

Ministry of Higher Education and Scientific Research
National Higher School of Statistics and Applied Economics

ENSSEA

MASTER'S THESIS

**Data Mining and Machine Learning for Customer Churn
Prediction
Case study: Telecom Company**

Author:
Yaicne AMMI

Supervisor:
Dr. Khaled ROUASKI

A Thesis Submitted in Partial Fulfillment for the Degree of Master in

APPLIED STATISTICS
Department of Applied Statistics and Econometrics



8th Promotion
2021

Ministry of Higher Education and Scientific Research
National Higher School of Statistics and Applied Economics

ENSSEA

MASTER'S THESIS

**Data Mining and Machine Learning for Customer Churn
Prediction
Case study: Telecom Company**

Author:
Yaicne AMMI

Supervisor:
Dr. Khaled ROUASKI

A Thesis Submitted in Partial Fulfillment for the Degree of Master in

APPLIED STATISTICS
Department of Applied Statistics and Econometrics



8th Promotion
2021

DEDICATIONS

I dedicate my dissertation work to my family and many friends. A special feeling of gratitude to my loving MOTHER and FATHER, whose words of encouragement and push for tenacity ring in my ears, I will always appreciate all they have done.

I dedicate this work to my paternal Grandfather who raised me and loved me. My maternal Grandmother and my Uncle HAKIM who have never left my side and are very special.

Furthermore, I dedicate this work to my young brother and sister; particularly my dearest brother, YANIS, who helped me when it was most needed.

I also dedicate this dissertation to my many friends who have supported me throughout the process. and give special thanks to my best friends OMAR, ISLAM and AIMEN for being there and for their help throughout the entire dissertation.

ACKNOWLEDGEMENTS

I would want to begin by thanking God, the all-powerful and merciful, for giving me the strength and patience to complete this simple task.

I want to express my heartfelt gratitude to my memory director, Dr. ROUASKI, for his invaluable advice, consistent guidance, and moral and educational support throughout this work.

I would want to thank all of my professors for their contributions throughout my academic career, particularly Mr. IHADAADEN, Mr. BENSALMA and Mme. LOUNICI. Our heartfelt gratitude goes to all of our ENSSEA teammates, especially Aimen.

Finally, thanks goes to everyone who helped bring this project to completion, no matter how close or far they were.

ABSTRACT

Predicting churn risk is critical for telecommunications providers because it represents a significant threat to their earnings. In this end-of-study project, we create an accurate predictive model that can assess and recognize subscribers that are more likely to leave the telecom operator, keeping the following goals in mind: Building the model, assessing the factors that influence attrition in order to initiate retention measures, and targeting the subscribers who are most likely to churn. To answer the initial questions, we employ Machine Learning algorithms (ensemble learning).

An empirical study is carried out by applying results from the literature to data provided by an anonymous telecommunications company. After preprocessing the data and running the models, it was discovered that XGBoost and LightGBM were the best-performing models, with XGBoost being the best. The evaluation was carried out using various metrics such as the confusion matrix, the ROC-curve, the F-score, the accuracy, and the AUC-scores.

By the end of our work, we have created a basic interpretation system to interpret the results that the model produced.

Keywords: Churn, Customer churn prediction, Telecommunication, Machine Learning, Data Mining, Supervised Learning, XGBoost.

RÉSUMÉ

Prédire le risque de désabonnement est essentiel pour les fournisseurs de télécommunications car il représente une menace importante pour leurs revenus. Dans ce projet de fin d'étude, nous créons un modèle prédictif précis qui peut évaluer et reconnaître les abonnés les plus susceptibles de quitter l'opérateur de télécommunications, en gardant à l'esprit les objectifs suivants : Construire le modèle, évaluer les facteurs qui influencent l'attrition afin d'initier des mesures de rétention, et cibler les abonnés qui sont les plus susceptibles de désert. Pour répondre aux questions initiales, nous utilisons des algorithmes d'Apprentissage Automatique (apprentissage d'ensemble).

Une étude empirique est réalisée en appliquant les résultats de la littérature aux données fournies par une entreprise de télécommunications anonyme. Après avoir prétraité les données et exécuté les modèles, on a découvert que XGBoost et LightGBM étaient les modèles les plus performants, XGBoost étant le meilleur. L'évaluation a été effectuée à l'aide de diverses mesures telles que la matrice de confusion, la courbe ROC, le F-score, la précision et les scores AUC.

À la fin de notre travail, nous avons créé un système d'interprétation de base pour interpréter les résultats produits par le modèle.

Mots clés: Churn, Prédiction du churn, Télécommunications, Apprentissage Automatique, Data Mining, Apprentissage Supervisé, XGBoost.

TABLE OF CONTENTS

Abstract.....	III
Table of Contents	V
List of Figures.....	VIII
List of Tables	X
List of Abbreviations	XI
General Introduction.....	2
Chapter I Concepts and Definitions	7
Introduction	7
1. Customer Relationship Management (CRM)	8
1.1. Defining CRM.....	8
1.2. History of CRM.....	8
1.3. CRM phases	8
2. Churn Phenomenon.....	11
2.1. Definitions	11
2.2. Churners types	12
2.3. Importance of churn	13
2.4. Factors of churn.....	14
2.5. Methods of reducing churn.....	16
3. RFM Analysis	18
3.1. RFM Basics	18
3.2. RFM Scoring	19
3.3. Segmentation	19
4. Data Mining	20
4.1. Definition.....	20
4.2. History of Data Mining	20
4.3. Process of Data Mining	21
4.4. Data mining in the CRM framework.....	21
5. Machine Learning	25

5.1. Definition.....	25
5.2. History of Machine Learning	25
5.3. Use cases of Machine Learning.....	26
Conclusion.....	27
Chapter II Data Mining & Machine Learning Processes	29
Introduction	29
1. Data Preprocessing Procedures.....	30
1.1. Data cleaning, Normalization and Transformation	31
1.2. Missing data	32
1.3. Sampling.....	33
1.4. Feature selection.....	33
1.5. Train – Test split.....	34
2. Modeling Process.....	35
2.1. Unsupervised Learning:	35
2.2. Supervised Learning:.....	38
3. Model evaluation	47
3.1. Cross-validation.....	47
3.2. Hyperparameter optimization.....	48
3.3. Evaluation Metrics	48
3.4. Model Interpretation.....	52
Conclusion.....	54
Chapter III Analysis & Results.....	56
Introduction	56
1. Exploratory Data Analysis.....	57
1.1. Dataset overview	57
1.2. Checking for missing values	58
1.3. Correlation inspection	58
1.4. Univariate analysis:	60
1.5. Multivariate analysis:	63
2. RFM Segmentation.....	65

2.1.	Creating the RFM table	65
2.2.	Standardizing the RFM table.....	65
2.3.	Unsupervised learning with K-means	66
3.	Preprocessing Data	68
3.1.	Standardization	68
3.2.	Train-Test Splitting	68
3.3.	Class imbalance	68
3.4.	Feature selection.....	70
4.	Modelling.....	71
4.1.	Model selection	71
4.2.	Hyperparameter optimization.....	73
5.	Model Evaluation and Interpretation	74
5.1.	Classification report	74
5.2.	ROC & Precision/Recall curves	74
5.3.	Model interpretation with SHAP values	75
	Conclusion	79
	General Conclusion	81
	Bibliography	86
	Appendices.....	92
	Appendix 01: Code for importing used libraries and packages	92
	Appendix 02: Code used for model selection.....	93
	Appendix 03: Features description.....	94

