much more complex than an artificial neuron. The mathematical version is inspired by the biological version, with many simplifications. With the mathematical version, it is possible to solve problems that have long been difficult to address with a computer.

Loss Loss is simply the difference between the desired and predicted values. The loss function, also called objective function, is used to calculate it. During the training, the optimizer will try to minimize this loss. Some popular examples of loss functions are the Mean Square Error (MSE) and Mean absolute error (MAE) which are used for regression or the Cross Entropy Loss which is commonly used for classification problems.

Optimizer An optimizer is an algorithm or function that modifies the attributes of the Neural Network, for example, weights and learning rate. In this way, it helps to reduce the

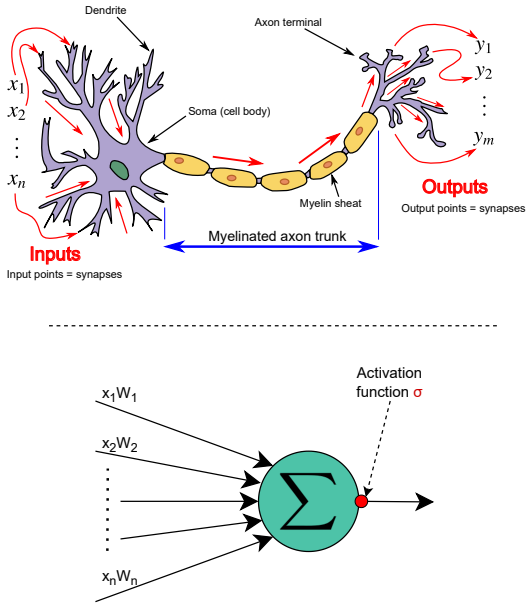


Figure 3.4: Biological and mathematical neuron. The biological neuron is much more complex than the artificial neuron.

overall loss and improve the accuracy. Since a Deep Learning model usually consists of millions of parameters, the problem of choosing the right weights for the model becomes difficult to solve. It is therefore necessary to choose an optimization algorithm adapted to the type of subject. Below are a few that are often used.

- Gradient Descent is a popular algorithm among optimizers. This optimization algorithm uses the gradient calculation to find out in which direction to change the parameters and reach the local minimum.
- Stochastic Gradient Descent SGD, The term stochastic refers to the randomness on which the algorithm is based. Unlike the default Gradient Descent which takes the whole data set at each iteration, in the Stochastic Gradient Descent, only randomly selected batches of data are taken. The batch designates the number of samples to be taken into account for the update of the model parameters.
- Adam (ADAptive Moment estimation), this optimization algorithm is a new extension of stochastic gradient descent to update the network weights while training. In contrast to maintaining a single learning rate throughout training in SGD, the Adam optimizer updates the learning rate of each network weight individually. Adam adds a first and second gradient moment and automatically adapts a learning rate for each parameter being optimized.

A vast amount of research has proposed various Deep Neural Networks architectures to solve natural language processing (NLP) tasks, such as machine translation [55, 51],

learning word embeddings [50], and document classification [54, 64]. DNNs have also had a substantial effect on the speech recognition community [41, 48].

The accessibility of open-source back-propagation frameworks [110, 111] has simplified network training, allowing customisation of network components and loss functions.

After a general overview of Machine Learning and Deep Learning main components, here is a presentation of some Machine Learning and Deep Learning models. These are Machine Learning methods that were relevant in this thesis.

3.2.2/ MACHINE LEARNING TO DEEP LEARNING MODELS

Extreme Gradient Boosting for classification Extreme Gradient Boosting (XGBoost) is an open-source library that provides an effective and efficient implementation of the Gradient Boosting algorithm [58]. It is worth presenting the latter a bit. Gradient Boosting represents a class of “ensemble” Machine Learning algorithms that can be used for predictive classification or regression modeling problems. The ensembles [18] are built from decision tree [60] models. Trees are added one by one to the ensemble and adjusted to correct the prediction errors made by the previous models. This is a type of ensemble Machine Learning model called boosting. The models are fitted using a chosen differentiable loss function and a gradient descent optimization algorithm. The technique is called ”Gradient Boosting” for this reason, as the loss gradient is minimized as the model is fitted, much like a Neural Network.

XGBoost has become the gold standard method and often the key component of winning solutions for classification and regression problems in Machine Learning competitions. Since this study involves a classification problem, it was deemed useful to test it in order to compare results with other approaches.

Here are some hyperparameters often used for its configuration:

- Number of estimators: The number of trees in the ensemble, which are often increased until no further improvement is seen.
- Maximum depth: The maximum depth of each tree. How far the decision tree can go.
- Eta α: The learning rate used to weight each model, often set to small values such as 0.3, 0.01, or smaller. The learning rate is a hyperparameter common to both Machine Learning and Deep Learning approaches. It provides the model with a scale to determine how much the model weights should be updated.

XGBoost can be used directly for classification and regression problems. In this thesis we tested the XGBoost Regressor API of the open-source Gradient Boosting implementation.

The configurations as well as the results will be presented in the contributions sections. A graphical representation can be found in Figure 3.5.

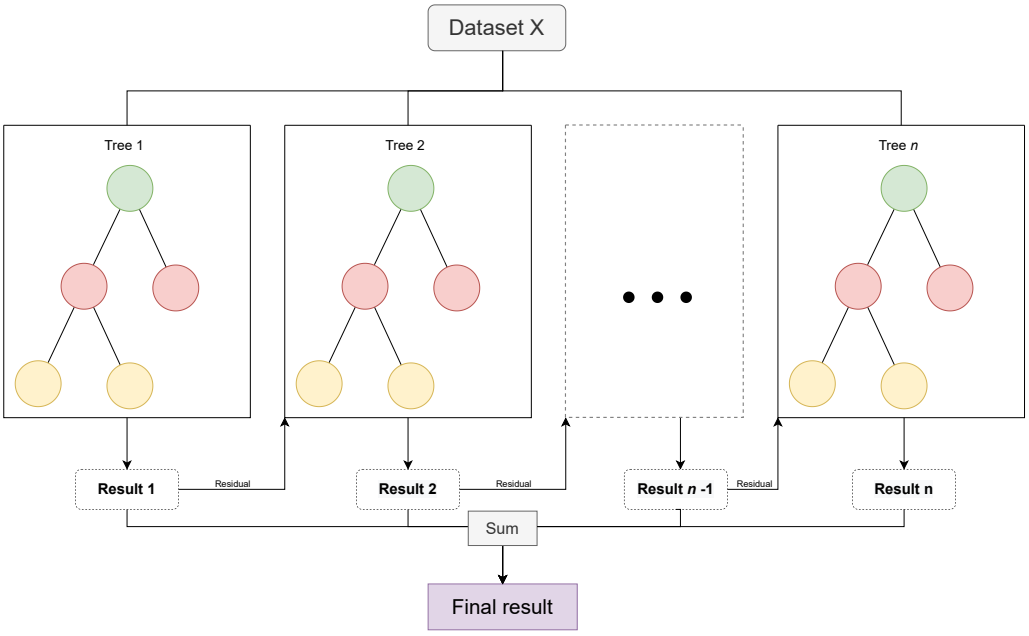


Figure 3.5: A simplified graphic representation of the XGBoost architecture. Each tree model in XGBoost minimizes the residual of its previous tree model.

MLP A Multi-Layer Perceptron (MLP) is a variant of the original Perceptron model proposed by Rosenblatt in 1950 [1]. MLP has proven itself in many fields with the emergence of Deep Learning [65]. It has one or more hidden layers between its input and output layers. The neurons are structured in layers, connections are always directed from lower layers to upper layers, and neurons in the same layer are not interconnected [68] as shown in Figure 3.6.

Convolutional Neural Networks A Convolutional Neural Network (CNN) is a Deep Learning Neural Network originally designed for arrays of data such as images. It seeks local relationships that are invariant across spatial dimensions [83]. CNN are very efficient in detecting patterns in the input image, such as circles, gradients, lines or even eyes and faces [83, 67].

Deep Convolutional Neural Networks (CNNs) have transformed computer vision [75]. In 2015, CNNs were used to achieve human-level performance in image recognition tasks [61].

To apply CNN to time series datasets, multiple layers of causal convolutions are used [67, 65, 42]. For instance, there are convolutional filters designed to ensure that only past information is used for prediction. For an intermediate feature at the hidden layer l, each

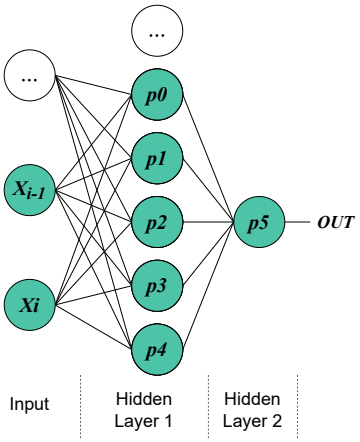


Figure 3.6: A graphical representation of a Multilayer Perceptron.

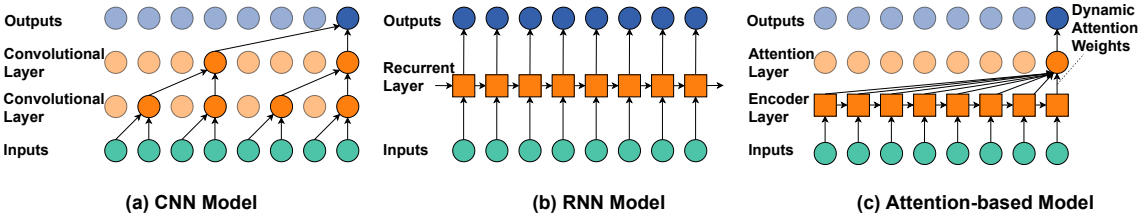


Figure 3.7: Illustrating temporal information using different encoder architectures.

causal convolutional filter takes the following form:

$$h_t^{l+1} = A\Big((\boldsymbol{W} \times \boldsymbol{h})(l,t)\Big), \tag{3.2}$$

$$(\boldsymbol{W} \times \boldsymbol{h})(l,t) = \sum_{n=0}^{k} \boldsymbol{W}(l,n)\boldsymbol{h}_{t-n}^{l} \tag{3.3}$$

where $h_t^l \in \mathbb{R}$ is an intermediate state at layer l at time t, $\times$ is the convolution operator, $\boldsymbol{W}(l,n) \in \mathbb{R}$ is a fixed filter weight at layer l, and $A()$ is an activation function, such as a sigmoid function, representing any architecture-specific nonlinear processing. For CNN that use a total of L convolution layers, we note that the encoder output is then $z_t = h_t^L$ [106]. Knowing that this is a one-dimensional situation, it can be seen that equation 3.3 is very similar to Finite Impulse Response (FIR) filters in digital signal processing [20]. Two key implications may follow for the temporal relationships learned by CNN. First, consistent with the spatial invariance assumptions for standard CNNs, temporal CNNs assume that the relationships are time-invariant — by using the same set of filter weights at each time step and over the whole time. Also, to perform prediction, CNNs are only able to use inputs within their receptive field, or defined lookback window. Thus, the size of the receptive field k will need to be carefully set to ensure that the model can use all relevant historical information. Interestingly, a single causal CNN layer with a linear activation function is equivalent to an auto-regressive (AR) model.

Dilated Convolutions The use of standard convolutional layers can be computationally challenging when long-term dependencies are important, as the number of parameters increases directly with the size of the receptive field. To overcome this problem, modern architectures frequently use dilated convolutional layers [67, 83], which extend equation 3.3 as follows:

$$(\boldsymbol{W} \times \boldsymbol{h}(l,t,d_l) = \sum_{n=0}^{\lfloor k/d_l \rfloor} \boldsymbol{W}(l,n)\boldsymbol{h}^l_{t-d_l n} \tag{3.4}$$

where $\lfloor . \rfloor$ is the floor operator and d_l is a layer-specific dilation rate. Dilated convolutions can hence be interpreted as convolutions of a down-sampled version of the lower layer features — reducing resolution to incorporate information from the distant past. Thus, by increasing the dilation rate with each layer, dilated convolutions can gradually aggregate information at different time blocks, allowing for more history to be used in an efficient manner. With WaveNet architecture [67] for instance, dilation rates are increased in powers of 2 with adjacent time blocks aggregated in each layer – allowing for 2^l time steps to be used at layer l as depicted in Figure 3.7.

Recurrent Neural Networks Recurrent Neural Networks (RNN) are essentially a type of Neural Network in which the output of the previous step is used as the input to the current step. RNNs were initially used in sequence modeling [65], with good results on a variety of natural language processing tasks [92]. In traditional Neural Networks, each input and output is independent, but in cases of predicting the next word in a sentence, the previous words are needed and therefore the previous words must be remembered. Similarly for time series, past values can influence the current value. This inspired the application of RNN for time series and the development of RNN-based architectures for time prediction [103, 98, 88].

RNN have a “memory” that retains all the information about what was calculated. A more detailed version can be seen in the Figure 3.8.

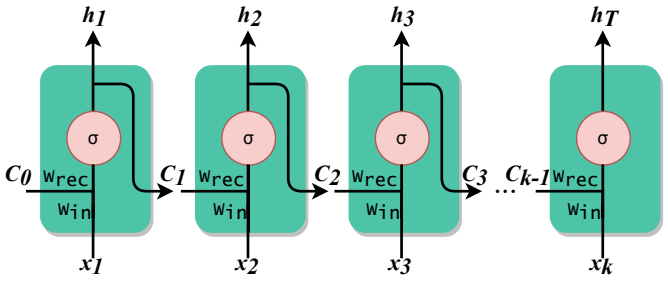


Figure 3.8: A graphical representation of RNN

The network has an input sequence of vectors [x(1), x(2),. . . , x(k)], and at a time step t,

there is an input vector x(t). Past information and learned knowledge are encoded in the network state vectors [c(1), c(2), ..., c(k-1)], and the network has an input state vector c(t-1) at time step t. The input vector x(t) and the state vector c(t-1) are concatenated to contain the complete input vector at time step t, [c(t-1), x(t)].

RNN uses the same parameters for each input as it performs the same task on all the inputs or hidden layers to produce the output. This reduces the complexity of parameters. The memory state is recursively updated with new observations at each time step as depicted in Figure 3.7. It is constructed in the following form:

$$z_t = v(z_{t-1}, y_t, x_t, s) \tag{3.5}$$

Where $z_t \in \mathbb{R}$, is the hidden internal state of the RNN, and $v()$, the learnt memory update function. For example, Elman's RNN [5] which is one of the simplest RNN versions, would look like this:

$$y_{t+1} = Y_y(\boldsymbol{W}_y z_t + \boldsymbol{b}_y), \tag{3.6}$$

$$\boldsymbol{z}_t = Y_y(\boldsymbol{W}_{z_1} z_{t-1} + \boldsymbol{W}_{z_2} y_t + \boldsymbol{W}_{z_3} x_t + \boldsymbol{W}_{z_4} s + \boldsymbol{b}_z) \tag{3.7}$$

Where W, b are the network's linear weights and biases, respectively, and $y(), z()$ are the network's activation functions. RNNs, unlike CNNs, do not require the explicit specification of a lookback window. From the standpoint of signal processing, the main recurrent layer in equation 3.7, looks like a nonlinear version of Infinite Impulse Response (IIR) filters.

Long Short-Term Memory (LSTM) Long Short-Term Memory (LSTM) networks are a type of recurrent Neural Network (RNN) that can learn order dependences in sequence prediction problems [11]. While introduced in the late 90's, LSTM models have only recently become a viable and powerful prediction technique for time-series. An LSTM rectifies a huge issue that recurrent Neural Networks suffer from: short term memory i.e. the inability to learn dependencies from long sequences.

Older version of RNN encounter limitations in learning long-range dependencies in the data, a problem known as "short term memory", which means the inability to learn dependencies from long sequences. The gradients carry information used in the RNN parameter update, and as the gradient continues to decrease, becoming smaller and smaller, the parameter updates become insignificant, implying that no real learning occurs [11]. Using a series of "gates" [12], each one with its own RNN, LSTM manages to keep, forget or ignore data points based on a probabilistic model. This is achieved by modulating a cell state c_t, which stores long-term information, via a series of gates. There is an illustration of LSTM shown in Figure 3.9.

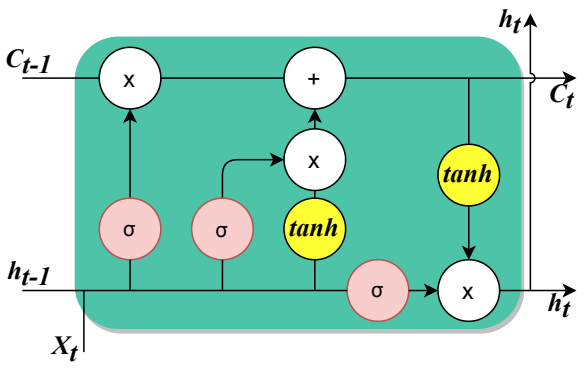


Figure 3.9: A graphical representation of LSTM memory cells.

Reservoir Computing A Reservoir computing is a form of Recurrent Neural Network in which the input data are transformed into spatio-temporal patterns in a high dimensional space called ”Reservoir” [26]. This Reservoir is then connected to the desired output by trainable units called Readout. The Reservoir is fixed and only the Readout is trained with a simple method such as linear regression. Thus the training cost is considerably reduced, allowing Reservoir Computing to achieve unprecedented learning speed compared to previous RNNs. Reservoir Computing models require few training data and are widely used for time series prediction, and even have very good scores on chaotic time series [86]. See Figure 3.10 for an illustration of the Reservoir Computing architecture.