LIST OF FIGURES

Figure 1: Research design	5
Figure 2: Phases of Customer Relationship Management.....	9
Figure 3: Churn taxonomy.....	13
Figure 4: Most significant retail revenue drivers	14
Figure 5: Factors of churn	15
Figure 6: Tactics for reducing churn	17
Figure 7: Pyramid model	19
Figure 8: History of data mining development.....	20
Figure 9: Data mining and customer lifecycle management.	22
Figure 10: The increase in predictive ability using data mining	24
Figure 11: Data preprocessing tasks.....	30
Figure 12. K-means clustering on 150 simulated point.....	36
Figure 13: The Elbow Method	38
Figure 14: A Decision Tree	42
Figure 15: Accuracy of some models in customer churn prediction.....	43
Figure 16: Cross Validation.....	47
Figure 17: Email spam detection	48
Figure 18: Confusion Matrix	49
Figure 19: ROC and AUC	51
Figure 20: SHAP (SHapley Additive exPlanation)	53
Figure 21: Feature names	57
Figure 22: Missing values check	58
Figure 23: Correlation Matrix of all features	59
Figure 24: Distribution of non-activity of users in days.....	61
Figure 25: Distribution of user_spendings	61
Figure 26: Distribution of user_lifetime.....	62
Figure 27: Distribution of user_intake	62

Figure 28: Customers spending habits	63
Figure 29: Spending frequency	63
Figure 30: Churn by different services	64
Figure 31: Elbow method	66
Figure 32: Recency, Frequency & Monetary Box-plots.....	66
Figure 33: Comparing churn percentage by cluster	67
Figure 34: Class imbalance in the training set.....	69
Figure 35: Phases of resampling.....	69
Figure 36: Feature selection's best features	70
Figure 37: Confusion matrices of top-performing models	72
Figure 38: Hyperparameter results	73
Figure 39: Classification report	74
Figure 40: ROC & Precision/Recall curves.....	75
Figure 41: SHAP Feature importance	76
Figure 42: SHAP Summary plot.....	77
Figure 43: SHAP individual 1	77
Figure 44: SHAP individual 2	78

LIST OF TABLES

Table 1: Comparing learning algorithms (**** stars for best, * star for worst)	40
Table 2: Training set.....	41
Table 3: Correlation of filtered features	60
Table 4: Sample from RFM table	65
Table 5: Scaled RFM table sample.....	65
Table 6: Standardized features sample	68
Table 7: Split data sizes	68
Table 8: ML algorithms comparison	71

LIST OF ABBREVIATIONS

AUC	Area Under Curve
CCP	Customer Churn Prediction
CRM	Customer Relationship Management
FN	False Negatives
FP	False Positives
FPR	False Positives Rate
kNN	k Nearest Neighbors
Max	Maximum
Min	Minimum
ML	Machine Learning
QoS	Quality of Service
Recall_STD	Recall Standard Deviation
RFM	Recency, Frequency and Monetary
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive explanation
Std_dev	Standard Deviation
SVM	Support Vector Machine
TN	True Negatives
TP	True Positives
TPR	True Positives Rate

GENERAL INTRODUCTION

GENERAL INTRODUCTION

Motivation and Background:

Nowadays, customers are increasingly interested in the quality of service (QoS) that firms may offer them. The services provided by many suppliers are not very distinctive, increasing rivalry among enterprises to maintain and improve their QoS.

Customer Relationship Management (CRM) solutions help businesses recruit new customers, maintain ongoing relationships with them, and boost customer retention for increased profitability.

Organizations nowadays face a variety of issues as a result of tough competition and changing markets. Churn is a trend in reducing customers and is a significant concern for organizations in several economic sectors, especially for the telecommunications industry.

Since the 1990s, the telecommunications sector has emerged as a dynamic critical area for the economic development of developed countries. This is due to tremendous technological improvement, as well as an increase in the number of network operators and the intense competition that has evolved; as a result of the cell phone market's saturation, the trend is no longer toward acquisition (acquiring new subscribers), but rather toward retention (keeping the consumers who already subscribed).¹

Customer churn can be an expensive risk if not effectively controlled, there are considerable costs involved with this occurrence, including lost revenue, retention costs, and so on, that is why considering churn as a priority issue is critical.

¹ Rob Mattison: « **The Telco Churn Management Handbook** », XiT Press, Illinois, USA, 2005, pp. 20.

This has led to the emergence of Customer churn prediction (CCP), a type of customer relationship management (CRM) in which a corporation attempts to develop a model that forecasts whether or not a customer intends to leave or reduce their purchases from a company. CCP is widely explored in various areas, especially the telecommunications industry, where companies use machine learning (ML) based methods for customer churn prediction.

The objective of the study:

We have discovered that churn is a real issue for telephone service providers. As a result, this study's primary goal is to put in place a precise predictive model capable of detecting clients at high risk of attrition before they leave their cycle of activity and then identify early warning signs through their behaviours.

Furthermore, we will continue to build a churn score that will allow us to give a churn likelihood to each client. This data will enable the telecommunication company to be proactive in the face of churn, which we hope will assist business experts in launching retention efforts by targeting profiles at a high risk of attrition.

Research questions and Hypotheses:

A solid understanding of the machine learning sector and its potential applications to the telecommunication industry is required to construct a suitable model. This study compares new methods to it. The core and sub-research questions are formulated below based on the objectives and the specific data type that we have:

- How can data mining be used to forecast client behaviour?

As a result of our investigation, we will be able to answer the following sub-questions:

- Can data mining techniques accurately forecast customer behaviour?
- What algorithms are most effective in customer churn prediction?
- What is the most crucial variable in predicting the customer churn given the dataset?

From these research questions, we formulate the following hypotheses:

H₁: We can rely on data mining to anticipate client behaviour.

H₂: Ensemble Learning methods are the most suited algorithms to predict customer churn.

H₃: The most crucial variable for predicting customer churn is the customer last account balance.

Structure of the thesis:

The structure of this thesis is as follows:

The first chapter will make a good compilation of the literature more widely and then narrow it down to get a good overview of three key concepts: CRM, Customer churn and CPP, RFM analysis, Data Mining and Machine Learning.

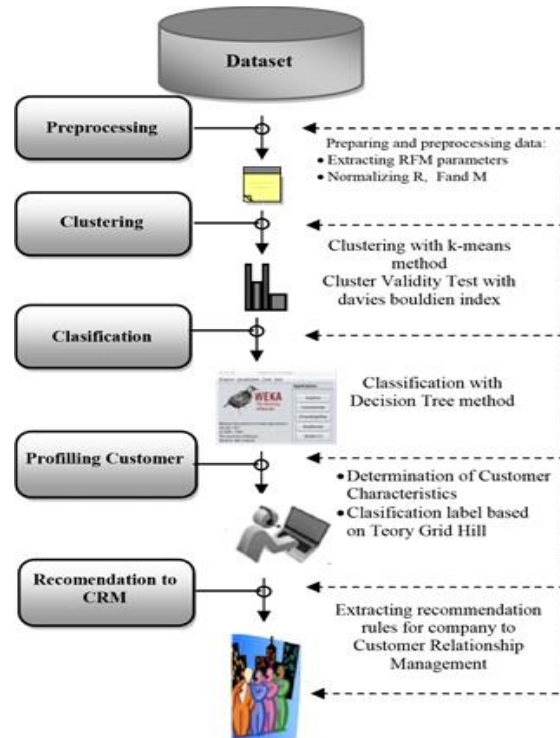
The second chapter will present a detailed review of the machine learning methodology and the best algorithms for CCP based on the literature review. More specifically, this chapter will discuss: Machine Learning and Data Mining, Supervised and Unsupervised learning, model Evaluation and Interpretation.

In the last chapter, we will practice the knowledge gained from the first two chapters, a precise model will be constructed, evaluated, and interpreted.

Research Methodology and Tools:

We chose to employ a well-developed method known as data mining techniques. The methodology for the practical study is depicted in Figure 1.

Figure 1: Research design



Source: Maryani and D. Riana, "Clustering and profiling of customers using RFM for customer relationship management recommendations," 5th International Conference on Cyber and IT Service Management, 2017, pp. 6

In order to accomplish such quest, we used **Python 3.7** with **Kaggle IDE** as an IDE, the code to import the used packages is in Appendix 01, we used the following packages:

- **Pandas** for cleaning and analyzing the data.
- **Matplotlib** and **Seaborn** for visualizing the data.
- **Scikitlearn** and **XGBoost** for building machine learning models.
- **Optuna** for hyperparameter optimization and **SHAP** for model interpretation.

CHAPTER I: CONCEPTS AND DEFINITIONS

CHAPTER I

CONCEPTS AND DEFINITIONS

Introduction

In increasingly competitive markets, companies have realized that they need to do more to keep customers. This has led to an increase in the level of competitiveness, hence the appearance of switchers.

This chapter will first start with a definition of the **CRM**, its history and its phases

We will then give a literature review to highlight the phenomenon of **Churn** and explain the environment that is favourable to its occurrence and its impact on the telecom sector.

After that, we will explain the concept of **RFM**, its utility and how to implement it in real-world problems.

Finally, we will define **Data Mining** and **Machine Learning**, look at their history and look at their utility in the churn problem.

1. Customer Relationship Management (CRM)

1.1. Defining CRM

CRM is an essential business strategy that seeks to establish and maintain profitable relationships with customers by developing and delivering superior value offers. It is based on high-quality customer data and is made possible by information technology.²

1.2. History of CRM

Historically, most businesses were located near the markets they served and knew their clients well. Face-to-face, even day-to-day, connections with clients would frequently occur, allowing knowledge of client requirements and preferences to grow.

However, as businesses have expanded in size, they have gotten increasingly distant from the customers they serve. The distance is not merely geographical, and the isolation is geographical; it may also be cultural.

Geographic and cultural isolation, along with company owner and management separation from customer interaction, means that many enterprises, especially smaller ones, lack the intuitive awareness and understanding of their consumers that can be found in micro-businesses like neighbourhood stores and hairdressing salons. This has increased the demand for improved customer-related data, the cornerstone of effective CRM.³

1.3. CRM phases

Customer management is made up of four components that can be represented as a phased model. All efforts are intimately tied with a thorough knowledge of the consumer, or customer insight⁴. These four phases are depicted in Figure 2.

² Francis Buttle, Stan Maklan : **Customer Relationship Management: Concepts and Technologies**, 4th. ed, Routledge, 2019, pp. 17

³ Ibid., pp. 18

⁴ Alexander H. Kracklauer, D. Quinn Mills, Dirk Seifert: **Collaborative Customer Relationship Management: Taking CRM to the Next Level**, 1st. ed, Springer-Verlag Berlin Heidelberg, 2004, pp. 04

Figure 2: Phases of Customer Relationship Management



Source: Alexander H. Kracklauer, D. Quinn Mills, Dirk Seifert: Collaborative Customer Relationship Management: Taking CRM to the Next Level, 1st. ed, Springer-Verlag Berlin Heidelberg, 2004, pp. 04

1.3.1. Identification:

Systematic customer management starts with identifying target groups and the gathering of quantitative and qualitative data on these groups. The corporation selects a consumer category that is most appealing to it.

Customer identification also examines the customers who have defected to the competition and how they might be re-acquired. Customer identification tools include customer segmentation, consumer market research, and consumer target group analysis.

1.3.2. Attraction:

A company's marketing efforts must always be weighed against those of its competitors. Benchmarking, promotions, and free samples are examples of customer attraction tools.

The methodical development of competitive advantages (price leadership or differentiation advantages) promotes better consumer attraction circumstances.

1.3.3. Retention:

One of the primary problems of customer management is the long-term retention of profitable clients.

Customer satisfaction is the consequence of a comparison process between the customer's expectations and perceptions. A customer's long-term impression of the manufacturer's added value leads to long-term client retention. Customer retention tools include one-to-one marketing, loyalty and bonus programs, personalization, and complaint handling.

Furthermore, a good shopping experience creates a favourable emotional association and establishes the framework for client loyalty.

1.3.4. Development:

Customer development aims to consistently increase transaction intensity, transaction value, and individual customer profitability. Increasing the customer's wallet share is performed by directing them to different product or service options.

Up- and cross-selling and product and service bundling are examples of customer development tools.

2. Churn Phenomenon

Nowadays, customers have multiple service providers to choose from. They can quickly switch services or even the provider. These customers are known as churned customers.

2.1. Definitions

2.1.1. Churn:

Churn is a word derived from the English phrase "change and turn." It is often used in marketing to describe a customer decline trend.

It is a marketing term that refers to a customer who switches from one company to another. He already has a client relationship with the focal firm, but he will go to the rival shortly. A preservation action is expected if the company wishes to keep him from leaving. Most churn definitions rely on a customer's product behaviour and a threshold set by a market law. If a customer's activity falls below the threshold (or equals zero), the customer is called a churner.⁵

The turnover rate or churn rate typically measures the phenomenon.

2.1.2. Churn rate:

When applied to a consumer base, the churn rate refers to the percentage of contracted clients or subscribers who leave a provider over a specified period. It may indicate consumer discontent, cheaper and better prices from the competition, more competitive sales or marketing, or factors related to the customer life cycle. The attrition rate is the opposite of the churn rate.

Churn rate is calculated with the following formula:

$$\text{churn rate} = \frac{\text{Number of Churned Customers}}{\text{Total Number of Customers}}$$

⁵ Nicolas Glady, Bart Baesens, Christophe Croux : « Modeling churn using customer lifetime value », **European Journal of Operational Research**, Vol. 197, Issue. 1, 2009 pp. 404-405

2.1.3. Attrition rate:

Attrition is the number of clients and/or customers maintained by a business for a given period. It is often used to represent brand loyalty among brand customers.