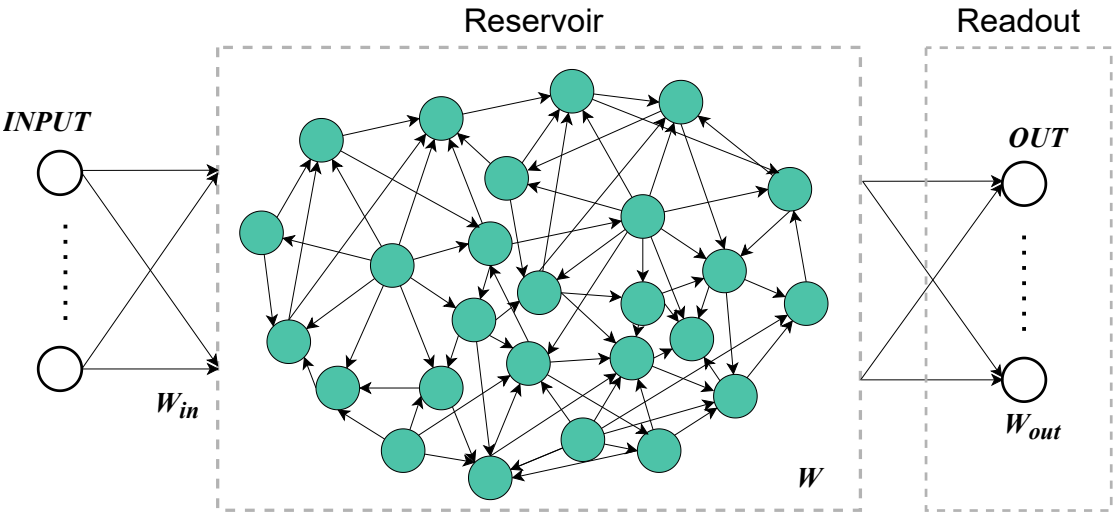


Figure 3.10: Simple illustration of Reservoir Computing architecture. The reservoir here is the high-dimensional space containing the space-time models. Only the Readout is trained with a simple method such as linear regression.

Attention family Transformer, Informer and Autoformer called Attention family networks are all fundamentally based on the mechanism of Attention [77]. Bahdanau et al. [51]

proposed the attention mechanism to address the bottleneck problem that arises when using a fixed-length encoding vector, as the decoder would have limited access to the information provided by the input. This is thought to be especially problematic for long and complex sequences, whose dimensionality would be forced to be the same as for shorter or simpler sequences. Among these 3 architectures, the Transformer was the first to be developed. Figure 3.11 shows an illustration of the original version of the Transformer. In

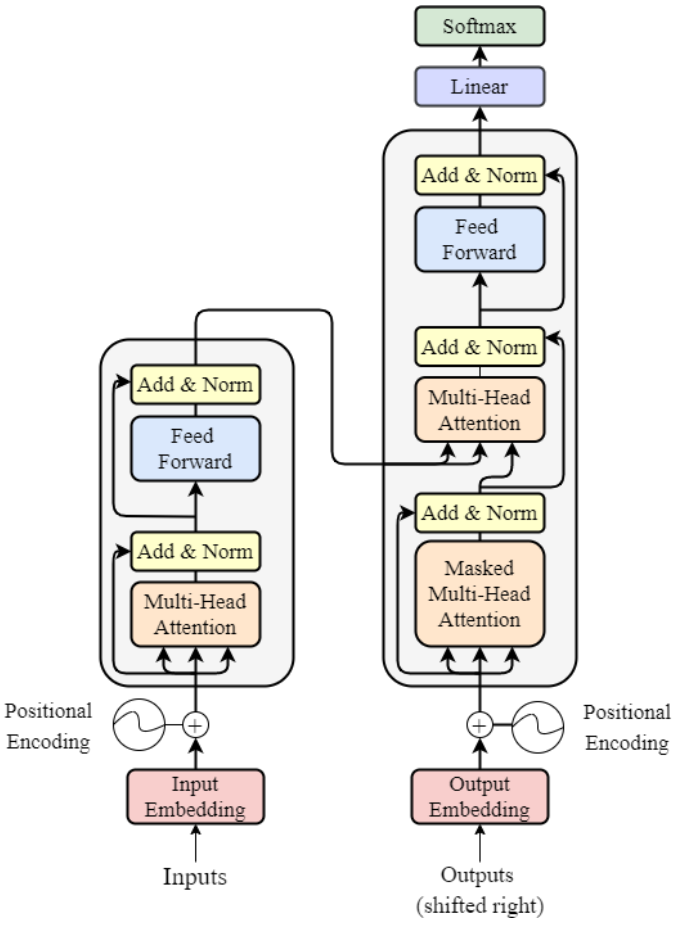


Figure 3.11: Original transformer model architecture

the following, the speciality of the Autoformer regarding the rest of Attention family will be explained.

Autoformer Autoformer is a decomposition architecture that embed the series decomposition block as an inner operator, which can progressively aggregate the long-term trend part from intermediate prediction [108]. Besides, there is an Auto-Correlation [108] mechanism to conduct dependencies discovery and information aggregation at the series level, which contrasts clearly from the previous attention family. See Figure 3.12 for an Autoformer illustration.

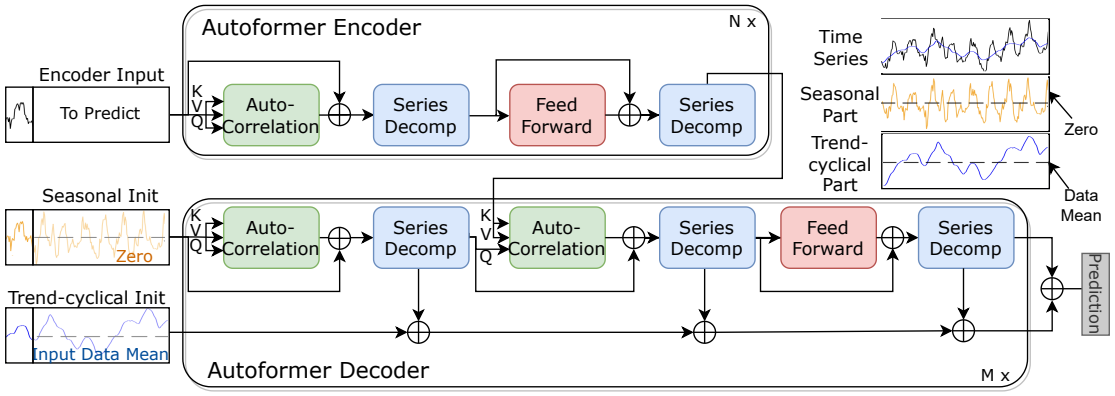


Figure 3.12: Autoformer architecture. The encoder eliminates the long-term trend-cyclical part by series decomposition blocks (blue blocks) and focuses on seasonal patterns modeling. The decoder accumulates the trend part extracted from hidden variables progressively. The past seasonal information from encoder is utilized by the encoder-decoder Auto-Correlation (center green block in decoder).

3.3/ MACHINE LEARNING APPLICATION ON TIME SERIES

Researchers have proposed hundreds of methods to solve the problem of accurately classifying and forecasting time series data [70].

Time series prediction has been a significant topic of academic research [106], integral to applications in fields such as biological sciences [47], climate modeling [95] and medicine [97], as well as finance [21] and commercial decision making in retail [71], to name a few.

Historically, in the literature, we can cite frequently used methods such as Dynamic Time Warping (DTW), the Nearest Neighbor classifier [70] and ensemble learning methods such as Random Forest and Support Vector Machine (SVM).

Following the non-Deep Learning classifiers for time series classification [70], there was Deep Learning [59] that has been successful in a variety of classification tasks that have motivated the recent use of Deep Learning models for time series classification [79]. Some publications have reported the use of CNN to predict chaotic and real time series. In 2015, Ding et al. adapted the dilated CNN architecture to a stock market prediction problem. In 2018, Hoermann et al. presented a deep CNN model for dynamic occupancy grid prediction with data from multiple sensors. We have seen more work using CNN architectures for classification rather than for prediction for multivariate time series. In the set of studies there is, for example, a study on the prediction of heart failure from heart rate data or activity from biometric time series [56].

There are also studies that compared LSTM to CNN for prediction. Livieris et al. [102] conducted a comparative study between CNN and LSTM for gold price time series pre-

diction and concluded that the LSTM gave the best results.

Wyffels et al. [35] conducted a comparative study on monthly time series prediction by Reservoir Computing, ARIMA. Several datasets were used (monthly electricity production, evolution of plastic, rubber, glass, metal products and machinery production). They showed that by slightly modifying the Reservoir Computing model, it has considerably increased its results and has given the best results in the comparison.

In other work, researchers have shown that Reservoir Computing techniques have difficulties in dealing with multiple time scales in time series. time scales in time series. Wyffels et al. have also shown that problems with multiple time scales can be solved in the area of time series prediction by time series decomposition.

In 2020, Bianchi et al. [100] compared Reservoir Computing (RC) with Deep Learning models (LSTM, GRU) for time series classification on a variety of datasets. In some cases RC performed better than the other models. They especially insisted on the execution time of the RC model.

In 2017, an Attention-based [77] DL architecture, specifically the Transformer have been successfully applied in time series predictions. After the Transformers, the Autoformers followed a while later. The Autoformer, based on the Attention mechanism, adds a decomposition function that has improved the results on time series prediction tasks.

Numerous Neural Network designs have arisen as a result of the diversity of time series challenges across numerous fields.

3.4/ FREQUENTLY USED METHODS AT COLRUYT

A clustering method named K-means, has been used a lot to meet the need in business. A brief presentation of this method will be made in this section. K-means is a clustering algorithm with many use cases in real life situations. This algorithm generates K clusters associated with a dataset, it can be done for various scenarios in different industries, including pattern detection, medical diagnosis, stock analysis, community detection, market segmentation, image segmentation, ... It provides quick insight into the data set under study by grouping similar data points that are close to each other. Data points from the same group called cluster will be close and similar to each other, while data points from other groups or clusters will be dissimilar. It is an unsupervised learning algorithm, which essentially means that the algorithm learns patterns from unlabeled data. This implies that you can train a model to create clusters on any given dataset without having to initially label the data. An illustration of clustering is shown in Figure 3.13.

The algorithm asks the user to specify the number of clusters K to search, and does not

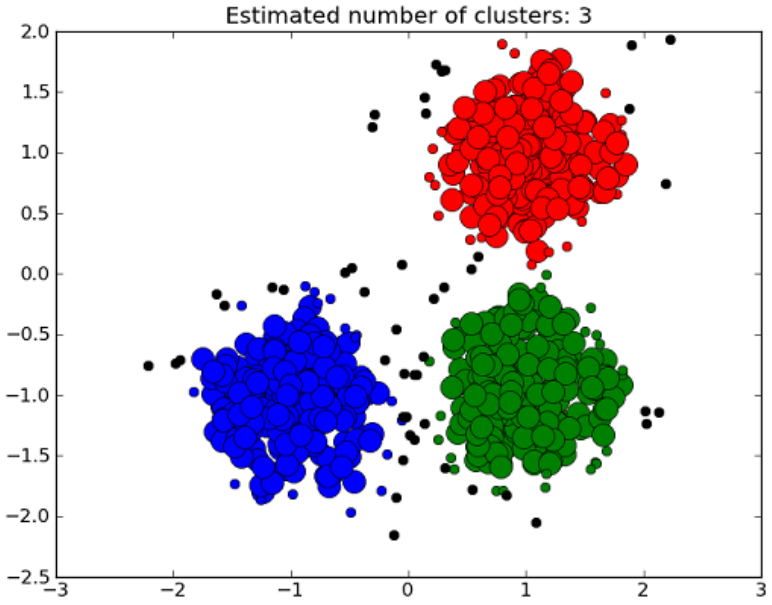


Figure 3.13: A graphical representation of clustering. In this example, a clustering algorithm has been applied to a dataset. The algorithm manages to organize them into 3 clusters.

learn it from the data. It is difficult to know if the value given to K is optimal. A deep knowledge of the domain allows to determine an ideal value of K. It can also be found using the elbow method.

The elbow method uses the sum of squared distances (SSE) to choose an ideal value of K based on the distance between the data points and their assigned clusters. Thus the K value where the SSE starts to flatten and we see an inflection point. When viewed, this graph looks a bit like an elbow, hence the name of the method. In Figure 3.14, the resemblance between the curve and an elbow can be seen, hence the name of the method.

3.5/ CONCLUSION

In this chapter, we first defined time series and what they represent. We saw the commonality of classical time series prediction approaches that usually try to decompose time series into seasonal variation, trend and irregular fluctuations. We have briefly defined the classification of time series and the problem of prediction.

We also presented the evolution of Machine Learning to Deep Learning techniques that have been involved in solving prediction problems, explaining the essential components of Machine Learning. We have presented the methods that have been used in this the-

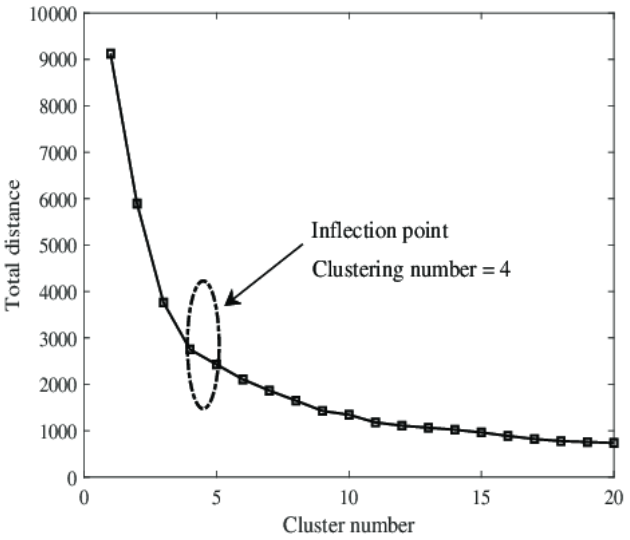


Figure 3.14: A graphical representation of the elbow method.

sis, such as Gradient Boosting, MLP, Convolutional Neural Networks, Recurrent Neural Networks, Long Short-Term Memory (LSTM), Reservoir Computing and Attention family approaches. Gradient Boosting is a Machine Learning method that proved to be successful in the contributions, and therefore deserved a place in this chapter. Reservoir Computing manages to give good results, but above all it sets unprecedented records on the execution time.

Next, we discussed advances in the time series prediction literature and briefly listed some classical methods such as Dynamic Time Warping, Nearest Neighbor Classifier [70], as well as Deep Learning methods.

Finally, we have seen K-means, a popular clustering method that we have used a lot to meet needs encountered in business during this thesis.

In the next chapter, we will look in more detail at the role that time series can play in predicting buying behavior and at the literature on predicting time series of retail data.

4

Forecasting in Retail

In this chapter, we will briefly review the literature around prediction on retail data in general. Then we will look more specifically at prediction within Colruyt.

4.1/ Sales forecasting in Retail industry

The sales forecasting process is used for estimating future sales. These processes, when accurate, help companies predict all kinds of behaviors and make important business decisions based on the results. Forecasts can be based on past sales data, economic trends, weather, industry comparisons, and more. Established businesses can more easily forecast future sales, which are based on past business data. New companies must base their forecasts on information that is not sufficiently verified, such as market research and competitive intelligence. New companies usually turn to forecasting service providers that are already more accurate but mostly based on aggregated data. Sales forecasting provides insight into the company's workforce, cash flow and resources. Forecasted sales data is crucial for companies to obtain investment capital [104]. There are numerous studies in the literature that model sales data using both straightforward and sophisticated methodologies in order to forecast future sales.

Already in 1996, Ansuj et al. [7] used the AutoRegressive Integrated Moving Averages (ARIMA) model with interventions and the Artificial Neural Network (ANN) model to analyze sales data covering a 10-year period (1979 to 1989). Compared to the ARIMA model, the ANN model's predictions were more accurate. They obtained a residual variation of 0.00227 for the ARIMA model, while the neural network model presented a residual variation of 0.00104. The selected neural network presented, for the last 12 months, better forecasts than the ARIMA model with interventions. The average absolute error of the forecast was 7.6486 and that of the ARIMA model with interventions was 9.8642. The sales time series exhibits a marked seasonality for which it was necessary to use 12 binary units (0 or 1) to determine the relative weight of each month. The model obtained

from the neural network was superior to the ARIMA model in both fit and prediction for the data analyzed.

In 2001, Alon et al. [16] conducted a comparative research between conventional techniques and ARIMA models with ANN models on US aggregate retail sales data. Based on the empirical results, they were able to deduce that for different forecast periods and different forecast horizons, ANN performed best, followed by Box-Jenkins [32] and Winters [3] exponential smoothing. They found that multiple regression with trend and seasonal dummy variables performed the worst. The ANN performed better than traditional statistical methods in the first period when economic conditions were relatively volatile. When macroeconomic conditions were relatively stable, the Box-Jenkins and Winters exponential smoothing models provided viable performance. Multi-step forecasts may be preferred under volatile macroeconomic conditions because new data may not add much useful information to the forecast model. Finally, in their view, the derived graphs show that the ANN model was able to capture dynamic nonlinear trends and seasonal patterns, as well as their interactions.

In 2013, Dwivedi et al [49] did a study that compared a kind of ANN called Adaptive Network-based Fuzzy Inference System (ANFIS) based on Takagi - Sugeno fuzzy inference system [76], with linear regression, neuro-fuzzy modeling and common ANN. The ANFIS method was found to be the most appropriate in their comparison.

In 2017, Aras et al. [69] provide a good overview of the literature and a comparative study on retail sales forecasting of ”an international furniture company, which has operated in Turkey’s retail sector for many years” between methods with different approaches like ARIMA and ARFIMA models, ETS (Error, Trend, Seasonal), Artificial Neural Networks (ANN) and Adaptive Network-based Fuzzy Inference System (ANFIS). According to them, it is almost impossible to know in advance which forecasting model will perform best for a given data set. No single model is best for all situations and circumstances. They emphasized that the term ”best” can be interpreted in many different ways, for example by saying that the best model is the one whose forecasts are both good and not very variable, whatever the time series analyzed. Based on this definition, they concluded that combined methods can be called ”best” in their context and can be used to forecast sales of any product in a retail company. In their view, they generally produce robust forecasting performance regardless of the time series considered. They describe the combination methods by three simple methods (simple average, trimmed, median) and three more sophisticated methods (LS weights, MSE weights, MSE ranks). The simple combination methods performed similarly or even better than the more complicated combination methods in their opinion, a finding consistent with the widely accepted claim in the existing literature.