Attrition rate is a loyalty metric expresses the proportion of customers who stay with a company from one time to another, the fiscal year is the most commonly used era. It aims to assess the profitability of customer recruitment activities.

2.2. Churners types

There are numerous categories of churners, the two main ones are voluntary and Involuntary churners.

2.2.1. Involuntary churners:

Involuntary churners are the easiest to spot. These are the customers that the company decides to exclude from the list of subscribers. As a result, this group involves churned individuals for fraud, non-payment, and inactive customers.⁶

2.2.2. Voluntary churners:

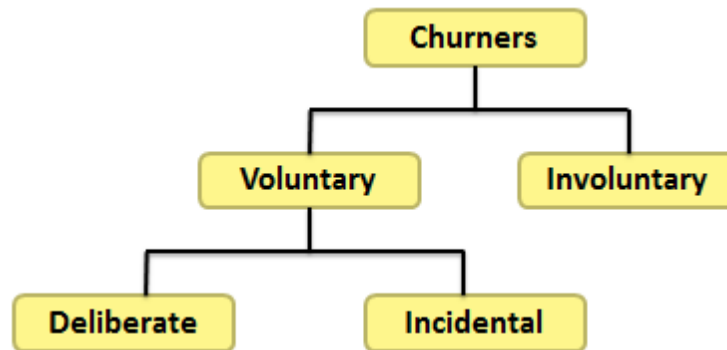
Voluntary churner is more difficult to identify; it happens when a customer decides to end service with the provider, it can be sub-divided into two main categories, incidental churn and deliberate churn, as shown in Figure 3.⁷

- **Incidental churn:** occurs when something happened in the consumers' lives rather than because they planned on it. For example, churn due to a change in financial situation, churn due to a change in venue, etc.
- **Deliberate churn:** occurs for various reasons, including technology (customers seeking newer or improved technology), economics (price sensitivity), service quality, social or psychological factors, and convenience.

⁶ Shaaban, E., Helmy, Y., Khder, A.E., & Nasr, M.M.: «A Proposed Churn Prediction Model», **International Journal of Engineering Research and Applications**, Vol. 02, Issue. 04, 2012, pp. 693

⁷ Ibid. pp. 693

Figure 3: Churn taxonomy



Source: Shaaban, E., Helmy, Y., Khder, A.E., & Nasr, M.M.: «A Proposed Churn Prediction Model», International Journal of Engineering Research and Applications, Vol. 02, Issue. 04, 2012, pp. 693

2.3. Importance of churn

Churn is a major problem for many businesses because it demonstrates how good (or bad) they are at keeping customers on board, and according to Forbes, it costs FIVE TIMES MORE to attract new clients than it does to retain existing ones.⁸

Here are some statistics about customer retention according to multiple studies conducted around the world:

- According to a Harvard Business School report, a 5% rise in consumer retention rates results in a 25% – 95% increase in earnings. And loyal clients account for the lion's share of a company's revenue – 65%.⁹
- Customer churn costs \$1.6 trillion per year due to poor service in the US.¹⁰
- According to KPMG, consumer retention is the primary driver of a company's revenue.¹¹, as shown in Figure 4.

⁸ Forbes, <https://www.forbes.com/sites/jiawertz/2018/09/12/dont-spend-5-times-more-attracting-new-customers-nurture-the-existing-ones/?sh=316cb5035a8e/> , [Cited: 25 May 2021]

⁹ Harvard Business School, <https://hbswk.hbs.edu/archive/the-economics-of-e-loyalty> , [Cited: 25 May 2021]

¹⁰ Accenture, <https://newsroom.accenture.com/news/us-companies-losing-customers-as-consumers-demand-more-human-interaction-accenture-strategy-study-finds.htm> , [Cited: 25 May 2021]

¹¹ KPMG, <https://home.kpmg/xx/en/home/insights/2020/01/home.html> , [Cited: 25 May 2021]

Figure 4: Most significant retail revenue drivers

MOST SIGNIFICANT RETAIL REVENUE DRIVERS



Source: <https://www.superoffice.com/blog/reduce-customer-churn/>

2.4. Factors of churn

Because of the relation between switching behavior and profitability, marketing academics have long been interested in consumer switching behavior, ten variables that are most likely to cause customers to switch suppliers ¹²:

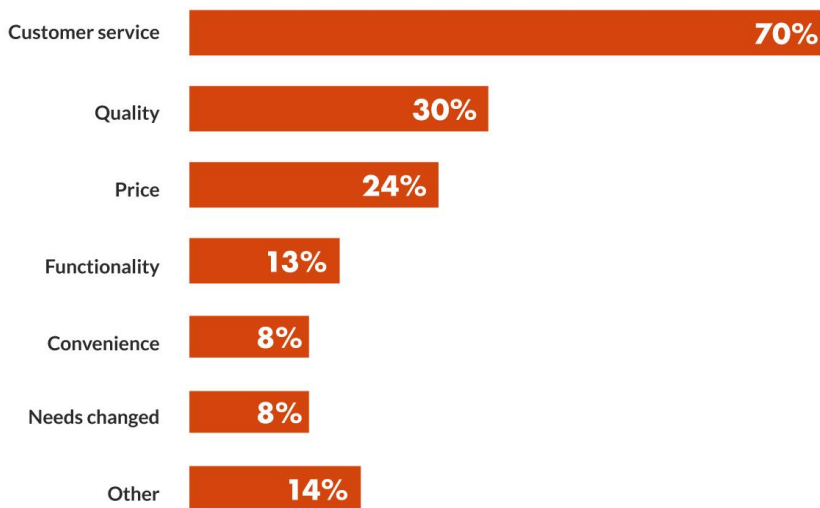
- Service encounter failure.
- Pushiness of direct sales people.
- Unreliability of direct sales people.

¹² Msweli, Pumela.: « Modeling Switching Behaviour of Direct Selling Customers.», **South African Journal of Economic and Management Sciences**. Vol. 7, Issue. 2, 2004, pp. 274

- The ethics of direct sales people.
- Rebate errors.
- Responsiveness to service failure.
- The quitting behavior of sales people.
- Competitors.
- Risk.
- Pricing of direct sales products.

Figure 5: Factors of churn

WHY DO CUSTOMERS LEAVE? (CUSTOMER VIEW)



Source: <https://www.superoffice.com/blog/reduce-customer-churn/>

2.5. Methods of reducing churn

There are several methods in marketing theory to reduce customer churn; many of these methods were developed in recent years with the emergence of online products and services, in which companies have more access to customer data, some of these methods are:

- **Investigate the causes of churn:** This can be done by conducting surveys with the customers across different platforms, and one of the best options is using phone calls. This way, you can demonstrate that you genuinely care, since studies show that approximately 68% of customers leave because they believe the company does not care about them.
- **Engage with your customers:** Give your customers reasons to keep coming back by demonstrating the day-to-day value of using your products, a practice known as relationship marketing.
- **Determine possible churners:** The most effective way to avoid churn is to prevent it from occurring in the first place, studies in the last years emphasize on using data mining and machine learning techniques to discover the most likely customers to leave.
- **Improve your services:** Obviously, giving better services is one of the best methods of keeping customers, because according to Forum Corporation's research¹³, churn due to poor service accounts for 70% of all churn, as illustrated in Figure 5.
- **Offer incentives:** Another great suggestion is to provide incentives, such as discounts and special offers, to customers who have been identified as being likely to defect, some of the techniques used in this area are illustrated in Figure 6. It is critical to note that offering incentives and discounts is widely regarded as the most effective tactic for reducing churn.

¹³ Achieveforum, <https://www.achieveforum.com/resources-research> , [Cited: 26 May 2021]

Figure 6: Tactics for reducing churn

Effective Tactics Used to Reduce Customer Churn



Source: <https://www.superoffice.com/blog/reduce-customer-churn/>

3. RFM Analysis

Direct marketing is, at its core, the technical knowledge of customer acquisition and contact. The central question is whether customer A deserves to be contacted again based on their previous purchase history. This question is equally applicable to direct mail, catalog, phone, field, or Internet contact. Customer segmentation is the process by which this decision is made.¹⁴

Not all customers purchased the same amount. Some have ordered more frequently, while others have ordered more recently. As a result, not all customers should be contacted with the same level of effort and expense. Recency, frequency, and monetary values are the foundations of direct marketing segmentation (RFM).¹⁵

3.1. RFM Basics

RFM analysis [5] is a three-dimensional method of categorizing or evaluating clients in order to discover the top 20%, or best, customers. It is based on the 80/20 principle, which states that 20% of consumers generate 80% of income.¹⁶

A customer segmentation model known as the pyramid model is used to group customers and do analysis. The pyramid model categorizes clients based on the revenue they generate, as seen in Figure 7. These value segments or categories are then employed in a variety of analytics. This method has the advantage of focusing the analytics on categories and terms that are immediately relevant to the business.¹⁷

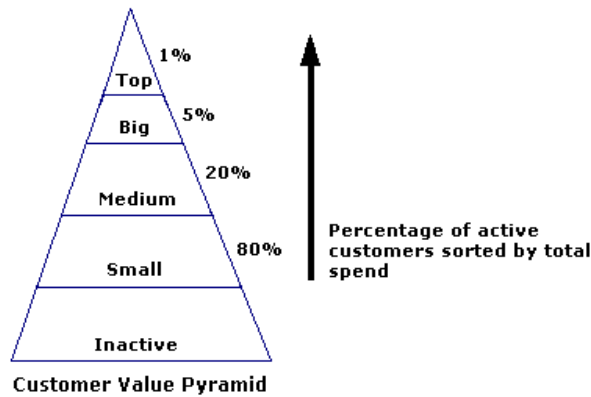
¹⁴ J R Miglautsch: Thoughts on RFM scoring, **Journal of Database Marketing & Customer Strategy Management**, Vol. 08, Issue. 1, 2000 , pp. 67

¹⁵ Ibid., pp. 67

¹⁶ Aggelis, V., & Christodoulakis, D.: Customer clustering using rfm analysis, **In Proceedings of the 9th WSEAS International Conference on Computers**,2005, pp.0 2.

¹⁷ Ibid., pp. 02

Figure 7: Pyramid model



Source: Aggelis, V., & Christodoulakis, D.: Customer clustering using rfm analysis, In Proceedings of the 9th WSEAS International Conference on Computers, pp. 02

3.2. RFM Scoring

The goal of RFM scoring is to forecast future behavior (in order to make better segmentation decisions). To allow for projection, it is necessary to translate customer behavior into numbers that can be used over time.¹⁸

Essentially, RFM analysis suggests that a customer with a high RFM score should normally conduct more transactions, resulting in a higher profit for the company¹⁹.

3.3. Segmentation

Clustering algorithms are categorized as undirected data mining tools. they are used to combine observable examples into clusters (groups).²⁰

K-means is the simplest and most used clustering algorithm. This algorithm takes as input a specified number of clusters, denoted by the letter k in its name. Mean is an abbreviation for average, as in the average location of all members of a specific cluster.²¹

¹⁸ J R Miglatsch: op. cit., pp. 67

¹⁹ Aggelis, V., & Christodoulakis, D.: op. cit., pp. 01

²⁰ Ibid., pp. 02

²¹ Ibid., pp. 02

4. Data Mining

With the significant increase in data and competitiveness among organizations over the years, data mining has emerged in order to gain insights from these massive datasets so managers can accurately predict the future of their businesses.

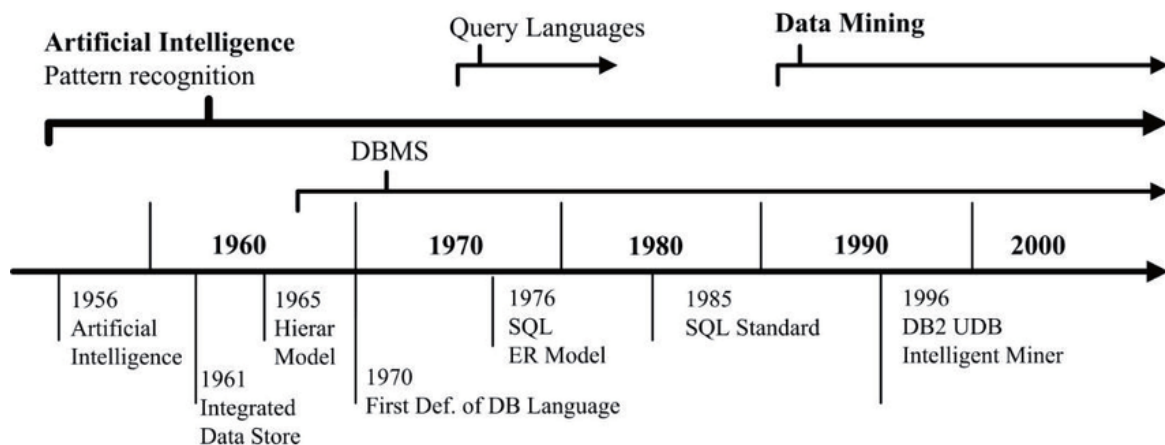
4.1. Definition

Data mining is the process of submitting various queries and pulling relevant information, patterns, and trends from enormous amounts of data that may be stored in databases.²²

4.2. History of Data Mining

Data mining can be traced back to the creation of artificial intelligence in the 1950s.²³ The evolution of data mining is depicted in Figure 8.

Figure 8: History of data mining development



Source: Gordan, Meisam & Ismail, Zubaidah & Ibrahim, Zainah & Hashim, Huzaifa: *Data Mining Technology for Structural Control Systems: Concept, Development, and Comparison*, intechopen, pp.03

²² Bhavani, Thuraisingham, :“**Data Mining: Technologies, Techniques, Tools and trends**”,1st. ed., CRC Press,1998, pp. 01

²³ Gordan, Meisam & Ismail, Zubaidah & Ibrahim, Zainah & Hashim, Huzaifa :**Data Mining Technology for Structural Control Systems: Concept, Development, and Comparison**, intechopen, pp.03

4.3. Process of Data Mining

Many people confuse data mining with another popular phrase, "Knowledge Discovery in Databases," or KDD. Others, on the other hand, see data mining as merely a necessary stage in the process of knowledge discovery in databases. Knowledge discovery is an iterative process that includes the following steps²⁴:

- **Data cleansing:** This is also referred to as data cleansing. Because we are dealing with irrelevant data at this phase.
- **Data integration** is the process of merging numerous data sources.
- **Data selection:** We must choose the data that is important to the analysis and extract it from the data collection.
- **Data transformation:** it is a step in which the selected data is turned into forms. That is suitable for the mining technique.
- **Data mining:** We must employ creative approaches to extract potentially relevant patterns.
- **Pattern evaluation:** Using specified metrics, interesting patterns reflecting knowledge are recognized in this procedure.
- **Knowledge presentation:** Knowledge is discovered and represented to the user, particularly during this phase. This critical stage employs visualization approaches. This aids users in comprehending and interpreting data mining results.