In 2020, Zunic et al [104] proposed and tested a Prophet-based framework on real data

available on 4TU.ResearchData. The data was obtained from a production environment in one of the largest retail companies in Bosnia and Herzegovina. Prophet [104] itself is built on an additive regression model with four primary components: a piecewise linear logistic growth curve trend; an annual seasonal component constructed using Fourier series; a weekly seasonal component created using dummy variables; and a list of important holidays provided by the user. By evaluating its performance in a real-world usage scenario, their framework demonstrated its ability to generate reasonably accurate monthly and quarterly sales forecasts, as well as great potential for classifying the product portfolio into several categories according to the expected level of forecast reliability: about 50% of the product portfolio (with a sufficiently long history) can be forecasted with MAPE (Mean Absolute Percentage Error) $< 30\%$ on a monthly basis, while about 70% can be forecasted with MAPE $< 30\%$ on a quarterly basis (of which 40% are forecasted with MAPE $< 15\%$). These approximately 40% of the product portfolio that can be forecasted with MAPE $< 15\%$ on a quarterly basis are primarily the subject company's top selling items. MAPE is similar to the Mean Absolute Error MAE, but normalized by true observation.

Retail sales data contain multiple seasonal cycles of different lengths. For example, the daily beer sales data presented in one experiment show weekly and annual cycles. Weekend sales are high, weekday sales are low, summer sales are high, winter sales are low, and Christmas sales are high. Beer sales are high around the date of a major soccer game. Some of the sales data depend on the nature of the business and the location of the businesses.

As can be seen, there are many studies that have shown the effectiveness of ANNs on retail sales data, but before this thesis, no study has applied an attention-based approach to retail sales data. Given the effectiveness of attention-based methods for time series on other data type, it stands to reason that it would be interesting to test it on retail data. This thesis will attempt to address this need.

4.2/ SALES FORECASTING AT COLRUYT FRANCE

In Section 2.3.2, we briefly presented the size and format of the Colruyt France data. However, prior to this thesis, the data from France was not fully exploited. In France, Linear regression techniques are often used by the Sales & Reporting departments to predict overall sales. RFM analysis is also often used by the marketing department to identify the company's best customers based on their consumption habits, but also to highlight the share of revenues coming from new customers, compared to regular customers. Since all data science profiles reside in Belgium, the forecasting via machine learning are usually made on the entire dataset, including the other subsidiaries. Considering the results of the machine learning models compared to the classical approaches

made in Belgium, Colruyt France has decided to apply the machine learning approach on the French dataset. These results then motivated the company to fund a research in this field. This thesis will treat it as the main topic.

4.3/ CONCLUSION

In this chapter, we have reviewed the state-of-the-art on general forecasting in the retail industry. Many studies have been conducted, comparing relatively deep or non-deep neural networks to classical techniques like ARIMA. The neural network based approaches generally gave the best results. There were also studies that tested the effectiveness of Prophet models (a method created by Facebook and built on Fourier series) on demand prediction in retail. According to these studies, the results were conclusive.

Given the performance of neural network based approaches and the fact that attention based approaches have been proven to predict time series in other domains, it would be interesting to test them on retail data. As no study has done so, we have attempted to do so in this thesis.

III

CONTRIBUTIONS

5

Churn detection using Machine Learning in the retail industry

The constant need to increase sales and profitability is a top priority in every company. Indeed, when the existing consumers stop purchasing from the company, its income tends to drop rapidly. For this reason, customer retention has been considered one of the most important issue in Customer Relationship Management, as it has been found to be less expensive than acquiring new consumers. The chance for future sales or even cross-selling is missed when a customer stops going to a particular shop. For such reasons companies have to be proactive and detect potential churners before they leave. This chapter shows how transactional data and Machine Learning can be relevant for the retail industry to forecast churns. To train the Machine Learning models, a sample of 5,115,472 records of consumers with a loyalty card was obtained from Colruyt's data warehouse. The results revealed that the Machine Learning models perform better than classical models.

5.1/ Introduction

Customer retention is a key challenge that is faced across different sectors. Since acquiring new customers is more expensive than retaining them [46], it has become essential for all companies to study the behavior of customers who churn, in order to avoid it in the future.

Customer could churn for several reasons. It can come from bad publicity on social medias or word-of-mouth, from similar products being sold at cheaper prices by the competition or the competition could also have better Customer Service or User Interface/Experience for online purchase [36]. Research reveals that attracting new clients is more expensive than customer retention [57] due to the marketing costs needed to bring new customers. Therefore the preservation of the existing client base has become es-

sential. Customers usually churn gradually and not suddenly, for such reasons, analyzing their historic purchasing patterns [22] could enable companies to detect a drop in their purchasing habits. When companies use loyalty cards, thousands of attribute values are stored for each buyer. Those data include useful knowledge which is often buried in the large array of raw data. It should be noticed that, these datasets contain mainly structured data that can be requested through SQL [13] and semi-structured [9] data such as Excel, JSON and CSV files.

Now, it is widely accepted that Machine Learning manages to perfectly extract hidden characteristics from raw data. Across many different fields, Machine Learning methods have been applied successfully, therefore it could be used to extract knowledge from those raw data.

This study uses a private dataset from Colruyt France, a retail company with 90 supermarkets (700 to 1200 m²), mainly located in the Franche-Comté region in France. This dataset represents purchases of 105,488 customers. These customers have so far produced a total of 5,115,472 rows of data.

The novelty of this study is the application of Machine Learning and Deep Learning Techniques on this type of data. Other main contributions are the possibility to bind personal data (age, length of the client relationship, gender, population of the city where the customer lives) to the sales time-series [31] and the data augmentation technique explained in Section 5.2.3. In the best scenario the model's precision is 75.60%.

5.1.1/ Definition of Churn

Churn is a marketing term that refers to a customer who has switched to a competing company or stopped purchasing from your company. Churn can be defined as customers who are likely to stop transacting with the firm in a given period [57]. It can also be defined as [44]: when the average basket of a customer, namely the average amount spent over a period, falls below a threshold over a predetermined period of time. The average basket is described formally in Equation 5.1.

Let $PurchaseAmount_C(i)$ be the amount spent by a customer C during a week i; and $n = |P|$ the number of weeks of a period P;

$$AverageBasket_C(P) = \frac{\sum_{i=1}^{n} PurchaseAmount_C(i)}{n} \tag{5.1}$$

It is difficult to pinpoint the specific moment when a customer would churn in the supermarket retail industry. Customers do not suddenly stop buying from the store, rather they partly defect. In fact, they tend to switch to a rival gradually [22]. To be more factual, in this context, a churner is defined by the following rule. Let $P1$ be the period of observa-

tion, where the customer's usual purchase amounts are examined. That will be the input to the future churn detection model.

Immediately after $P1$, there is $P2$ which is the period of evaluation, where a change in customer buying habits can be seen. $P2$ will be unknown to the prediction model. Throughout this study, $P2$ is permanently set to 12 to match the requirement of the marketing service. Which means the evaluation period is always 3 months.

The labelling technique will be as follows. If the average purchase during $P2$ is $< 20\%$ of the average purchase during $P1$, then, that customer will be labelled as a churn, if not, he/she will be labelled as a non churner. This 20% is the allowed drop, and is called the reduction factor. It means that the churn could be partial or total. The formal definition of churn is:

Let α be the reduction factor, and C a customer. C is considered as churner if and only if :

$$AverageBasket_C(P1) < \alpha \times AverageBasket_C(P2) \tag{5.2}$$

Some examples of churners and non churners are shown respectively in Figure 5.1 and Figure 5.2.

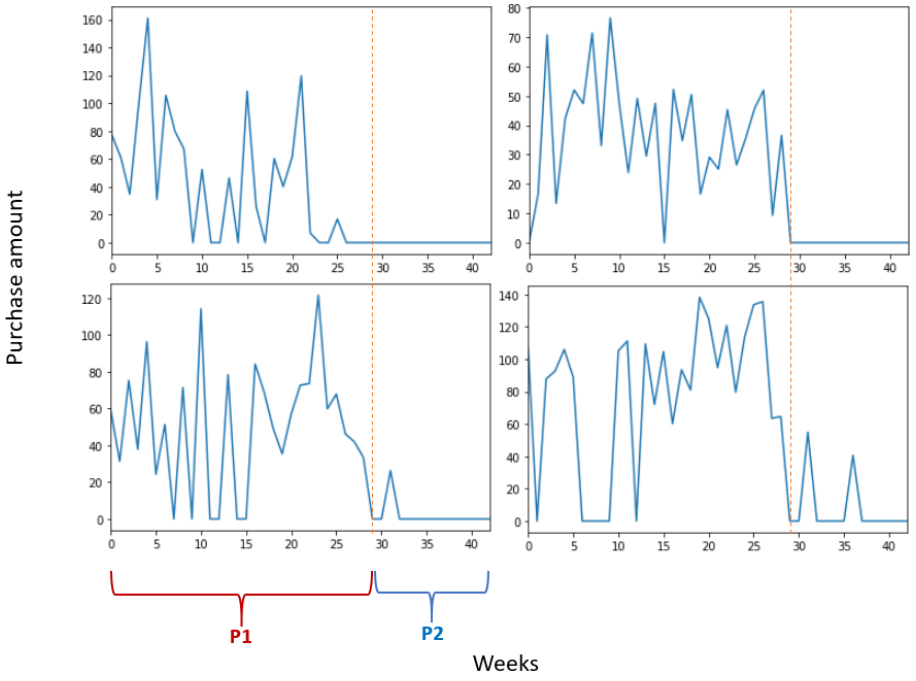


Figure 5.1: Four examples of churners. During period $P1$ they used to buy each weeks, and they have totally or partially stopped buying during $P2$.

From these definitions and observations we can deduce three churn cases.

1. *Simple churn cases* These cases are negative slopes [101] with lower noises, as shown in Figure 5.3. Let S the slope for the time-series values as $\hat{y}_i = S\,x_i + c$. For

each i the error $e_i = |y_i - \hat{y}_i|$ have to respect $e_i \leq \frac{y_i}{n}$, n being the number of periods or the length of the time-series. In other words, the noise does not distort the trend. This kind of churners are relatively easy to detect using a simple linear regression with a threshold.

2. *Churn cases with hidden variables* These cases are similar to the previous ones with the difference that they include hidden variables, namely promotions, soccer championships, weather, or any other external event that impacts the consumer's habit. These hidden variables significantly distort the trend, creating a bias in the

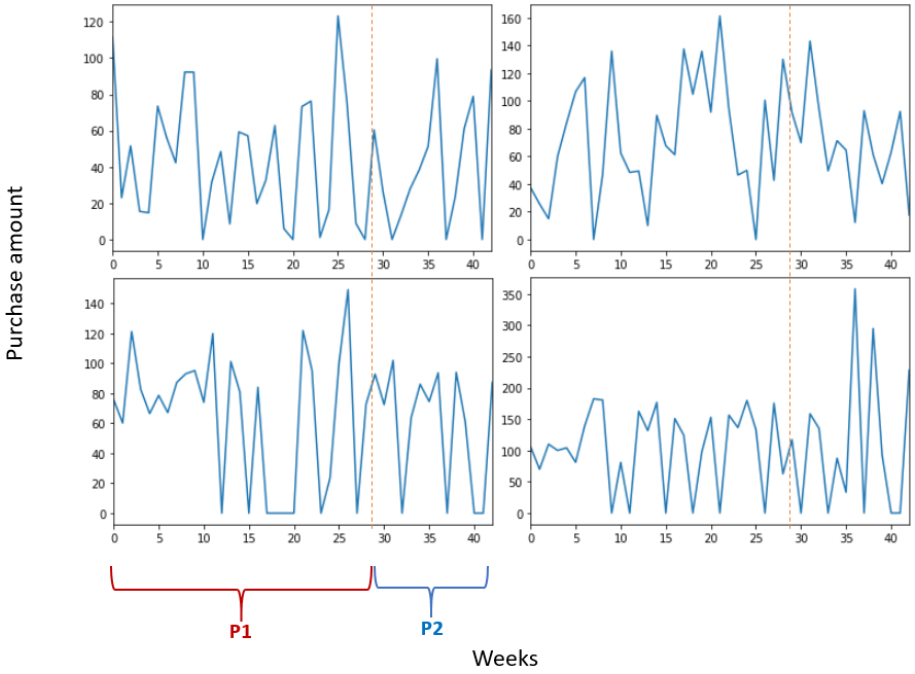


Figure 5.2: Four examples of non churners. They used to buy each weeks, both during $P1$ and $P2$. There was no dramatic change in purchasing habits

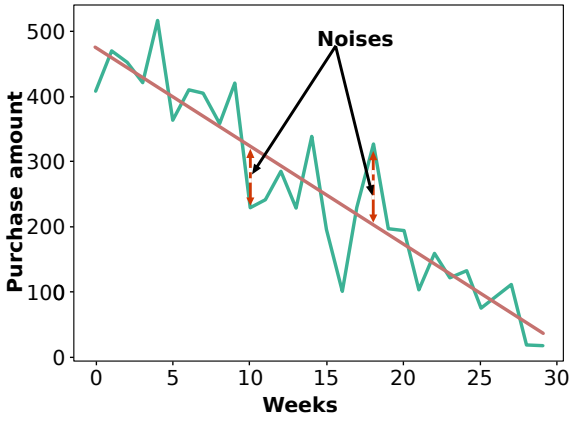


Figure 5.3: Churner with lower noise, here the slope can be seen despite the noise.

classification by a linear regression model. The proposed model for these cases have to be able to read from the input, an extra information about a potential hidden variable. This hidden variable can be formatted as a time-series too.

3. *Difference between churn case and sporadic case* The sales records often include 20% to 30% of customers who come to buy in our stores occasionally. In contrast to regular customers, those who come periodically, i.e. once or several times a week or even every two weeks. The regular customers consider our stores as their main, they will make their biggest purchases there. Even if, when needed they can also make small purchases in a competitor's store closer to them or more easily accessible. Basically, churners are regular customers who are becoming sporadic. This study does not focus on the sporadic cases.

5.1.2/ RELATED WORKS

During the past few years, there have been a lot of research in the field of churn prediction. The fact that customer retention is much more economical than customer acquisition was explained in the work [46]. There has been a lot of studies on churn that compared Machine Learning approaches to classical approaches on data from different domains, namely, telecommunication [93], banking [34] and online [17] subscription. On the other hand, there are just a few studies that specifically target the retail industry, which has certain unique characteristics in terms of consumer interactions and life cycles. RFM analysis has been used for decades in churn detection analyses [82]. Subsequently, it was recommended by some authors that the results were improved by combining RFM with (K-means) methods. In 2017, Dingli [73] compared Restricted Boltzmann Machine (RBM) model to Convolutional Neural Network (CNN) model. At that time RBM achieved the best score with F-measure (β=0.5) of 77% and a Precision of 74%. Since then, convolutional networks have evolved, as have deep networks.

However, the context and the nature of the data can have an effect on the outcome, pushing to use one to privilege one approach over another. For example, in certain domains such as telecommunications [93], the customer rarely subscribes to more than one operator simultaneously. While in retail industry, the consumer may have a main store, where he/she buys periodically, and others where he/she buys sporadically. This study focuses on the retail sector in all its particularities. Also, the data augmentation technique proposed in this study has not been seen in any other similar study.

5.2/ METHODOLOGY

5.2.1/ DATA ACQUISITION

As previously stated, this research includes 5,115,472 rows of sales data, distributed in 105,488 customers from a supermarket chain company in France. A pseudonymization was applied on the entire dataset, which allows to bind personal info (age, length of the client relationship, gender, population of the city where the customer lives) legally to the sales time-series [31] in this research. This can be seen in Figure 5.4. From the raw data, we obtain a data without the sensitive information (which might make someone able to find the original customer without going through a formal identification service). This resulting data can be used for training and testing purposes of classification and forecasting models. Then, the result of these classifications will be returned to the identification service, which will provide it to the marketing department for possible customer recalls.

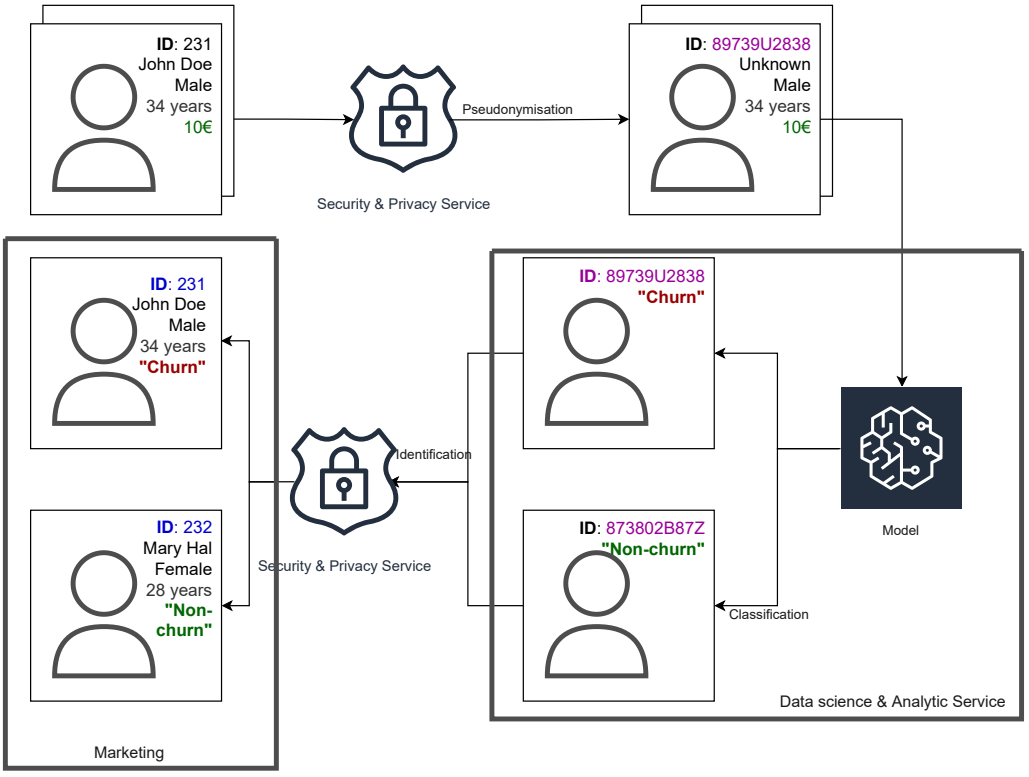


Figure 5.4: The process of pseudonymization and re-identification. These occur respectively, before and after a classification on customer data. This example illustrates a classification case, but it can be generalized to forecasting and other types of analysis.