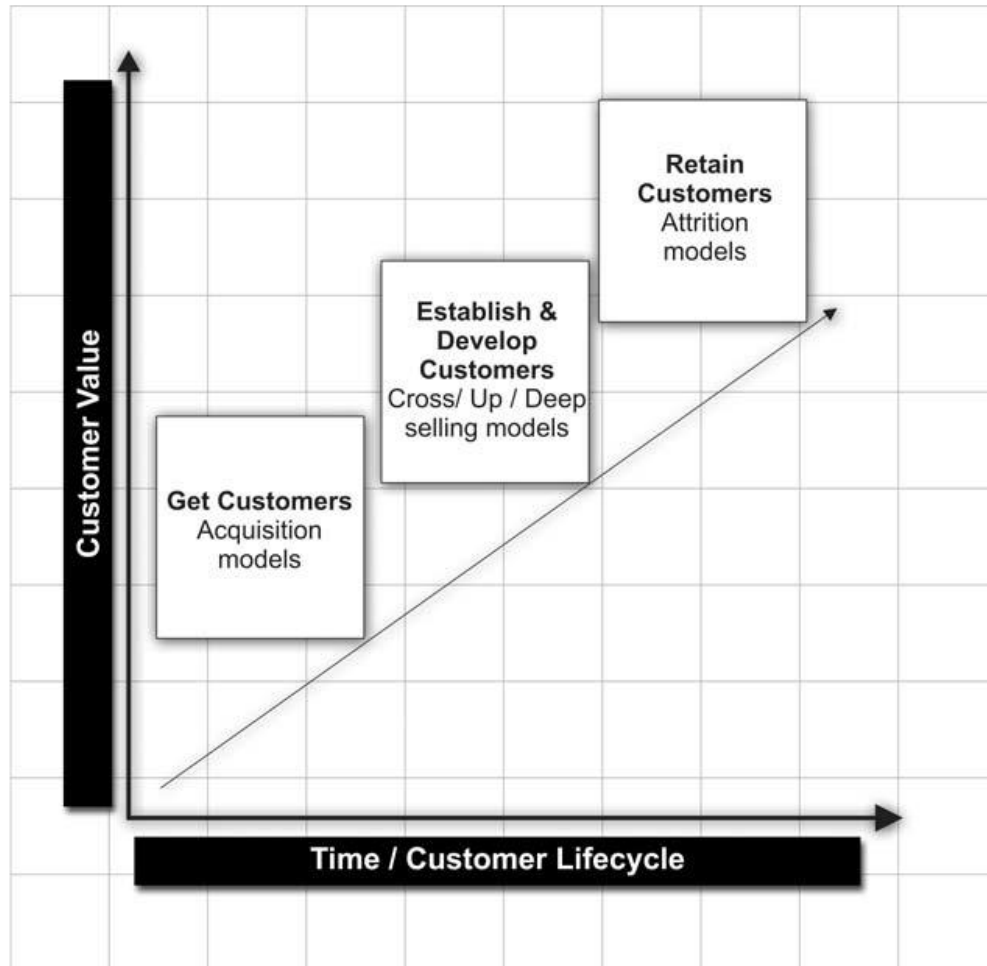
4.4. Data mining in the CRM framework

Data mining can provide valuable customer insight, which is essential for developing an efficient CRM strategy, it can support an 'individualized' and optimized customer management throughout all stages of the customer lifecycle, from acquisition and

²⁴ Han, Jiawei, Micheline Kamber, and Jian Pei : **Data Mining: Concepts and Techniques**, 3rd ed, Morgan Kaufmann Publishers, 2012, p.06

establishing a better relationship to preventing attrition and regaining lost customers, As demonstrated in Figure 9 , data mining models can assist in all of these activities.²⁵

Figure 9: Data mining and customer lifecycle management.



*Source: Konstantinos Tsipstis, Antonios Chorianopoulos: **Data Mining Techniques in CRM: Inside Customer Segmentation**, 1st. ed, Wiley, 2010, pp. 05*

²⁵ Konstantinos Tsipstis, Antonios Chorianopoulos: **Data Mining Techniques in CRM: Inside Customer Segmentation**, 1st. ed, Wiley, 2010, pp. 04

More specifically, the marketing activities that can be aided by data mining include Customer segmentation, Direct marketing campaigns, Market basket and Sequence analysis.²⁶

4.4.1. Customer segmentation:

Segmentation is a process of splitting a clientele into distinct and internally similar groups in order to build distinctive marketing strategies based on their characteristics. There are numerous segmentation methods based on the precise criteria or attributes utilized for segmentation, clustering algorithms can examine behavioral data, find natural consumer groupings, and provide a solution based on observed data patterns. If data mining models are effectively constructed, they can find groups with distinct profiles and features, leading to rich segmentation schemes with business meaning and value.²⁷

4.4.2. Direct marketing campaigns:

Direct marketing campaigns are used by marketers to send a message to their customers by mail, the Internet, e-mail, telemarketing (phone), and other direct channels in order to prevent churn (attrition) and drive customer acquisition and purchase of add-on items. Acquisition efforts, in particular, seek to attract new and potentially important clients away from the competitors, retention programs strive to keep key clients from leaving the firm.²⁸

To optimize subsequent marketing campaigns, the classification models listed below are used:

- **Acquisition models:** These can be used to identify potentially profitable new consumers by locating “clones” of valuable existing customers in external contact databases.
- **Cross-/deep-/up-selling models:** These can expose existing clients' purchasing power.