The rows were split into groups where each group identified a customer and is formatted as a 2 dimensions array. Outliers such as customers with 2 active profiles on the same name and address were removed. The length of $P1$ and $P2$ periods 5.1.1 in weeks was discussed with the Marketing service of our company. Multiple lengths were tested to

finally converge to the ones which produce the highest accuracy. After setting α (the reduction factor) to 0.2, the labelling technique discussed in 5.1.1 was applied.

The dataset was then split into a train and test sets. See table 5.1 for an overview of the dataset.

Table 5.1: An overview of the dataset. Sporadic customers were not considered during this study.

Labels	Values
Number of customers	105,488
Number of sporadic customers	42,636
Total number of rows	5,115,472
Number of customers labelled as Non churner	61,259
Number of customers labelled as Churner	1,593

5.2.2/ HIGHLY IMBALANCED DATASET

One of the main issues with the dataset used for this study was the class imbalance. When there is an imbalance between classes, the learning algorithm basically tends to forecasts only the majority class to minimize the error. For example, in the dataset used for this study, there were 60,000 active customers compared to just 2,000 churners. In other words, 96% of active consumers compared to just 4% of churners, showing a classic case of class imbalance.

5.2.3/ RE-CALIBRATION TECHNIQUES FOR IMBALANCED DATASETS

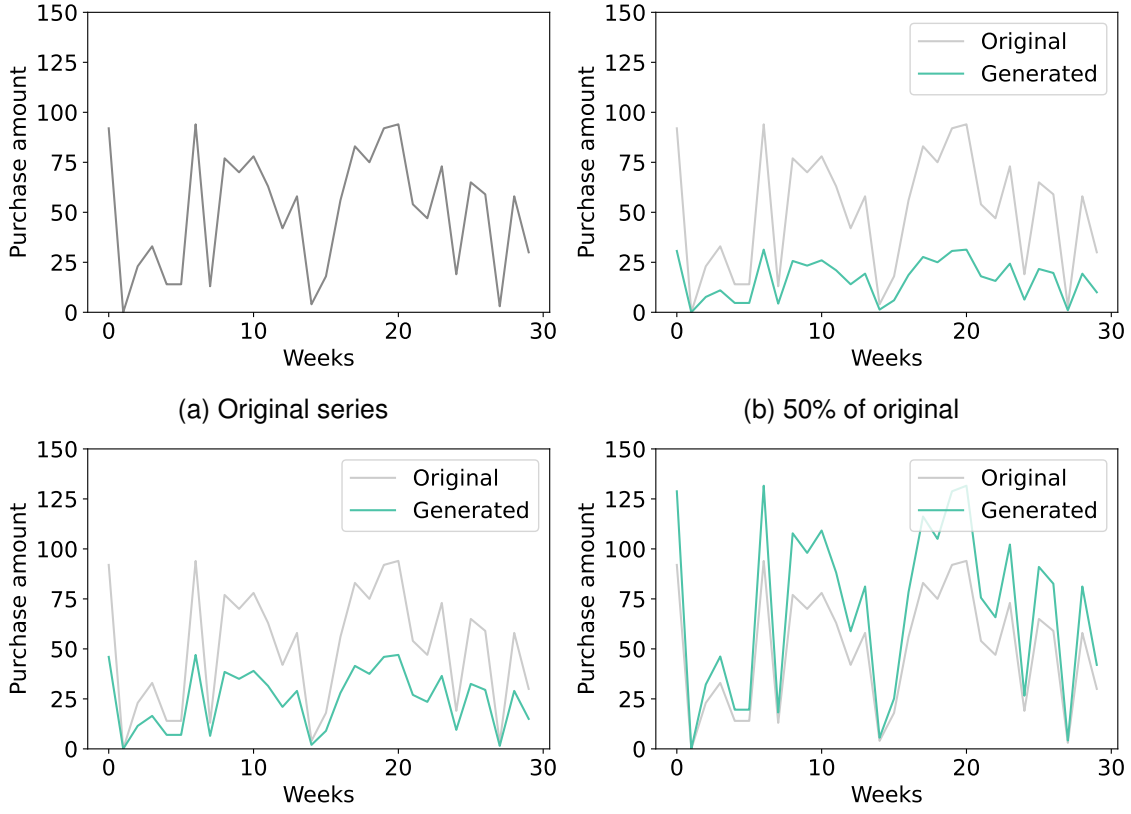
Here are two techniques used to try to rebalance the data in our context.

Data augmentation by scaling It consists in generating other time-series to artificially increase the number of samples in one class by changing the scale of the original time-series. From a customer who always buys around 100 euros every week, we can generate a virtual one who will buy around 50 euros every week and another one who will buy around 200 euros every week. The latter two will have exactly the same dynamics as the first. The idea is to artificially create new customers with the same behavior as the original one, but at a different scale. A representation of original and scaled versions can be seen in Figure 5.5. This is a technique that worked and reduced the imbalance in the dataset studied. The data went from ”highly imbalanced” to just ”imbalanced”. We go from a ratio of 96/4 to a ratio of about 70/30.

Resampling This resampling technique was used to finalize the correction of the imbalance. It is a widely adopted technique for addressing the class imbalance issue. It can be done in two ways: either over-sampling or under-sampling can be used [19]. With under-sampling, just a subset of the majority class is used to train our models. In this study, a random sample of inputs was excluded from the set of active clients, so that the number of churners and the number of non churners becomes equivalent.

5.2.4/ Evaluation Metrics

Precision, recall, and F-measure are used as measurement tools in this chapter to measure the reliability of the various prediction models [38]. TP and FP respectively denote True Positive and False Positive samples, while True Negative and False Negative samples are respectively denoted as TN and FN. The Recall is the proportion of Churn customers that were correctly identified. The recall is intuitively the ability of the classifier to find all the positive samples.



(a) Original series

(b) 50% of original

(c) 33% of original

(d) 140% of original

Figure 5.5: Data augmentation illustration, original data with the generated versions by changing scale. In grey, the original series, and in green the generated versions.

$$Recall = \frac{TP}{TP + FN} \tag{5.3}$$

The Precision is the proportion of the predicted Churners that were correct. The precision is intuitively the ability of the classifier not to label as positive a negative sample.

$$Precision = \frac{TP}{TP + FP} \tag{5.4}$$

F-Measure / F-beta score weighs Recall more than Precision by a factor of β. When β equals 1.0, Recall and Precision are equally important.

$$F_\beta = (1 + \beta)\frac{\text{Precision} \cdot \text{Recall}}{\beta \cdot \text{Precision} + \text{Recall}} \tag{5.5}$$

In this context, the goal is to get as many TN as possible without spamming a lot of customers. The TP is necessary but not the priority. Which leads us to use the F-beta score metric, with β equal to 0.5.

5.2.5/ HYPERPARAMETERS

In this section, the hyperparameters [89] configurations used will be presented along with the reasons for these choices.

The MLP was $200\times200\times1$ in size, with a dropout of 0.2 separating each layer and a Dense layer of size 1 at the output. A range of (10^{-10} to 10^{-2}) was chosen to vary the learning rate, and the model learned better with a learning rate of 10^{-3}. Optimizers such as SGD, Adam, and Nadam have been tested for each of the neural network models. NAdam is an extension of Adam explained in Section 3.2, it adds the Nesterov accelerated gradient (NAG) or Nesterov momentum, which is an improved type of momentum [91]. NAG is an extension to momentum where the update is performed using the gradient of the projected update to the parameter rather than the actual current variable value. This has the effect of slowing down the search when the optima is located rather than overshooting, in some situations.

The XGBoost had 50 estimators, with a “Logistic regression for binary classification” as objective function. A range of (5 to 50) was chosen to vary the maximum depth, and the model learned better with a maximum depth of 10.

For the LSTM model, there were 3 stacked layers of LSTM, with a dropout of 0.2 separating each layer and a Sigmoïd activator at the output. The model learned better with a learning rate of 10^{-7}.

5.3/ EXPERIMENTAL RESULTS

In this study we used cross-validation for a more effective evaluation and comparison of the models. Cross-validation [39] is frequently used in the evaluation of regression and classification models. Applying it to the time-series or other naturally ordered data adds some complexity because of the chronology of events. Two techniques can be considered.

The first one is to define a time interval, where data should be retrieved for all customers, then apply the k-fold technique [33]. To do so, the dataset must be divided into k equal parts, called folds, where $k - 1$ folds will be the training set and the remaining fold will be the test or validation set. Repeat it k times with different remaining fold as test set each time. The final score averages the validation results at the end of each iterations. In this context, several k were tested, the best result was obtained with k = 4.

The second technique consists in choosing 2 time intervals where to extract the data. There will be two subsets then, and each of them will be divided in two. Thus, four subsets are obtained. On these subsets the k-fold technique can be used. It will therefore be called 4-fold.

In the following, some Machine Learning approaches — MLP (Multilayer perceptron), Extreme Gradient Boosting, LSTM — that have been proven to work for churn prediction in general, was compared with a Linear Regression model as baseline. See Section 3 for an overview of these models. Efficiency and reliability were considered during the comparison.

Linear regression is known in statistics as a linear approach to modeling the relationship between a scalar response and one or more explanatory variables (also called dependent and independent variables) [4]. Linear regression is a lightweight statistical tool useful for gaining insights into customer behavior, understanding the business and factors influencing profitability [45]. Although linear regression has limited applicability in the business world, as it can only work when the dependent variable is continuous in nature, it is still a well-known technique in situations where it can be used [25].

The dataset which includes time-series with some additional information for each customer, was used to evaluate the performance of the tested classifiers. As previously exposed, it contains 61259 samples labelled as loyal customers and 1593 labelled as churners, for a total of 62,852 samples.

Linear Regression It can be seen in tables 5.2 and 5.3 that the results of the linear regression are unsatisfactory, even if there are some individuals for whom the churn is detectable by simple linear regression as mentioned in Section 5.1.1-1. , it is unfortunately

not the case for the majority. It is worth noting that the predictions are slightly better when the model is given a longer period at input. Table 5.2 shows results with $P1 = 8$ and Table 5.3 shows results with $P1 = 30$. With a longer period at input, the accuracy of the results suddenly becomes sensitive to the variation of the threshold. A period of 8 weeks was chosen at the beginning as input, to match the requirement of the marketing service. But, by varying the parameters in a naive way, and by looking for a better accuracy in the linear regression's predictions, parameters converged to a 30 weeks period and a threshold of 0.4. With these parameters we obtain a precision of 67%.

Table 5.2: Precision and F-measure for the linear regression model with $P1 = \mathbf{8}$ weeks predictions. (Highest values in bold, the couple (Precision/F-measure) with the largest values is also in bold)

Threshold	Precision (%)	F-measure β=0.5 (%)
-1.0	56.09	55.27
-0.8	55.72	54.50
-0.6	55.43	53.85
-0.4	56.04	54.05
-0.2	55.97	53.47
-0.0	62.34	**57.52**
0.2	63.07	57.51
0.4	63.36	57.18
0.6	**63.78**	**57.05**
0.8	63.44	56.34
1.0	63.52	55.87

MLP The results in Table 5.4 show that using 200 neurons in the hidden layer captured the slope information including noise, necessary for a good classification. Beyond 200 neurons, there is only a slowing down during training without a real improvement of the predictions. The model then reaches its best prediction at 10 epochs of training with, with an Nadam optimizer. The best result was with a precision of 73.30% and an average F-measure (β=0.5) of 72.21%, and starts to overfit beyond that.

XGBoost The results in 5.5 show that using 50 estimators with a logistic regression for binary classification as objective function, the XGBoost model was able to capture the slope information with the noise slightly better than the MLP. By varying the maximum depth and the number of estimators, it was observed that beyond 70 estimators and a maximum depth of 15 there was only a decrease in the prediction accuracy. The best prediction (precision $= 73.63\%$, F-measure $= 74.45\%$) was achieved with 50 estimators and a maximum depth of 10.

Table 5.3: Precision and F-measure for the linear regression model with $P1$ = **30** weeks predictions. (Highest values in bold, the couple (Precision/F-measure) with the largest values is also in bold)

Threshold	Precision (%)	F-measure β=0.5 (%)
-1.0	64.19	**65.27**
-0.8	64.70	64.87
-0.6	65.94	64.61
-0.4	**67.19**	**64.13**
-0.2	67.52	62.40
-0.0	66.66	58.90
0.1	67.40	58.24
0.4	**67.84**	54.02
0.6	67.47	50.62
0.8	65.48	44.93
1.0	66.51	42.60

LSTM It was determined that the LSTMs outperformed the other detection approaches in this study. Precisely because LSTM was designed to be able to learn order dependence in sequence prediction problems such as analyzing customer sales data over time to detect potential churners. The best prediction (Precision = 73.70%, F-measure = 75.60%) was achieved with 3 stacked layers of LSTM, with an Nadam optimizer.

5.4/ CONCLUSION

The application of Machine Learning models on customer sales data for churn prediction has been discussed in this chapter. A total of 4 statistical and Machine Learning models have been compared in this study, focusing on predictive performance. With the reputation that precedes Machine Learning models, it is conceivable to guess that they might outperform conventional approaches, but this study explicitly shows the gap in Section 5.3. The re-calibration techniques improved the dataset balance, allowing the models to achieve these precisions.

From this study, the company can better anticipate the reaction of churn customers. These customers who are likely to stop buying in its stores. The company will also be able to think of marketing campaigns adapted to these customers.

Future studies may use other methodologies such as Transformers [77]. Since recent Machine Learning models have proven to be better at handling more data, the exploitation of external data such as weather, neighborhood details, average per capita income, will be of major interest.

Table 5.4: Precision and F-measure (averages) for MLP ($200 \times 200 \times 1$) predictions. (Highest values in bold)

Epochs	Resampling	P1 length	Precision (%)	F-measure β=0.5 (%)
10	Yes	8	70.01	70.60
10	Yes	30	**73.30**	**72.21**
10	No	8	50.04	55.60
10	No	30	52.40	56.51
50	Yes	8	68.80	**70.40**
50	Yes	30	**69.67**	69.38
50	No	8	50.02	55.60
50	No	30	52.20	57.70

Table 5.5: Precision and F-measure (averages) for eXtreme Gradient boosting (50 estimators and a maximum depth of 10) predictions. (Highest values in bold)

Epochs	Resampling	P1 length	Precision (%)	F-measure β=0.5 (%)
10	Yes	8	71.14	71.22
10	Yes	30	**73.63**	**74.45**
10	No	8	52.01	54.77
10	No	30	53.73	56.92
50	Yes	8	69.25	70.68
50	Yes	30	**71.89**	**70.71**
50	No	8	51.31	55.78
50	No	30	52.66	58.07

Table 5.6: Precision and F-measure (averages) for LSTM ($100 \times 50 \times 25$) predictions. (Highest values in bold)

Epochs	Resampling	P1 length	Precision (%)	F-measure β=0.5 (%)
10	Yes	8	70.86	70.92
10	Yes	30	**73.70**	**75.60**
10	No	8	51.87	55.98
10	No	30	52.78	57.32
50	Yes	8	69.17	70.42
50	Yes	30	**72.31**	**71.28**
50	No	8	52.14	55.91
50	No	30	53.41	58.53

6

Overstock prediction using Machine Learning in retail industry

Success in supply-chain relies on good stock management. It is quite simple to guess that there will be an increase in demand for a type of product, or rather reluctance over a period of time, but it becomes complicated to know in advance the exact or optimal number of products to order to avoid stock-outs and at the same time overstocking. This chapter shows how transactional data can be used with Machine Learning to forecast demand in the retail industry. To train the Machine Learning models, a sample of records of receipt data was obtained from the largest Belgian supermarket chains's data warehouse. The results revealed that the Machine Learning the models manage to learn the seasonality effect and make better predictions.

6.1/ Introduction

Forecasting demand is one of the biggest challenges in the retail industry, which essentially answers the question: What is the probability distribution of the demand for an product or product family, over a given time period starting from a date in the future? Among its numerous advantages, a predictive forecast is a crucial enabler for lower costs thanks to better planned inventory and fewer write-off items, as well as for a better customer experience by reducing out-of-stock situations [71]. Manufacturers, distributors, retailers and other businesses are constantly looking for more accurate predictions to reduce uncertainty in decision-making. Accurate demand forecasting [24], especially in retail, leads to well-informed purchasing, inventory management, scheduling, capacity management, and assortment planning decisions. A common definition of demand forecasting refers to the practice of estimating future customer demand over a predetermined period of time

using historical data and other information. [71].

The most often used methods to forecast demand try to identify seasonality and trend in time series, and create correlations between the variable of interest and other independent variables.

However, these methods struggle to perform effectively when the dependent variable also depends on external variables. The techniques presented in this document manage to deal with this situation.

It is now, widely accepted that Machine Learning manages to perfectly extract hidden characteristics from raw data. Across many different fields, Machine Learning methods have been applied successfully. It could therefore be used to extract knowledge from those raw data.

This study uses a private dataset from Colruyt France, a retail company with 90 supermarkets (700 to 1200 m^2), mainly located in the Franche-Comté region, France. This dataset represents only in-store purchases, i.e. sales receipts. The dataset does not directly concern information related to the loyalty card and it focuses on two major product families, the dairy family which includes (milk, cheese, yogurt, ice cream, ...) and the fish family which includes products derived from fish.

The novelty of this study is training Machine Learning and Deep Learning models on this type of dataset with a limited number of samples.

In the best scenario the model's Mean Squared Error (MSE) is equal to 0.43. This result was obtained with an Autoformer model.

Demand is forecasted based on past time series data, which means that demand prediction is basically a time series forecasting problem. The state-of-the-art of time series forecasting has been presented in Section 3.1.2.

Overstock in retail industry, usually means having too much stock in a store that has not sold. One consequence of overstocking in some supermarkets is that the latter is obliged to make a special promotion called 'anti-waste' on products to make them more attractive. In anti-waste promotions, the company agrees to sell excess products at a loss just before their expiration date instead of throwing them away. This is often the case with short-life products. Even an anti-waste promotion cannot prevent some products from ending up in the trash because customers were unable to purchase all of the discounted products before their expiration date. Anti-waste is very different from traditional promotion which is more about attracting new customers.

The remainder of the chapter is structured as follows. Section 6.2 explains the methodology, from data acquisition to evaluation of models. Section 6.2.1 details the different models used and Section 6.3 presents the obtained results and compares the different

techniques used to tackle the problem. Finally, Section 6.4 summarizes the conclusions drawn from the chapter.

6.2/ METHODOLOGY

This research includes Colruyt sales data from the year 2017 to 2022. In Table 1, there is a representation of the structure of the dataset in its raw state.

Table 6.1: A sample of the input dataset

product_ID	date	quantity	unit_price
229490008301	2015-02-02	738	1.2363
229490008301	2015-02-04	366	1.0368
229490008301	2015-02-05	521	1.1131

The product ID includes information about the category it belongs to. Therefore, the products were grouped into categories. Two main categories were experimented in this study. These two main product families were the ones that usually generate the most anti-waste. Namely the dairy family which includes (milk, cheese, yogurt, ice cream, . . .) and the fish family which includes products derived from fish. See Table 6.2 for an overview of the dataset.

Table 6.2: An overview of the dataset.

Labels	Values
Number of products	150
Time period	2017 - 2022
Granularity	days

6.2.1/ MODELS

In the following, we compare some Machine Learning approaches that have been proven to work for time series forecasting in general. We consider models based on attention, such as Autoformer, Transformer and Informer, with the models that gave the best results in the previous chapter, namely XGBoost and LSTM.

Detailed description of such models can be seen in section 3.1.2.

6.2.2/ HYPERPARAMETERS

In this section, the hyperparameters [89] configurations used will be presented along with the reasons for these choices.

For the XGBoost, we have used the Scikit-learn [37] library and its hyperparameter tuning function which considers the cross validation, to automate the search for the best configuration. We provided this function with the splitting of the time series which common to all the tested models. The search ranges were $[0.005, 0.05]$ for the learning rate, $[5, 30]$ for the maximum depth, $[50, 1000]$ for the number of estimators. The configurations then converged to a learning rate of 0.03, a maximum depth of 100.

The LSTM model as well as the other attention-based models took advantage of an automatic adjustment of the learning rate and an early stopping system. The best of LSTM models, contained 8 layers of LSTM with 16 hidden states, a dropout equal to 0.25. The search ranges were $[1, 10]$ for the number of layers, $[4, 32]$ for the hidden state dimension, $[0.2, 0.5]$ for the dropout. The learning rate started at 0.1 and could be adapted during training, it could go down to 10^{-10}.

The attention-based models, because of their similarity to each other, were able to benefit from the same search ranges overall. Some of them still gave better results on some ranges than on others. The search ranges were $[4, 35]$ for the number of heads, $[2, 5]$ for the number of encoding and decoding layers, $[10, 25]$ for the moving average, $[1, 3]$ for the attention factor. The activation functions Rectified Linear Unit (ReLU) [81] and Gaussian error linear units (GELU) [66] have been tested. The models gave better results with the GELU activation function.

6.3/ EXPERIMENTAL RESULTS

Mean Square Error (MSE) and Mean Absolute Error (MAE) are used as metrics in this chapter to measure the reliability of the various prediction models [52]. When these metrics are used on scaled outputs, it helps to reduce the bias that can occur when comparing results between different models.

Cross-validation [39] and Forward-chaining [99] are used to evaluate the models in this chapter. Cross-validation [39] is frequently used in the evaluation of regression and classification models. Applying it to the time-series or other naturally ordered data adds some complexity because of the chronology of events. To prevent data leakage when working with time series data, special care must be taken when splitting the data. To accurately simulate the “real world forecasting environment, in which we stand in the present and forecast the future” [15], the forecaster must withhold all data about events occurring

chronologically after the events used to fit the model. Rather than using k-fold cross-validation, we use hold-out cross-validation for time series data, in which a subset of the data (split temporally) is reserved for validating model performance. As shown in Figure 6.1, the test set data follows the training set chronologically. Similarly, the validation set follows the training subset chronologically.

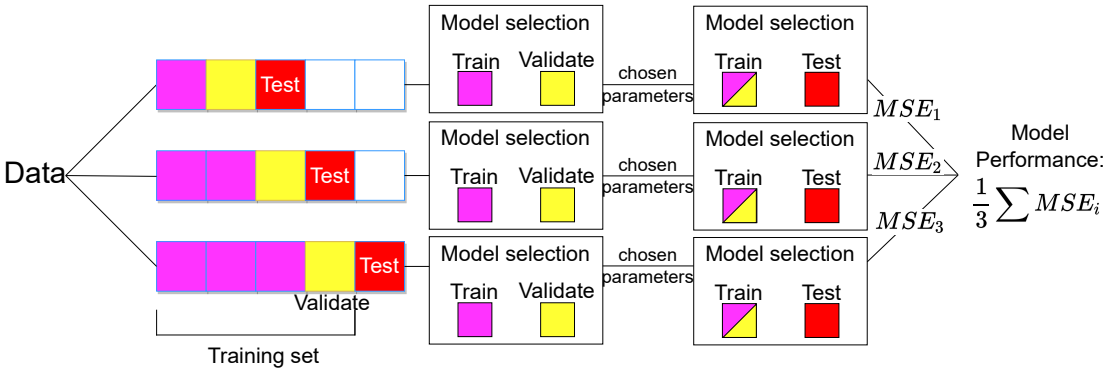


Figure 6.1: Forward-chaining illustration. Creating many train/test splits and average the errors over all the splits to produce a better estimate of model prediction error.

The dataset used to evaluate the performance of the tested forecasting models contains daily purchases of 150 products divided into 2 main categories.

Table 6.3: Results on two product families with predicted length as 1, 7, 30, 60. (Lowest values in bold)

	Models	Autoformer		Informer		Transformer		XGboost		LSTM	
	Prediction length	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Fish	1	**0.43**	**0.36**	0.50	0.42	0.54	0.43	0.67	0.47	0.59	0.45
	7	**0.52**	0.40	0.53	0.39	**0.52**	**0.39**	0.60	0.45	1.08	0.77
	30	**0.53**	0.44	0.69	0.48	**0.53**	**0.38**	0.71	0.50	1.67	0.99
	60	**0.55**	0.45	0.83	0.58	0.56	**0.40**	0.70	0.50	1.79	1.03
Dairy	1	**0.43**	0.37	0.44	0.39	0.47	**0.36**	0.68	0.46	0.54	0.44
	7	**0.48**	**0.37**	0.49	0.38	**0.48**	0.38	0.60	0.46	0.98	0.73
	30	0.52	0.45	0.64	0.46	**0.48**	**0.36**	0.72	0.51	1.52	0.94
	60	**0.51**	0.43	0.78	0.56	0.53	**0.39**	0.75	0.53	1.64	0.98

LSTM

The results in Table 6.3 show that LSTM captured the patterns but encounter difficulties against noise. All of the tested models were able to understand the seasonality in the data on a global scale. However, when trends are combined with seasonality, the LSTM encounters difficulties, as shown in Figure 6.2. On the product category A, the LSTM

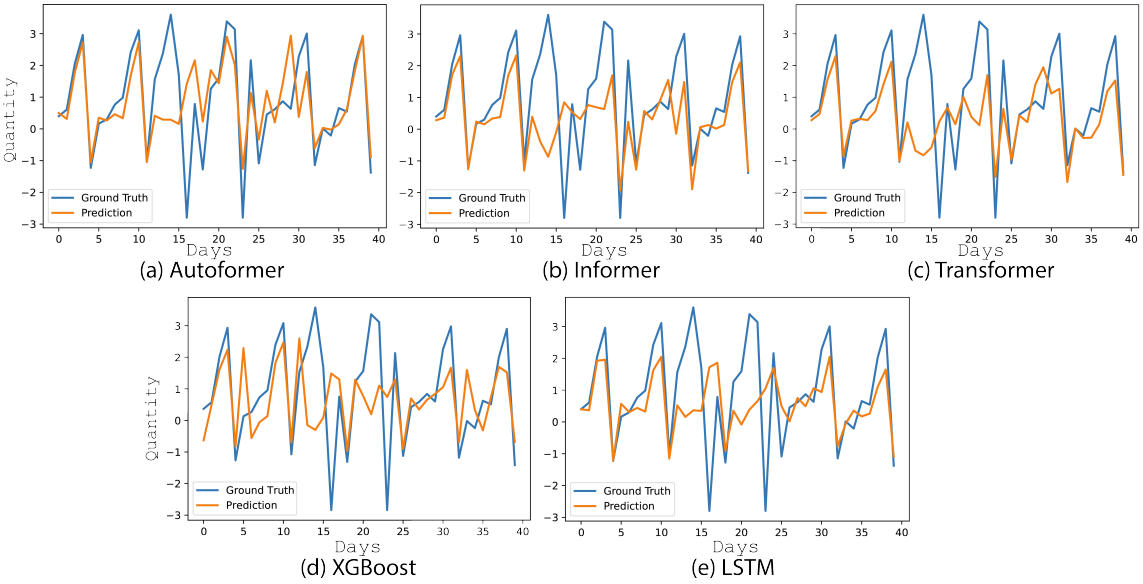


Figure 6.2: Prediction cases from the Product category A dataset under the input-12-predict-1 setting

model reaches its best prediction with $MSE = 0.59$ and $MAE = 0.45$. The number of epochs has been limited to 100 with an Early stopping [94], an Auto adjust learning rate [109] configuration enabled, and an ADAM optimizer.

XGBoost

The results in Table 6.3 show that using 100 estimators with a "Regression with squared loss" as objective function, the XGBoost model was able to capture some patterns. By varying the maximum depth and the number of estimators, it was observed that beyond 200 estimators and a maximum depth of 25 there was only a decrease in the prediction accuracy. The best prediction $MSE = 0.60$ and $MAE = 0.45$ was achieved with 200 estimators and a maximum depth of 20.

Autoformer vs Attention family

It was determined that Autoformer globally outperformed the other models in this study. Precisely because with the series decomposition blocks, Autoformer can aggregate and refine the trend-cyclical part from series progressively. It was also designed to facilitates the learning of the seasonal part, especially the peaks and troughs. This verifies the necessity of the decomposition architecture. The best results of the Autoformer, Informer and Transformer were obtained with 8 heads when predicting on 1 day, and 30 when predicting on 7 to 60 days, with 3 encoding layers and 2 decoding layers, an ADAM

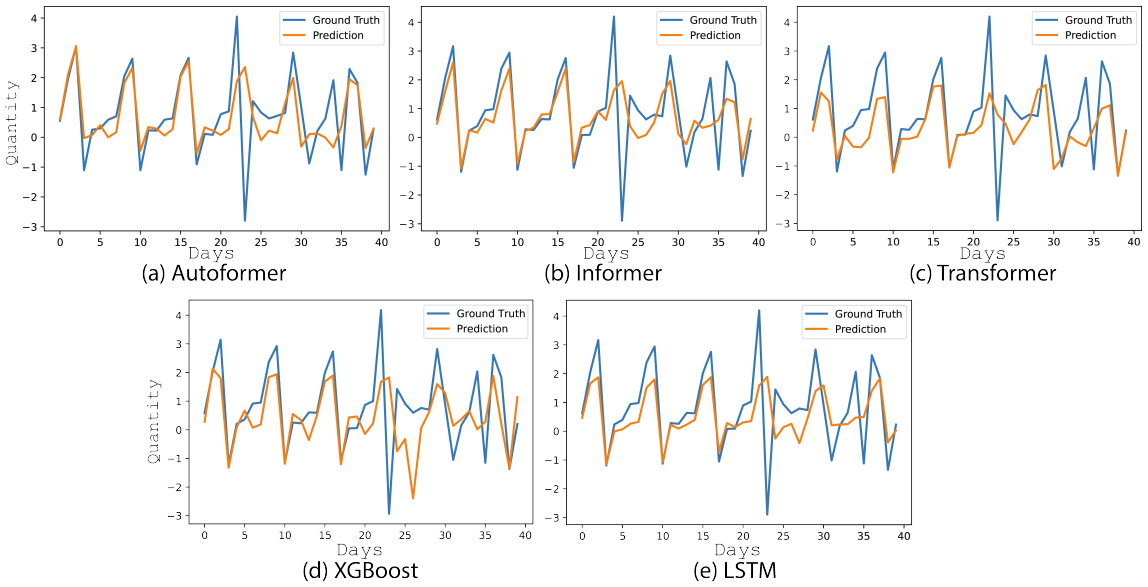


Figure 6.3: Prediction cases from the Product category B dataset under the input-12-predict-1 setting

optimizer, a moving average equal to 13, an attention factor equal to 3. This confirms the observation made in [108] — datasets with an obvious periodicity tend to perform better with a high factor. Full details about the comparison can be found in Table 6.3.

6.4/ CONCLUSION

In this chapter we have seen the application of Machine Learning models to predict overstocking. We addressed the overstocking prediction problem, which can be formulated as a demand forecasting problem. We compared 5 prediction approaches, including Deep Learning approaches such as LSTM, Autoformer, Informer, Transformer and a Machine Learning approach, namely XGBoost. Since LSTM and XGboost performed well in the previous chapter, they were taken up and re-adapted for the regression problem and then compared to attention-based models. The best results with the MSE metric were generally observed with the Autoformer and the best results with the MAE metric were generally observed with the Transformer.

In next chapter, we use other external and public data such weather to increase accuracy.

7

IMPROVED DEMAND FORECAST WITH WEATHER DATA

This chapter presents research on feeding different Machine Learning models (Gradient Boosting, LSTM, Autoformer,...) with weather forecasts. The goal here is to use weather forecasts that are more than 90% accurate within 5 days to improve the prediction of demand for some products which are considered as weather sensitive. These are products whose quantities purchased by consumers vary considerably according to the weather. In each case tested, the modified algorithms always perform better than their original version. We have also added to the previously tested models a new model that has caught our attention thanks to its learning speed. This is the Reservoir Computing [26].

7.1/ INTRODUCTION

As seen in the Chapter 6, demand forecast is very essential in the retail industry, and every retail company is looking for techniques to increase its accuracy as much as possible. In this chapter we will use the weather forecasts to try to achieve this.

To minimize the discrepancy between weather forecasts and actual temperatures experienced by consumers, this study was conducted by city. The weather forecast of the city where a store is located is used as an external feature to feed the demand forecast model. For this experiment, we mainly selected weather sensitive product families.

The novelty of this study is the possibility to bind future predicted weather to the sales time-series [31] as explained in Section 7.2 and train Machine Learning models. A model has also been added to the list of models seen in the previous chapter, it is the Reservoir Computing [26].

The remainder of the chapter is structured as follows. Section 7.2 explains the methodology we adopted, from data acquisition to models evaluation. Section 7.3 presents the obtained results and compares the various techniques used to tackle the problem. Finally,

Section 7.4 provides a summary of the insights gained from this chapter.

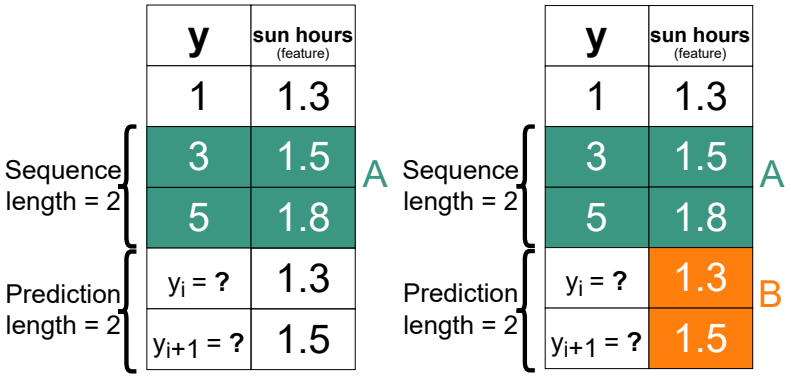
7.2/ METHODOLOGY

This experiment was mainly aimed at 2 groups of products that are sensitive to the weather. The first group includes products often purchased in summer for barbecues (meat, beer with or without alcohol, rosé wines and aperitifs). The second group is made up of products in demand in winter (cheeses, foie gras and potatoes). The input data format of the models remains the same as in the previous chapter.

As illustrated in Figure 7.1, there is the default forecasting technique and the edited version. In the following lines, we will see in detail what the edited version consists of.

By default, when we need to forecast a value for tomorrow, we use today's and previous days values with the extra features as input. The weather can be used as an external feature, it will work like any other external feature, but it can be used for much more since the weather can also be considered as a known future (with a low risk of error when not far in the future). The idea behind this work is to use this known future to improve predictions at dates further than one day. Let Sequence length be the number of days we go back in history, and Prediction length be the number of days the model will predict.

In the Figure 7.1a, only "A" cells are used as input to predict y_i and y_{i+1}, but in the edited version Figure 7.1b, "A" and "B" cells are used to predict y_i and y_{i+1}. The sun hour represents the number of hours of sunshine per day. These weather data are available on Open Weather Map website.



(a) Default (b) Edited

Figure 7.1: Default forecasting technique version and the edited version. The edited version takes into account future data known during training and testing. Here, "sun hours" is the used weather feature. By default, the models consider just the part A on the figure, to predict yi and $yi + 1$, the idea here is to allow the models to consider also the part B in orange to predict yi and $yi + 1$. This will be done during the training phase as well as during the test.

Figure 7.2 illustrates the conceptual difference between the predictions made by a default model and a modified model. The example in this figure shows sales prediction of a product family that is strongly influenced by weather. The modified prediction model will try to take into account the weather changes that will take place at a time interval $[i, i+x]$, in order to make a prediction at a time $i+x$. Here, i is the current date, and x is the number of days predicted after i. The default model will instead stop at i. In a case where the weather does not fluctuate much, or if the fluctuation is not new to the model and has already been learned during training, the default model may perform as well as the modified model. This unfortunately does not happen often, hence the interest of the modified model.