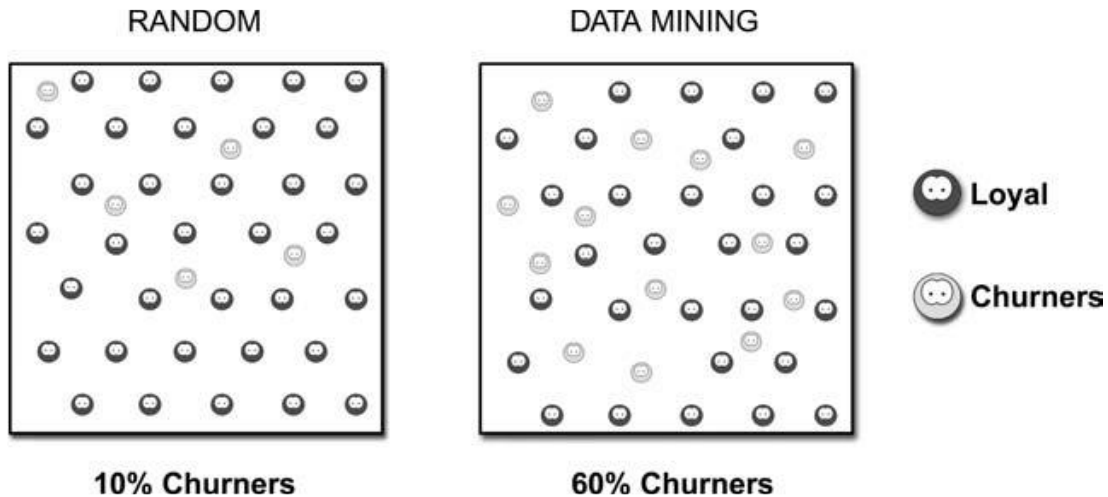
²⁶ Konstantinos Tsitsis, Antonios Chorianopoulos: Op. cit., pp. 04-05

²⁷ Ibid., pp. 05

²⁸ Ibid., pp. 05-06

- **Voluntary attrition or churn models:** These detect early churn signals and identify such clients with an increased likelihood to leave voluntarily, the efficiency of such models is demonstrated in Figure 10.

Figure 10: The increase in predictive ability using data mining



Source: Konstantinos Tsipstis, Antonios Chorianopoulos: Data Mining Techniques in CRM: Inside Customer Segmentation, 1st. ed, Wiley, 2010, pp. 07

4.4.3. Market basket and Sequence analysis:

Data mining and association models in particular can be used to identify related products typically purchased together. These models can be used for market basket analysis and for revealing bundles of products or services that can be sold together.

5. Machine Learning

5.1. Definition

Machine learning is about designing algorithms that automatically extract valuable information from data.²⁹

It studies algorithms and techniques in order to automate solutions to complex problems that are hard to program using conventional programming methods,

Machine Learning typically tries to extract the most relevant patterns from vast data using various statistical and machine learning methods. It usually involves several procedures such as data cleaning, data integration, data selection, data transformation, pattern discovery, pattern evaluation, and knowledge presentation.³⁰

5.2. History of Machine Learning

One of the first use cases of the "Machine Learning" expression was in 1959, after the publication of a paper by IBM in the IBM Journal of Research and Development Authored by IBM's Arthur Samuel, where the paper invested the use of machine learning in the game of checkers “to verify the fact that a computer can be programmed so that it will learn to play a better game of checkers than can be played by the person who wrote the program”³¹

Arthur Samuel is widely considered as the one who used and defined machine learning in the form we now know today, although it was not the first publication to use the term “machine learning” per se.³²

²⁹ Deisenroth, Marc Peter ; Faisal, A. Aldo ; Ong, Cheng Soon: **Mathematics for Machine Learning** : Cambridge University Press, 2020, p.11

³⁰ Han, Jiawei, Micheline Kamber, and Jian Pei : op. cit., 2012, p.01

³¹ Arthur Samuel, Some Studies in Machine Learning Using the Game of Checkers, **IBM Journal of Research and Development**, Vol. 3, Issue. 3, 1959.

³² Oliver Theobald : **Machine learning for absolute beginners : a plain English introduction**, 2nd ed, Scatterplot Press, 2017, pp. 08

5.3. Use cases of Machine Learning

Machine learning admits an extensive set of practical applications, which includes the following:

- **Product recommendation:** most of the recommendation systems of companies like Amazon and YouTube are built using machine learning
- **Personalized marketing:** Marketing is becoming more digitalized every day, especially in online marketing, where marketers can use ML to target potential customers based on their characteristics
- **Spam detection:** Where companies use ML to detect and prevent spammed emails from being opened by the users.
- **Fraud detection:** Companies use ML to prevent the occurrence of fraudulent transactions in the banking industry.
- **Voice assistants:** Companies like Google and Apple use advanced ML techniques in their voice assistants' programs to make them understand what the consumer says
- **Process automation:** Many processes in the enterprise can be automated using ML and therefore increase efficiencies, such as risk assessments, demand forecasting, customer churn prediction, and others.
- **Dynamic pricing:** Which is used often in the travel industry, such as flying companies and hotels, they use ML algorithms to set the prices based on the time of booking, where most customers know that the sooner, the cheaper.

This list is by no means exhaustive, and there are much more cases where we could use ML and get outstanding results.³³

³³ Algorithmia, <https://algorithmia.com/blog/machine-learning-use-cases> , [Cited: 28 april 2021.]

Conclusion

Today, every business is affected by **Churn**. Indeed, all products are constantly developed and improved, and with the progression of market acquisitions, one can say that all companies are on the same level in terms of product quality, which encourages customers to switch brands easily and not be intimidated by the fact that they are loyal to a specific brand.

Customers are an organization's most valuable asset. There can be no commercial opportunities unless there are satisfied clients who remain loyal and grow their relationship with the organization, **CRM** (Customer Relationship Management) is an approach for developing, managing, and sustaining long-term customer connections.

To fully understand **Churn** management, this study suggests that the primary influencing factors for churn are those related to competition, technology, regulators, and customers.

The techniques used for customer retention may vary from customer to customer, that's why segmenting customers is crucial in dealing with churn, **RFM** analysis is a technique used to determine the top most profitable customers to the company.

Data Mining and **Machine Learning** are two quite similar disciplines that allow the full exploitation of the data and discover hidden patterns and insights that may benefit businesses greatly.

The following chapter will discuss **Data Mining** and **Machine Learning** strategies for understanding and forecasting future client behavior.

CHAPTER II: DATA MINING & MACHINE LEARNING PROCESSES

CHAPTER II

DATA MINING & MACHINE LEARNING PROCESSES

Introduction

Machine Learning is gaining popularity in diverse enterprise environments due to the exponential rise in data volume and the rapid spread of data in this era of the internet and social media. Organizations should make strategic decisions using a complete data collection, including organized and unstructured data from internal and external sources.

This chapter will look at the three stages of developing a predictive model: **Data Preprocessing**, **Modeling Process**, and **Model Evaluation**.

Firstly, we will see the different techniques used in the **Data Preprocessing** stage and examine the data mining and machine learning industry standards.

Secondly, according to previous studies, we will dive into the Modeling Process and study the state-of-the-art algorithms in unsupervised and supervised learning, particularly the most efficient algorithms used in customer churn prediction.