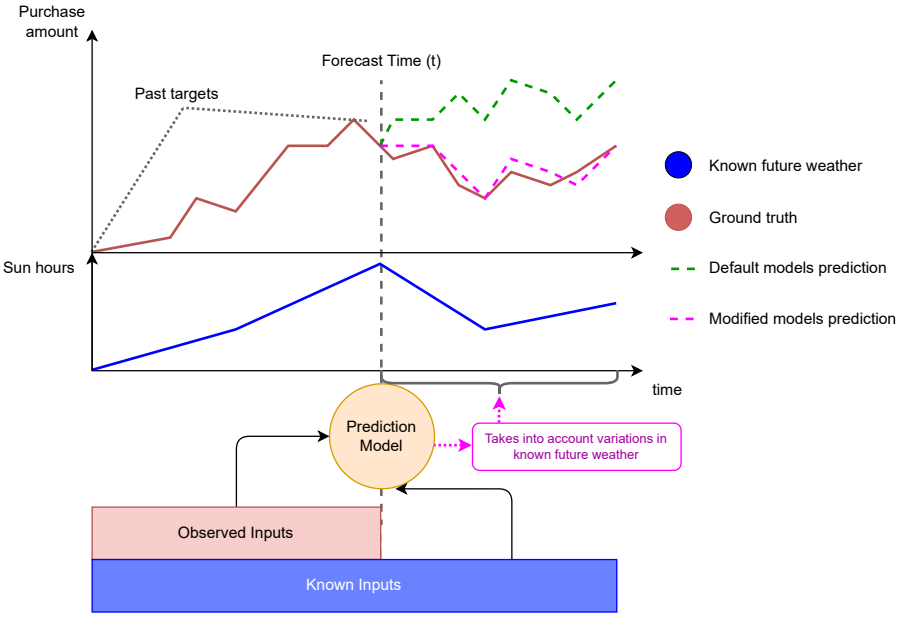


Figure 7.2: MSE by Prediction length

This study is based on an implementation of models from the Attention family [108], including Autoformer, Informer, Transformer, and the Reservoir Computing model. The source code of Haixu Wu et al. [108] has been reused and edited to add the ability to associate future weather data with the application history. Weather is a special kind of feature because it is also a prediction with a risk of error. The average error between predicted and measured weather has been calculated throughout this study. Nowadays, a five-day forecast can predict the weather with an error of about 10% and a seven-day forecast can predict the weather with an error of about 20%. In contrast, a 10-day or longer forecast is accurate only half the time.

7.3/ Experimental Results

In a first step, we used the measured values of the weather to train the model. This is equivalent to considering that the error of the weather prediction is insignificant, which is not true, but it simplifies the problem in a first step. In fact, the forecast weather data are not recorded by default. The forecasts are constantly updated and replaced by the measured values as they come in. The available past datasets then contain measured values. These datasets were used in this research to test the edited model and this quickly gave very impressive results. In parallel a robot was developed during this study to start recording the predicted values on a daily basis to train the model with. The results were slightly less good as expected but still very interesting.

As it can be seen on Figure 7.3, for the improved version, the MSE remains preserved and degrades very slightly over time, while it can be seen that for the default version, as soon as we go to a Prediction length greater than 1, the MSE increases considerably.

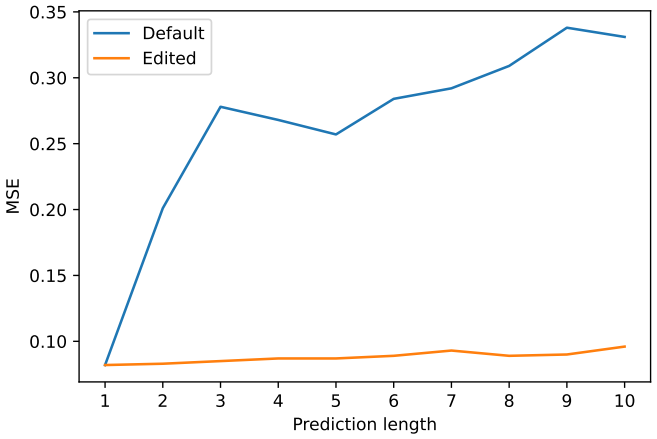


Figure 7.3: Average MSE per prediction length for the edited and default Autoformer models. The MSE remains preserved for the edited version and degrades slightly over time, unlike the default version.

It is worth noting that when the Prediction length equals to 1 day, the behavior is the same for the default version and the modified version. They work exactly the same way and have the same result. The gap starts to widen from Prediction length = 2.

The results of this comparison can be seen in detail in Table 7.1. It shows the results of the default models and the edited models that manage to consider the known future data during training and testing. Figures 7.4, 7.5, 7.6, 7.7 show the predictions of the Autoformer models (the default and edited version), for different Prediction lengths.

It can be seen from the two tables that the results of the Reservoir Computing model are very close to those of the Autoformer. Even if it does not overcome the Autoformer, it deserves its place in this table because of its learning speed. The average training time

required for Reservoir Computing models to achieve these results on this dataset was around 15 seconds. It takes about 8 mins for the Autoformer (being itself known for its training speed compared to its predecessors) to converge to its best results.

Table 7.1: Comparison table of default and edited models. Edited models incorporate known future data in their predictions. The known future here is the weather forecasts.

Models		Autoformer		XGboost		LSTM		Reservoir Computing	
	Prediction length	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Default	1	**0.13**	**0.21**	0.17	0.22	0.15	**0.21**	0.14	0.24
	4	**0.14**	**0.26**	0.17	0.29	0.17	0.28	0.15	0.26
	7	**0.17**	**0.27**	0.21	0.32	0.2	0.3	0.18	0.28
	14	**0.23**	**0.3**	0.28	0.32	0.27	0.33	0.24	0.31
Edited	1	**0.13**	**0.21**	0.17	0.22	0.15	**0.21**	0.14	0.24
	4	**0.14**	**0.22**	0.15	0.23	0.15	0.23	0.14	0.24
	7	**0.14**	**0.25**	0.16	0.26	0.16	0.25	0.15	0.25
	14	**0.14**	**0.25**	0.16	0.27	0.16	0.27	0.15	0.26

7.4/ CONCLUSION

In this chapter we have seen how to improve overstock prediction by editing the default behavior of some forecasting models framework. By adding the known future weather data, it was found that the results considerably improved. We tested predictions from $i + 1$ to $i + 14$ (i being the day when we position ourselves to run the prediction) with various types of models. The modification really impacted the results. The performances of the modified models start to degrade from $i + 10$. In future work, we intend to perform further experiments and test the models with several external features while avoiding overlearning as much as possible.

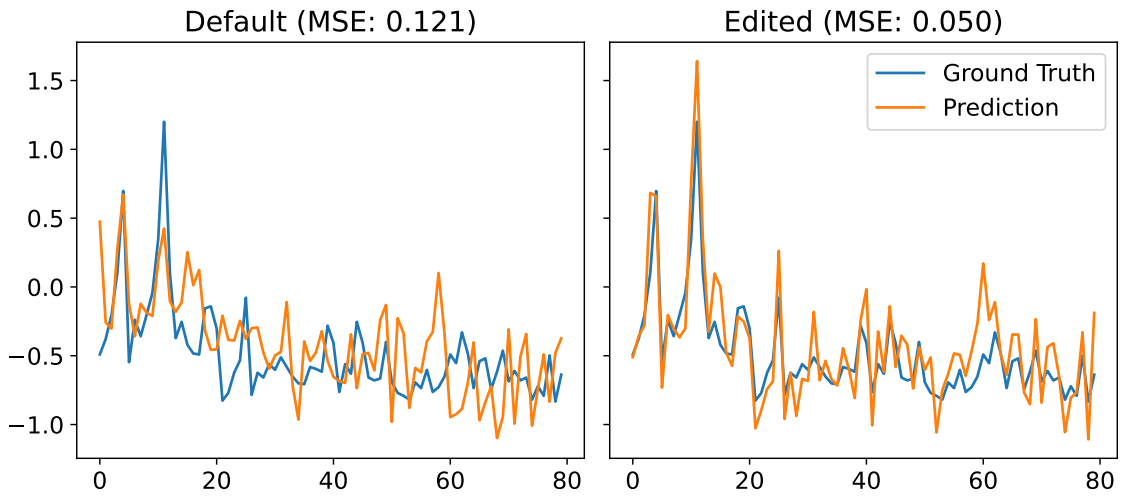


Figure 7.4: Result with Prediction length = 2, x-axis represents the days and y-axis represents the scaled purchase amounts

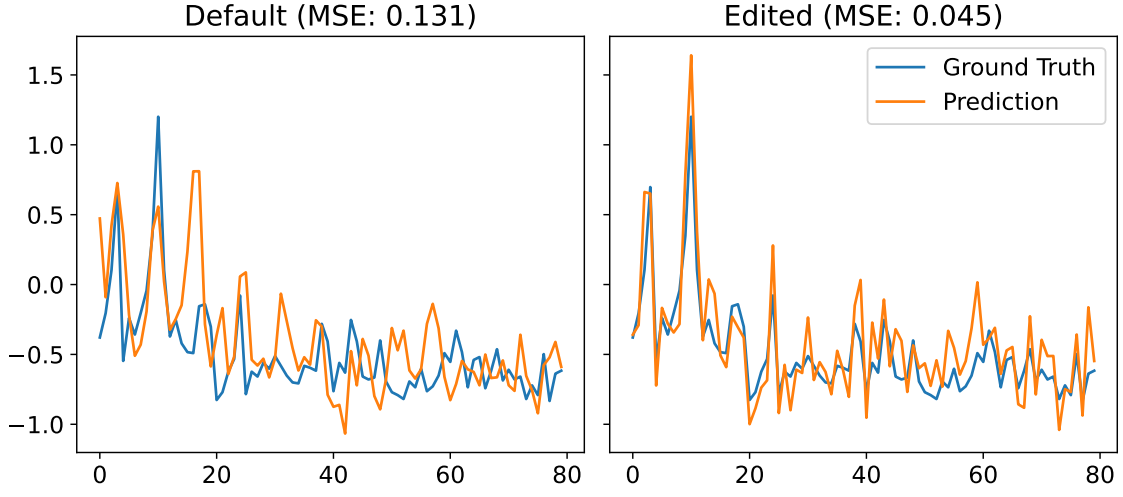


Figure 7.5: Result with Prediction length = 3, x-axis represents the days and y-axis represents the scaled purchase amounts

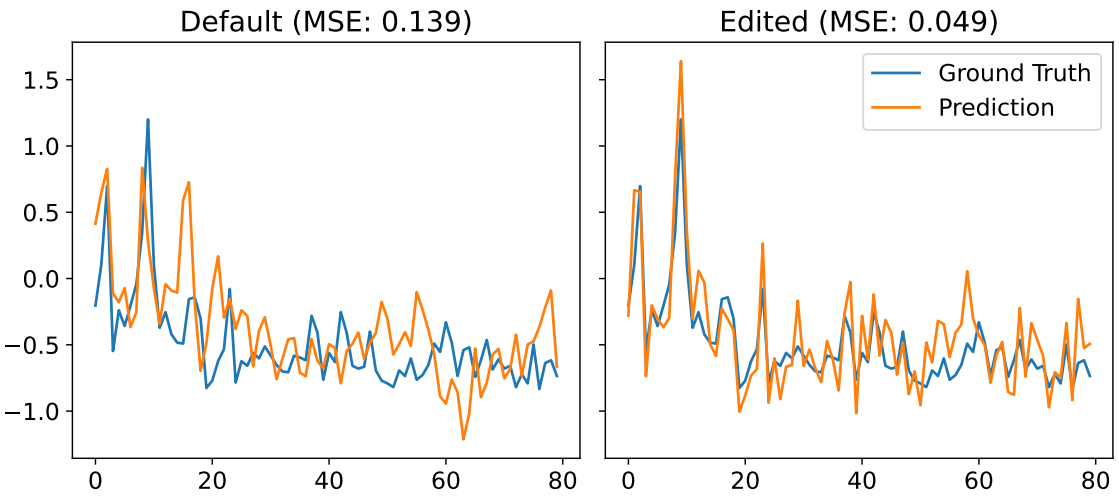


Figure 7.6: Result with Prediction length = 4, x-axis represents the days and y-axis represents the scaled purchase amounts

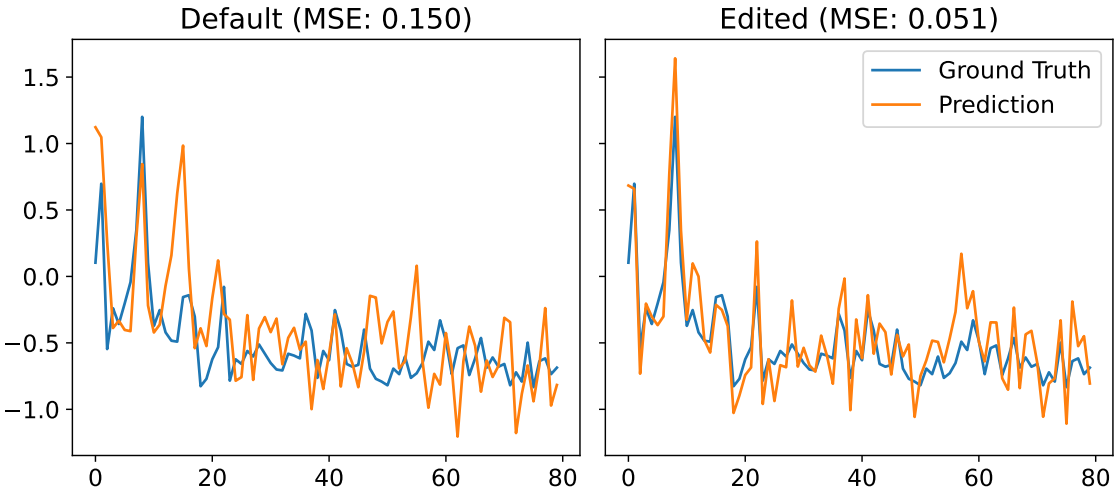


Figure 7.7: Result with Prediction length = 5, x-axis represents the days and y-axis represents the scaled purchase amounts

8

Clustering application and Impact analysis in retail

8.1/ Introduction

This chapter briefly presents some works that are more related to Colruyt. Simpler and more classical approaches have been used in these works. They have made it possible to meet practical needs encountered in companies. We will first present how clustering algorithms have been applied to better understand the evolution of stores along several axes such as attendance, turnover, or axes more related to its environment such as its competitive pressure, its surface. . . Then we see how an event, internal or external to the store, could impact the performance of this store. Here, an internal event means that the store itself makes a change within itself, and an external event means that the store does not really have control over the event, in this case, the store undergoes the change and tries rather to adapt. An example of an internal event would be a price decrease or increase of a product due to a decision of the management group. An example of an external event could be the opening of a new competing store near the Colruyt store.

8.2/ Clustering of Stores

A collaborative work was carried out with the company's expansion department in order to understand the existing stores on several axes such as the geographical position, the competitive environment, the perfect neighbors. A perfect neighbor here represents an external activity such as a high school or a cinema, which is close to a Colruyt store and gathers people. These people are usually potential customers of the store.

The results of these analyses will be used to better choose the location of a new store.

8.2.1/ DEFINITION

BOUNDARY ADDRESS

The exact address of a customer is a sensitive data, which can lead to the identification of a customer, reason why it cannot be used in an analysis. Each city is then subdivided into several areas, and in the database, the address of the customer (who has filled in his address when creating his loyalty card), becomes the address of this area. A customer will be on average about 2 mins drive from the center of his area. In Figure 8.1, we can see the division of the cities into areas. The real address of the customer remains unknown for the analyses. The center of the zone to which the customer belongs will be used to calculate the driving distance from the customer to the store.

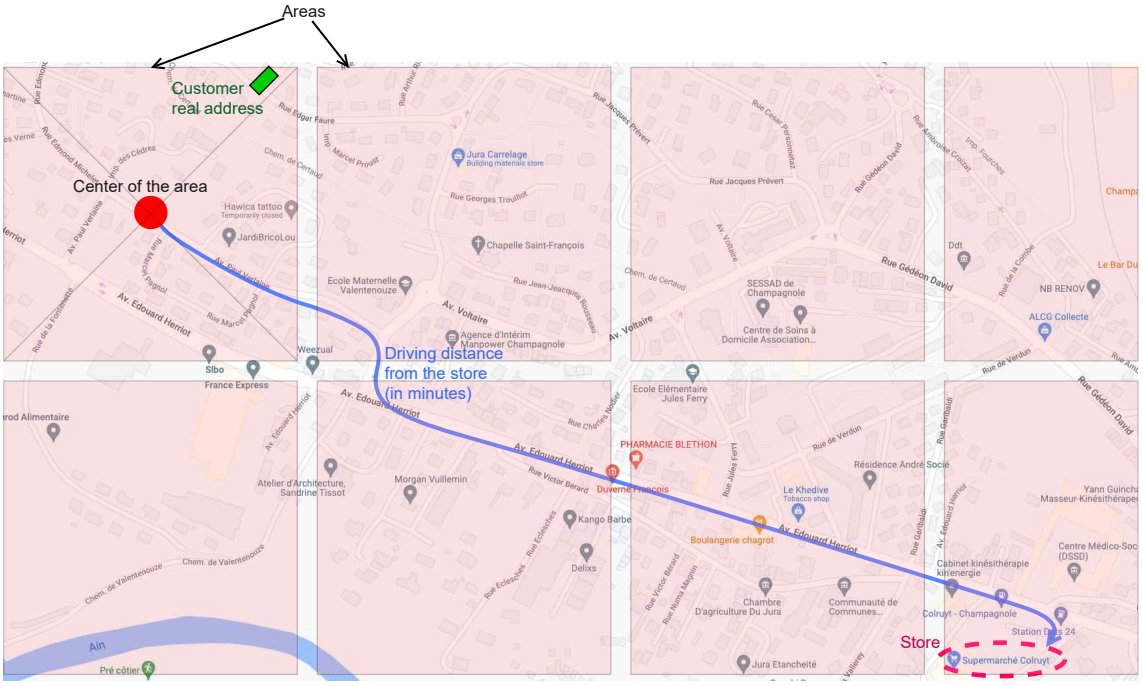


Figure 8.1: The area split of the cities. Each customer is represented by the center of its area. which represent each customer in the area. The actual address of the customer is unknown for the analyses. Here, for example, the center of the area is used to compute the driving distance from the customer to his store.

PERFECT NEIGHBOR

The perfect neighbor is an external activity, such as a high school, a movie theater, a gas station near a Colruyt store. These perfect neighbors attract a lot of people and they are usually potential customers of the Colruyt store. A high school for example will automatically have a lot of students, who might want to buy a sandwich in the store

between lunch and dinner, or they can go shopping in the Colruyt supermarket for student parties.

8.2.2/ CLUSTERING CRITERIA

The clustering analyses involve many criteria, like:

The characteristics of the store These are the store location, its postcode and size. The size information includes the style or the version of the store (G1 to G4, from small and simple to large and modern). There is also the opening date of the store which is important to know the maturity of the store.

The changes compared to last year Here the changes for an attribute like the turnover means the percentage change of the turnover compared to last year. It can be formulated as $((y2 - y1)/y1) * 100$ (where $y1$ is the start value and $y2$ is the end value). There are, the percentage change of the turnover per store, then the turnover only related to the "Pickup Service" [85] by store (orders placed on the internet and picked up in store).

Calculated values These are the distribution of driving distance, and the competitive pressure. For example, the distribution of driving distance is a value that depends on a long calculation process. First, it is necessary to know how long it takes to drive from a customer's "boundary address" to the store. This analysis only concerns customers with a loyalty card, which represents about 75% of the total number of customers. We will then have to divide the turnover in distinct intervals $([0, 5][5, 10]...[25, 30])$, which represent ranges of driving distance. The idea here is to be able to know how this distribution varies between stores and to be able to explain it.

There is also the competitive pressure, which consists in summing up the surfaces of the competitors around a store by type of store, if it is a discounter, an organic Store, a supermarket, a hypermarket or a proximity store. The calculations are made on public data. The surface areas of every store can be found on the map.

The characteristics of the store's city These include the urbanization of the city where the Colruyt store is located, the surface of that city, whether it is a city in the mountains or in the plains. All these information will allow to better guide the Machine Learning models.

8.2.3/ LESSONS

The result of this analysis is used to better choose the location of the next store. The work consisted in clustering the existing stores, in order to detect the factors which have an influence on the evolution of a store.

In Figure 8.2, we can see a representation of clustering of the stores on 2 main criteria, the changes of the turnover and the customers' attendance. The algorithm used here is K-means, a popular clustering algorithm.

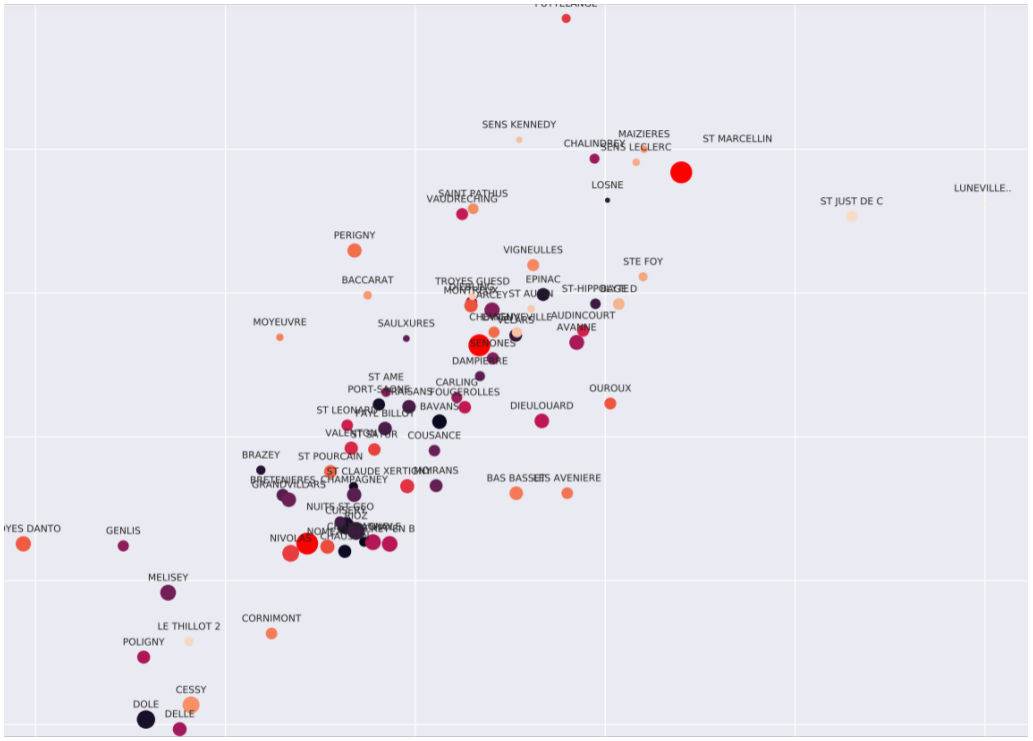


Figure 8.2: Stores clustering using K-means algorithm. On the graph, 2 criteria are considered for the clustering. The changes of the turnover in x-axis, and the changes of the customers' attendance in y-axis.

The result of the clustering is analyzed with the business, to draw the best lessons. Thanks to this result, we can draw conclusions like, the stores with a low competitive pressure that have a close customer base (the majority of the turnover is done through customers that live less than 5 minutes by car from the Colruyt store) have generally the best turnover evolutions.

Another conclusion is that some competing stores were at the same time, perfect neighbors. They are stores that attract customers who potentially end up visiting the Colruyt store next door, looking for certain products.

8.3/ IMPACT ANALYSIS ON SALES

This involves working with the business to estimate the impact of an event on sales. Like creating an additional service in a store like a bakery, or right next to a store, like a gas station. In some cases, it is a price increase on a product family. The idea here to solve the problem is to make a comparison over several years of sales, while taking into account the noise.

Use case: Setting up a bakery Colruyt has a bunch of stores in France. And all these stores, even if they are fundamentally similar, sometimes have small differences in appearance. For example, there are some stores with bakeries commonly called ”hot spots”, and other stores without these hot spots. The latter are often very expensive to install. The idea is to find out after calculation, if the investment is really worth it. The complexity of this problem lies in comparing things that are comparable, and eliminating as much as possible the biases that might exist.

8.4/ CONCLUSION

This chapter has provided a brief overview of some of the business-related work that has involved simpler and more familiar methods. These works have been used to meet practical business needs.

We have seen how clustering algorithms were used to study Colruyt stores and draw conclusions. These conclusions can be used to make big decisions within the company, such as the creation of a new store. Then we looked at the topic of estimating the impact of certain events, both internal and external, on the turnover of a store.

IV

Conclusion & Perspectives

9

Conclusion & Perspectives

9.1/ Conclusion

In this thesis, we compared prediction algorithms in order to propose the algorithms with the best results in predicting purchasing behavior. We also developed and proposed strategies to improve the accuracy of these models.

This dissertation is divided into two parts: the first discusses the scientific background of time series classification, forecasting, clustering techniques, as well as their application in retail, while the second one presents the contributions that have been made in this thesis.

The first part started with a general presentation of the retail industry. Before the advent of Machine Learning, the retail industry had already started to analyze purchasing behaviors and had even developed techniques such as Recency, Frequency, Money (RFM), to evaluate and use the results of these analyses, which were presented. We have seen that customer loyalty has always been a key issue in companies and retail was not the exception. Customer loyalty was discussed at length, we saw active customer loyalty resulting from a rational attachment or preference, passive customer loyalty resulting mainly from personal factors such as routine, or laziness to change. We then reviewed the regulations that govern the analysis and processing of this type of data.

A more in-depth presentation on the time series was made. In this sense, we saw that time series were a representation of data, which could be used to model data to then solve real problems encountered in business, related to these data.

As Deep Learning approaches have recently been proven in different domains, we presented their fundamental components. We discussed their effectiveness on time series classification problems. We introduced Gradient Boosting and Reservoir Computing which has shown convincing results on the tests performed. The state-of-the-art on time series prediction was briefly presented with the common point of classical approaches which generally try to decompose time series into seasonal variation, trend and irregular fluctuations. We were interested in how Deep Learning techniques have evolved to

attempt to solve this classification and forecasting problem, from MLP to attention-based approaches. Then we briefly looked at K-means which is a popular clustering algorithm. This algorithm is used to better apprehend some tasks encountered in business. The results of this algorithm were used to make decisions such as opening a new store.

The first part ends with the state-of-the-art on general forecasting in the retail industry. There was many studies that compared relatively deep or non-Deep Neural Networks to classical techniques such as ARIMA. Neural Network-based approaches have generally given the best results. Still, some studies have also successfully tested Prophet (a Facebook method, based on Fourier series) on retail demand prediction. Among all these studies, attention-based approaches which have already been successful in predicting many time series in other fields, were absent. This has been satisfied in the contributions of this thesis.

The second part was dedicated to the contributions. We first viewed the application of Machine Learning models on customer sales data for churn prediction. Churn is a marketing term for a customer who has moved to another company or who has gradually stopped buying from your company. The formal definition agreed upon in this study was to consider two periods P1 and P2, then a churner is a customer who was a regular buyer from the company during P1, but did not buy enough from the company during P2. In other words, his average basket during P2 is less than 20% of the average basket during P1. The study was performed on a dataset of 5,115,472 records of consumers with a loyalty card at Colruyt. Re-calibration techniques were used to improve the balance of the data set, allowing the models to achieve good accuracies. The re-calibration techniques were, data augmentation by scaling and Resampling. A comparison of models such as LSTM, XGBoost, MLP and Linear Regression as a baseline, was performed for the churn prediction. The LSTM had overall the best results followed by the XGBoost. The LSTM model achieved its maximum accuracy (Precision = 73.30%, F-measure = 72.21%) by accepting 30 days of input data. As a result of this study, the company was able to better anticipate the reaction of churned customers. From this study, the company was able to better anticipate churn, and propose offers to potential churners to keep them.

Next, we have seen the application of Machine Learning models to predict overstocking. Overstocking means having too much inventory in a store that has not been sold. It sometimes forces companies to sell excess products at a loss just before their expiration date instead of throwing them away. This is often the case for short-life products. This study focused on 150 products with a short use-by date, including the family of dairy products such as milk, cheese, yogurt, ice cream. We have tackled the overstock prediction problem, which is purely a demand forecasting problem. A comparative study was done on the LSTM, Autoformer, Informer, Transformer and XGBoost) to see which one would give the best results. The best results with the MSE metric were generally observed with the

Autoformer and the best results with the MAE metric were generally observed with the Transformer.

To finish the contributions, we discussed how to improve overstock forecasting by modifying the default behavior of the framework of the tested forecasting models. This study used the same data format as the overstock study, but considered two other product categories. These are weather-sensitive product families, which include products often purchased in summer for barbecues such as meat, beer with or without alcohol, rosé wines and aperitifs, and products commonly purchased in winter such as cheeses, foie gras and potatoes. By adding functionality to read known future weather data, the results improved significantly. It worked a little better than expected. The prediction tested range was $i + 1$ to $i + 14$ (i being the day we are positioned to run the prediction) with different model types. The modification had a real impact on the results. The performance of the modified models started to degrade from $i + 10$.

9.2/ PERSPECTIVES

There are many perspectives to consider in a constantly developing environment. The main objective of this thesis was to suggest solutions for purchasing behavior prediction, from receipt and loyalty card data.

In the following, some of the various approaches that should be considered for future research are described as possible perspectives.

NLP Each week, comments on local and regional events, with their predicted and actual impacts, are sent to functional groups within the company. These natural language comments contain a wealth of flat, unstructured information that often does not follow a common format. Another future work would be to use Natural Language Processing (NLP), to classify and reuse the result to improve other predictions such as demand forecasting and churn detection.

More product categories A future work could be to push further the demand forecasting, exploring more product categories. Look for categories that will be sensitive to other seasonalities that are not weather related. It can be seen in the sales data that outside Summer and Christmas, some products, beer for example, have other high demand periods. The idea here is to improve the accuracy of predictions by making predictions per product category.

Reservoir computing In this thesis, Reservoir computing has shown a considerable learning speed, but we were not able to explore in detail its hyperparameterization. One of the future works could be to do so in order to take advantage of its full potential.

Open data We could also consider testing the models with several external characteristics while avoiding as much as possible the overfitting of the models. Next, consider adding open data and testing Deep Learning models that can incorporate and learn from a variety of data types.

Global turnover Last but not least, there is currently prediction of the global turnover which is done using Data mining in Belgium. A future work could be to use the previous proposals listed in this section of perspectives such as NLP, Open data and Reservoir computing to predict global turnover through Deep Learning.

PUBLICATIONS

CONFERENCE PAPERS

- Kodjo Agbemadon, Raphael Couturier, David Laiymani. *"Churn Detection Using Machine Learning in the Retail Industry"*. In **2nd International Conference on Computer, Control and Robotics (ICCCR) 2022**, Accepted in 2022.

SUBMITTED PAPERS

- Kodjo Agbemadon, Raphael Couturier, David Laiymani. *"Overstock prediction using machine learning in retail industry"*. In **3rd International Conference on Computer, Control and Robotics (ICCCR) 2023**. Submitted in 2022.

BIBLIOGRAPHY

[1] ROSENBLATT, F. **The perceptron: a theory of statistical separability in cognitive systems (Project Para)**. Cornell Aeronautical Laboratory, 1958.

[2] PARZEN, E. **An approach to time series analysis**. *The Annals of Mathematical Statistics 32*, 4 (1961), 951–989.

[3] CHATFIELD, C. **The holt-winters forecasting procedure**. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 27*, 3 (1978), 264–279.

[4] AALEN, O. O. **A linear regression model for the analysis of life times**. *Statistics in medicine 8*, 8 (1989), 907–925.

[5] ELMAN, J. L. **Finding structure in time**. *Cognitive science 14*, 2 (1990), 179–211.

[6] WHITE, H., AND OTHERS. **Artificial neural networks**. Blackwell Cambridge, Mass., 1992.

[7] ANSUJ, A. P., CAMARGO, M., RADHARAMANAN, R., AND PETRY, D. **Sales forecasting using time series and neural networks**. *Computers & Industrial Engineering 31*, 1-2 (1996), 421–424.

[8] SNOW, R. E., CORNO, L., AND JACKSON III, D. **Individual differences in affective and conative functions.**

[9] ABITEBOUL, S. **Querying semi-structured data**. In *International Conference on Database Theory* (1997), Springer, pp. 1–18.

[10] DOWLING, G. R., AND UNCLES, M. **Do customer loyalty programs really work?** *Sloan management review 38* (1997), 71–82.

[11] HOCHREITER, S., AND SCHMIDHUBER, J. **Long short-term memory**. *Neural computation 9*, 8 (1997), 1735–1780.

[12] HOCHREITER, S. **The vanishing gradient problem during learning recurrent neural nets and problem solutions**. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 6*, 02 (1998), 107–116.

[13] MELTON, J. **Database language sql**. In *Handbook on Architectures of Information Systems*. Springer, 1998, pp. 105–132.

[14] CHATFIELD, C. **Time-series forecasting**. Chapman and Hall/CRC, 2000.

[15] TASHMAN, L. J. **Out-of-sample tests of forecasting accuracy: an analysis and review**. *International journal of forecasting 16*, 4 (2000), 437–450.

[16] ALON, I., QI, M., AND SADOWSKI, R. J. **Forecasting aggregate retail sales:: a comparison of artificial neural networks and traditional methods**. *Journal of retailing and consumer services 8*, 3 (2001), 147–156.

[17] KEAVENEY, S. M., AND PARTHASARATHY, M. **Customer switching behavior in online services: An exploratory study of the role of selected attitudinal, behavioral, and demographic factors**. *Journal of the academy of marketing science 29*, 4 (2001), 374–390.

[18] DIETTERICH, T. G., AND OTHERS. **Ensemble learning**. *The handbook of brain theory and neural networks 2*, 1 (2002), 110–125.

[19] ESTABROOKS, A., JO, T., AND JAPKOWICZ, N. **A multiple resampling method for learning from imbalanced data sets**. *Computational intelligence 20*, 1 (2004), 18–36.

[20] LYONS, R. G. **Understanding digital signal processing, (2004)**. *Motoyama T.(2005), Study on Variable Resolution Imaging in a Microscope OCT System, null 336* (2004).

[21] ANDERSEN, T. G., BOLLERSLEV, T., CHRISTOFFERSEN, P., AND DIEBOLD, F. X. **Volatility forecasting**, 2005.

[22] BUCKINX, W., AND VAN DEN POEL, D. **Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual fmcg retail setting**. *European journal of operational research 164*, 1 (2005), 252–268.

[23] YANG, Q., AND WU, X. **10 challenging problems in data mining research**. *International Journal of Information Technology & Decision Making 5*, 04 (2006), 597–604.

[24] ABURTO, L., AND WEBER, R. **Improved supply chain management based on hybrid demand forecasts**. *Applied Soft Computing 7*, 1 (2007), 136–144.

[25] ECKERSON, W. W. **Predictive analytics**. *Extending the Value of Your Data Warehousing Investment. TDWI Best Practices Report 1* (2007), 1–36.

[26] SCHRAUWEN, B., VERSTRAETEN, D., AND VAN CAMPENHOUT, J. **An overview of reservoir computing: theory, applications and implementations**. In *Proceedings of the 15th european symposium on artificial neural networks. p. 471-482 2007* (2007), pp. 471–482.

[27] CUNNINGHAM, P., CORD, M., AND DELANY, S. J. **Supervised learning**. In *Machine learning techniques for multimedia*. Springer, 2008, pp. 21–49.

[28] DEMOULIN, N. T., AND ZIDDA, P. **On the impact of loyalty cards on store loyalty: Does the customers' satisfaction with the reward scheme matter?** *Journal of retailing and Consumer Services 15*, 5 (2008), 386–398.

[29] MEIER, O., AND PACITTO, J.-C. **Le groupe colruyt: une réussite hors normes au sein de la grande distribution**. *Gestion 32*, 4 (2008), 28.

[30] ZORGATI, H. **Degré d'importance des actions de fidélisation: Les clients des grandes et moyennes surfaces tunisiennes**. *La Revue des Sciences de Gestion: Direction et Gestion 43*, 229 (2008), 103.

[31] COWPERTWAIT, P. S., AND METCALFE, A. V. **Introductory time series with R**. Springer Science & Business Media, 2009.

[32] PANKRATZ, A. **Forecasting with univariate Box-Jenkins models: Concepts and cases**. John Wiley & Sons, 2009.

[33] RODRIGUEZ, J. D., PEREZ, A., AND LOZANO, J. A. **Sensitivity analysis of k-fold cross validation in prediction error estimation**. *IEEE transactions on pattern analysis and machine intelligence 32*, 3 (2009), 569–575.

[34] MUTANEN, T., NOUSIAINEN, S., AND AHOLA, J. **Customer churn prediction–a case study in retail banking**. In *Data Mining for Business Applications*. IOS Press, 2010, pp. 77–83.

[35] WYFFELS, F., AND SCHRAUWEN, B. **A comparative study of reservoir computing strategies for monthly time series prediction**. *Neurocomputing 73*, 10-12 (2010), 1958–1964.

[36] BRAUN, M., AND SCHWEIDEL, D. A. **Modeling customer lifetimes with multiple causes of churn**. *Marketing Science 30*, 5 (2011), 881–902.

[37] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., AND OTHERS. **Scikit-learn: Machine learning in python**. *the Journal of machine Learning research 12* (2011), 2825–2830.

[38] POWERS, D. M. **Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation**. *Journal of Machine Learning Technologies* (2011).

[39] BERGMEIR, C., AND BENÍTEZ, J. M. **On the use of cross-validation for time series predictor evaluation**. *Information Sciences 191* (2012), 192–213.

[40] ESLING, P., AND AGON, C. **Time-series data mining**. *ACM Computing Surveys (CSUR) 45*, 1 (2012), 1–34.

[41] HINTON, G., DENG, L., YU, D., DAHL, G. E., MOHAMED, A.-R., JAITLY, N., SENIOR, A., VANHOUCKE, V., NGUYEN, P., SAINATH, T. N., AND OTHERS. **Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups**. *IEEE Signal processing magazine 29*, 6 (2012), 82–97.

[42] KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. **Imagenet classification with deep convolutional neural networks**. *Advances in neural information processing systems 25* (2012).

[43] LIN, J., WILLIAMSON, S., BORNE, K., AND DEBARR, D. **Pattern recognition in time series**. *Advances in Machine Learning and Data Mining for Astronomy 1*, 617-645 (2012), 3.

[44] MIGUÉIS, V. L., VAN DEN POEL, D., CAMANHO, A. S., AND E CUNHA, J. F. **Modeling partial customer churn: On the value of first product-category purchase sequences**. *Expert systems with applications 39*, 12 (2012), 11250–11256.

[45] MYERS, R. H., MONTGOMERY, D. C., VINING, G. G., AND ROBINSON, T. J. **Generalized linear models: with applications in engineering and the sciences**, vol. 791. John Wiley & Sons, 2012.

[46] RAI, A. K., AND SRIVASTAVA, M. **Customer loyalty attributes: A perspective**. *NMIMS management review 22*, 2 (2012), 49–76.

[47] STOFFER, D. S., AND OMBAO, H. **Special issue on time series analysis in the biological sciences**, 2012.

[48] DAHL, G. E., SAINATH, T. N., AND HINTON, G. E. **Improving deep neural networks for lvcsr using rectified linear units and dropout**. In *2013 IEEE international conference on acoustics, speech and signal processing* (2013), IEEE, pp. 8609–8613.

[49] DWIVEDI, A., NIRANJAN, M., AND SAHU, K. **A business intelligence technique for forecasting the automobile sales using adaptive intelligent systems (anfis and ann)**. *International Journal of Computer Applications 74*, 9 (2013).

[50] MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. S., AND DEAN, J. **Distributed representations of words and phrases and their compositionality**. *Advances in neural information processing systems 26* (2013).

[51] BAHDANAU, D., CHO, K., AND BENGIO, Y. **Neural machine translation by jointly learning to align and translate**. *arXiv preprint arXiv:1409.0473* (2014).

[52] CHAI, T., AND DRAXLER, R. R. **Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature**. *Geoscientific model development 7*, 3 (2014), 1247–1250.

[53] LÄNGKVIST, M., KARLSSON, L., AND LOUTFI, A. **A review of unsupervised feature learning and deep learning for time-series modeling**. *Pattern Recognition Letters 42* (2014), 11–24.

[54] LE, Q., AND MIKOLOV, T. **Distributed representations of sentences and documents**. In *International conference on machine learning* (2014), PMLR, pp. 1188–1196.

[55] SUTSKEVER, I., VINYALS, O., AND LE, Q. V. **Sequence to sequence learning with neural networks**. *Advances in neural information processing systems 27* (2014).

[56] ZHENG, Y., LIU, Q., CHEN, E., GE, Y., AND ZHAO, J. L. **Time series classification using multi-channels deep convolutional neural networks**. In *International conference on web-age information management* (2014), Springer, pp. 298–310.

[57] CAO, J., YU, X., AND ZHANG, Z. **Integrating owa and data mining for analyzing customers churn in e-commerce**. *Journal of Systems Science and Complexity 28*, 2 (2015), 381–392.

[58] CHEN, T., HE, T., BENESTY, M., KHOTILOVICH, V., TANG, Y., CHO, H., AND OTHERS. **Xgboost: extreme gradient boosting**. *R package version 0.4-2 1*, 4 (2015), 1–4.

[59] LECUN, Y., BENGIO, Y., AND HINTON, G. **Deep learning**. *nature 521*, 7553 (2015), 436–444.

[60] SONG, Y.-Y., AND YING, L. **Decision tree methods: applications for classification and prediction**. *Shanghai archives of psychiatry 27*, 2 (2015), 130.

[61] SZEGEDY, C., LIU, W., JIA, Y., SERMANET, P., REED, S., ANGUELOV, D., ERHAN, D., VANHOUCKE, V., AND RABINOVICH, A. **Going deeper with convolutions**. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 1–9.

[62] TAMBORINI, É. **Les programmes de fidélisation en grande distribution sont-ils efficaces face à des consommateurs de plus en plus exigeants?** PhD thesis, Simply Market, boulevard Edgard Kofler, 38500 Voiron, 2015.

[63] YAN, Y., CHEN, M., SHYU, M.-L., AND CHEN, S.-C. **Deep learning for imbalanced multimedia data classification**. In *2015 IEEE international symposium on multimedia (ISM)* (2015), IEEE, pp. 483–488.

[64] GOLDBERG, Y. **A primer on neural network models for natural language processing**. *Journal of Artificial Intelligence Research 57* (2016), 345–420.

[65] GOODFELLOW, I., BENGIO, Y., AND COURVILLE, A. **Deep learning**. MIT press, 2016.

[66] HENDRYCKS, D., AND GIMPEL, K. **Gaussian error linear units (gelus)**. *arXiv preprint arXiv:1606.08415* (2016).

[67] OORD, A. V. D., DIELEMAN, S., ZEN, H., SIMONYAN, K., VINYALS, O., GRAVES, A., KALCHBRENNER, N., SENIOR, A., AND KAVUKCUOGLU, K. **Wavenet: A generative model for raw audio**. *arXiv preprint arXiv:1609.03499* (2016).

[68] RAMCHOUN, H., IDRISSI, M. A. J., GHANOU, Y., AND ETTAOUIL, M. **Multilayer perceptron: Architecture optimization and training.** *IJIMAI 4*, 1 (2016), 26–30.

[69] ARAS, S., DEVECI KOCAKOÇ, İ., AND POLAT, C. **Comparative study on retail sales forecasting between single and combination methods**. *Journal of Business Economics and Management 18*, 5 (2017), 803–832.

[70] BAGNALL, A., LINES, J., BOSTROM, A., LARGE, J., AND KEOGH, E. **The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances**. *Data mining and knowledge discovery 31*, 3 (2017), 606–660.

[71] BÖSE, J.-H., FLUNKERT, V., GASTHAUS, J., JANUSCHOWSKI, T., LANGE, D., SALINAS, D., SCHELTER, S., SEEGER, M., AND WANG, Y. **Probabilistic demand forecasting at scale**. *Proceedings of the VLDB Endowment 10*, 12 (2017), 1694–1705.

[72] DE BAYNAST, A., LENDREVIE, J., AND LÉVY, J. **Mercator-12e éd.: Tout le marketing à l'ère digitale**, vol. 1. Dunod, 2017.

[73] DINGLI, A., MARMARA, V., AND FOURNIER, N. S. **Comparison of deep learning algorithms to predict customer churn within a local retail industry**. *International journal of machine learning and computing 7*, 5 (2017), 128–132.

[74] GAMBOA, J. C. B. **Deep learning for time-series analysis**. *arXiv preprint arXiv:1701.01887* (2017).

[75] KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. **Imagenet classification with deep convolutional neural networks**. *Communications of the ACM 60*, 6 (2017), 84–90.

[76] RAJURKAR, S., AND VERMA, N. K. **Developing deep fuzzy network with takagi sugeno fuzzy inference system**. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (2017), IEEE, pp. 1–6.

[77] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł., AND POLOSUKHIN, I. **Attention is all you need**. In *Advances in neural information processing systems* (2017), pp. 5998–6008.

[78] VOIGT, P., AND VON DEM BUSSCHE, A. **The eu general data protection regulation (gdpr)**. *A Practical Guide, 1st Ed., Cham: Springer International Publishing 10*, 3152676 (2017), 10–5555.

[79] WANG, Z., YAN, W., AND OATES, T. **Time series classification from scratch with deep neural networks: A strong baseline**. In *2017 International joint conference on neural networks (IJCNN)* (2017), IEEE, pp. 1578–1585.

[80] YAN, Y., CHEN, M., SADIQ, S., AND SHYU, M.-L. **Efficient imbalanced multimedia concept retrieval by deep learning on spark clusters**. *International Journal of Multimedia Data Engineering and Management (IJMDEM) 8*, 1 (2017), 1–20.

[81] AGARAP, A. F. **Deep learning using rectified linear units (relu)**. *arXiv preprint arXiv:1803.08375* (2018).

[82] ALEKSANDROVA, Y. **Application of machine learning for churn prediction based on transactional data (rfm analysis)**. In *18 International Multidisciplinary Scientific Geoconference SGEM 2018: Conference Proceedings* (2018), vol. 18, pp. 125–132.

[83] BAI, S., KOLTER, J. Z., AND KOLTUN, V. **An empirical evaluation of generic convolutional and recurrent networks for sequence modeling**. *arXiv preprint arXiv:1803.01271* (2018).

[84] CARON, M., BOJANOWSKI, P., JOULIN, A., AND DOUZE, M. **Deep clustering for unsupervised learning of visual features**. In *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 132–149.

[85] JIN, M., LI, G., AND CHENG, T. **Buy online and pick up in-store: Design of the service area**. *European Journal of Operational Research 268*, 2 (2018), 613–623.

[86] PATHAK, J., HUNT, B., GIRVAN, M., LU, Z., AND OTT, E. **Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach**. *Physical review letters 120*, 2 (2018), 024102.

[87] POUYANFAR, S., SADIQ, S., YAN, Y., TIAN, H., TAO, Y., REYES, M. P., SHYU, M.-L., CHEN, S.-C., AND IYENGAR, S. S. **A survey on deep learning: Algorithms,**

techniques, and applications. *ACM Computing Surveys (CSUR) 51*, 5 (2018), 1–36.

[88] RANGAPURAM, S. S., SEEGER, M. W., GASTHAUS, J., STELLA, L., WANG, Y., AND JANUSCHOWSKI, T. **Deep state space models for time series forecasting**. *Advances in neural information processing systems 31* (2018).

[89] SCHRATZ, P., MUENCHOW, J., ITURRITXA, E., RICHTER, J., AND BRENNING, A. **Performance evaluation and hyperparameter tuning of statistical and machine-learning models using spatial data**. *arXiv preprint arXiv:1803.11266* (2018).

[90] SILVA, D. F., GIUSTI, R., KEOGH, E., AND BATISTA, G. E. **Speeding up similarity search under dynamic time warping by pruning unpromising alignments**. *Data Mining and Knowledge Discovery 32*, 4 (2018), 988–1016.

[91] TATO, A., AND NKAMBOU, R. **Improving adam optimizer**.

[92] YOUNG, T., HAZARIKA, D., PORIA, S., AND CAMBRIA, E. **Recent trends in deep learning based natural language processing**. *ieee Computational intelligenCe magazine 13*, 3 (2018), 55–75.

[93] AHMAD, A. K., JAFAR, A., AND ALJOUMAA, K. **Customer churn prediction in telecom using machine learning in big data platform**. *Journal of Big Data 6*, 1 (2019), 1–24.

[94] LIANG, H., ZHANG, S., SUN, J., HE, X., HUANG, W., ZHUANG, K., AND LI, Z. **Darts+: Improved differentiable architecture search with early stopping**. *arXiv preprint arXiv:1909.06035* (2019).

[95] MUDELSEE, M. **Trend analysis of climate time series: A review of methods**. *Earth-science reviews 190* (2019), 310–322.

[96] SEN, R., YU, H.-F., AND DHILLON, I. S. **Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting**. *Advances in neural information processing systems 32* (2019).

[97] TOPOL, E. J. **High-performance medicine: the convergence of human and artificial intelligence**. *Nature medicine 25*, 1 (2019), 44–56.

[98] WANG, Y., SMOLA, A., MADDIX, D., GASTHAUS, J., FOSTER, D., AND JANUSCHOWSKI, T. **Deep factors for forecasting**. In *International conference on machine learning* (2019), PMLR, pp. 6607–6617.

[99] AMINANTO, M. E., BAN, T., ISAWA, R., TAKAHASHI, T., AND INOUE, D. **Threat alert prioritization using isolation forest and stacked auto encoder with day-forward-chaining analysis**. *IEEE Access 8* (2020), 217977–217986.

[100] BIANCHI, F. M., SCARDAPANE, S., LØKSE, S., AND JENSSEN, R. **Reservoir computing approaches for representation and classification of multivariate time series**. *IEEE transactions on neural networks and learning systems 32*, 5 (2020), 2169–2179.

[101] HOBAN, R. A. **Introducing the slope concept**. *International Journal of Mathematical Education in Science and Technology* (2020), 1–17.

[102] LIVIERIS, I. E., PINTELAS, E., AND PINTELAS, P. **A cnn–lstm model for gold price time-series forecasting**. *Neural computing and applications 32*, 23 (2020), 17351–17360.

[103] SALINAS, D., FLUNKERT, V., GASTHAUS, J., AND JANUSCHOWSKI, T. **Deepar: Probabilistic forecasting with autoregressive recurrent networks**. *International Journal of Forecasting 36*, 3 (2020), 1181–1191.

[104] ZUNIC, E., KORJENIC, K., HODZIC, K., AND DONKO, D. **Application of facebook's prophet algorithm for successful sales forecasting based on real-world data**. *arXiv preprint arXiv:2005.07575* (2020).

[105] JANIESCH, C., ZSCHECH, P., AND HEINRICH, K. **Machine learning and deep learning**. *Electronic Markets 31*, 3 (2021), 685–695.

[106] LIM, B., AND ZOHREN, S. **Time-series forecasting with deep learning: a survey**. *Philosophical Transactions of the Royal Society A 379*, 2194 (2021), 20200209.

[107] RUIZ, A. P., FLYNN, M., LARGE, J., MIDDLEHURST, M., AND BAGNALL, A. **The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances**. *Data Mining and Knowledge Discovery 35*, 2 (2021), 401–449.

[108] WU, H., XU, J., WANG, J., AND LONG, M. **Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting**. *Advances in Neural Information Processing Systems 34* (2021).

[109] TONG, Q., LIANG, G., AND BI, J. **Calibrating the adaptive learning rate to improve convergence of adam**. *Neurocomputing* (2022).

[110] ABADI, M., AGARWAL, A., BARHAM, P., BREVDO, E., CHEN, Z., AND CITRO, C. **Tensorflow: Large-scale machine learning on heterogeneous systems**. Software available from tensorflow.org. Available from:.

[111] PASZKE, A., GROSS, S., MASSA, F., LERER, A., BRADBURY, J., AND CHANAN, G. **Pytorch: An imperative style, high-performance deep learning library**. In *Advances in Neural Information Processing Systems 32*. p. 8024–8035.

List of Figures

List of Tables

Document generated with LaTeX and:
the LaTeX style for PhD Thesis created by S. Galland — http://www.multiagent.fr/ThesisStyle
the tex-upmethodology package suite — http://www.arakhne.org/tex-upmethodology/

SPIM
école doctorale **sciences pour l'ingénieur et microtechniques**

Title: Prediction of purchasing behavior using Deep Learning techniques

Keywords: Retail industry, Purchasing behavior prediction, Churn prediction, Overstock prediction, Time series classification, Deep Learning

Abstract:

Recently, large retail companies often have the same suppliers, although some focus more on their own private label. The competitiveness of the retail sector relies, in part, on anticipating consumer reaction and optimising the supply chain. The receipts generated during purchases produce a large volume of data. The processing of these data can allow numerous analyses and predictions, particularly on purchasing behavior and supply chain improvement. The objective of this thesis is to predict purchasing behavior. To achieve this, we propose machine learning methods to detect the factors that lead to a change in purchasing behavior. More precisely, we have worked on the following challenges:
1) The application of machine learning models on sales dataset to identify customers at risk of no longer buying in stores. This is known as "churn". This study uses linear regression as a baseline to compare machine learning models such as gradient boosting, MLP (Multilayer perceptron) and LSTM (Long Short Term Memory). LSTM outperformed the other approaches due to the fact that they were designed to be able to learn order dependence.
2) The application of machine learning models to predict overstocking and wastage of short-life products, including fresh and dairy products. A total of 5 machine learning models were compared in this study, including gradient boosting, LSTM, Transformer, Informer, and AutoFormer. The latter was able to give the best results thanks to the time series decomposition into trend and seasonality components.
3) We have also worked on feeding different machine learning models (Gradient Boosting, LSTM and Autoformer) with priori-known future weather data, in order to improve demand forecasting. In each case tested, the modified algorithms always perform better than their original versions. The work presented in this thesis was the result of a collaboration between the retail company Colruyt France and the department of computer science of complex systems (DISC) of the FEMTO-ST laboratory under a CIFRE contract.

UBFC UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ
Université Bourgogne Franche-Comté
32, avenue de l'Observatoire
25000 Besançon, France