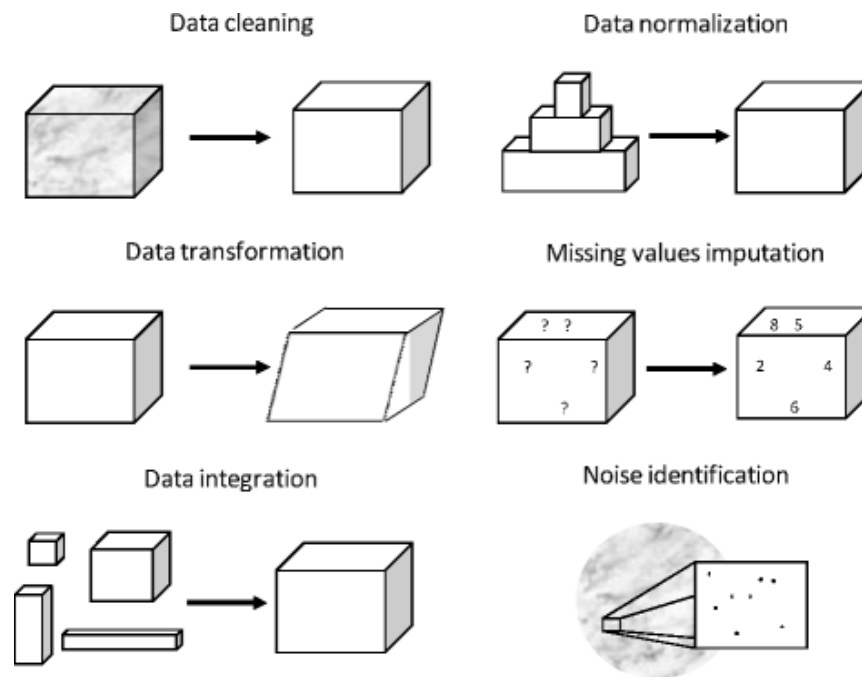
Lastly, we will look at the different methods and techniques used in **Model Evaluation** and **Model Interpretation**.

1. Data Preprocessing Procedures

Data preprocessing is increasingly becoming one of the hottest topics in machine learning and data mining, more and more researchers are paying more attention to data preprocessing to improve their models due to the importance of these techniques in increasing the performance of ML models.

Since most data will likely be imperfect, containing inconsistencies and redundancies which makes it unusable in ML models³⁴, that's where data preprocessing becomes handy using different techniques such as data cleaning, normalization, transformation, missing data imputation, sampling, feature selection and many other methods to improve the quality of the data, as shown in Figure 11.

Figure 11: Data preprocessing tasks



Source: Chu Xu.: Data Cleaning. In: Sakr S., Zomaya A.Y. (eds) Encyclopedia of Big Data Technologies. Springer, 2019, pp. 01

³⁴ García, S., Ramírez-Gallego, S., Luengo, J. et al. :Big data preprocessing: methods and prospects, **Big Data Analytics**, Vol. 1, Issue. 9, 2016, pp. 03

After the application of a successful data preprocessing stage, the final data set obtained can be regarded as a reliable and suitable source for any machine learning algorithm applied afterwards.

1.1. Data cleaning, Normalization and Transformation

1.1.1. Data cleaning:

Data cleaning is used to refer to all kinds of tasks and activities to detect and repair errors in the data.³⁵

According to the principle of garbage in, garbage out (GIGO for short), bad data can truly harm machine learning. Bad data consists of missing data, outliers, skewed value distributions, redundancy of information, and features not well explicated³⁶.

1.1.2. Normalization:

Some machine learning algorithms that are based on distance measures or gradient descent such as K-means and K-Nearest-Neighbors are sensitive to the scale of the numeric values, rescaling the distribution becomes crucial,

Rescaling transforms the range of the values of the variables to prevent bias towards values that are on different scales, this can be done in two ways:

- **Statistical standardization (z-score normalization):** This is done by subtracting the mean and then dividing the result by the standard deviation, this will likely transform most of the values to the range between -3 and 3.

$$X' = \frac{X - \text{mean}}{\text{std_dev}}$$

- **Min-Max transformation:** This is done by subtracting the minimum value of the feature and then dividing by the range of the same feature (which is the maximum value minus the minimum value), this will rescale all the values to a

³⁵ Chu Xu. : **Data Cleaning**. In: Sakr S., Zomaya A.Y. (eds) Encyclopedia of Big Data Technologies. Springer, 2019, pp. 01

³⁶ John Paul Mueller, Luca Massaron : **Machine Learning For Dummies** , New Jersey, John Wiley & Sons Inc, 2016, pp. 220

range between 0 and 1.

This method gives better results compared to standardization when the standard deviation of the feature is too small.³⁷

$$X' = \frac{X - \min}{\max - \min}$$

1.1.3. Transformation (feature construction / engineering):

Sometimes, the raw data obtained from various sources may not include the features needed to implement ML algorithms, we can surpass this problem by creating our own features to get the desired output, these features will be constructed from existing data.³⁸

The construction of new features of a basic feature is a technique called feature engineering/construction, the new generated features may lead to the creation of more concise and accurate classifiers. In addition, the discovery of meaningful features contributes to the better comprehensibility of the produced classifier, and a better understanding of the learned concept³⁹

1.2. Missing data⁴⁰

In most cases, data sets wouldn't be complete and there would be some missing data, missing values make it difficult for ML algorithms to learn during the training process, some learning algorithms do not know how to deal with missing values and report errors in both training and test phases, that's why something must be done with missing data.

Sometimes, you can just delete or repair missing data just by guessing a likely replacement value, a good rule of thumb is to drop a feature if more than 90% of its values are missing.

There are multiple strategies to deal with missing data, these strategies may differ when dealing with qualitative or quantitative variables, some of these strategies are listed below:

³⁷ John Paul Mueller, Luca Massaron : op. cit., pp. 226 - 227

³⁸ Ibid., pp. 227

³⁹ Sotiris Kotsiantis, Dimitris Kanellopoulos, P. E. Pintelas: Data Preprocessing for Supervised Learning. **International Journal of Computer Science**. Vol. 1, Issue. 1, 2006, pp. 115.

⁴⁰ John Paul Mueller, Luca Massaron : op. cit., pp. 221 - 223

- Replace missing values with a computed constant like the mean or the median value
- Replace missing values with 0, which works well with regression models and standardized variables
- Interpolate the missing values when they are part of a series of values tied to time
- Impute their value using the information from other predictor features

1.3. Sampling

“Class imbalances” is a phenomenon that is present often in customer churn prediction cases,

However, when building a model with this kind of imbalanced data it leads to problems such as improper evaluation metrics, lack of data (absolute rarity), relative lack of data (relative rarity), data fragmentation, inappropriate inductive bias, and noise.⁴¹

The random sampling method is used to change the data distribution in order to minimize the imbalance class problem which is caused due to the lack of availability of data. It minimizes the imbalance data distribution which is caused due to unavailability of the target class data in the dataset (i.e., customer churn).⁴²

1.4. Feature selection

Feature and variable selection are methods for extracting as much information as possible from a large number of variables. Since the number of variables and data has grown as a result of more sophisticated data collection, it is important to include only the most critical and usable variables in the model being built. The key goals of selection are to improve predictive efficiency, make predictions faster and more efficiently, and gain a deeper and more accurate understanding of the predictive process. Adding extra variables to the model adds complexity and can lead to overfitting, but leaving out critical variables results in lower predictive efficiency. And because of its simplicity, scalability, and

⁴¹ Burez, J. and Van den Poel, D. :Handling class imbalance in customer churn prediction, **Expert Systems with Applications**, Vol. 36, 2009 pp. 4626–4627.

⁴² A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar: Customer churn prediction in telecommunication industry using data certainty, **Journal of Business Research**, Vol 94, 2019, pp. 292-293

empirical success, many variable selection algorithms use variable ranking as a primary or auxiliary selection mechanism.⁴³

There are three main objectives of feature selection:

- improving the prediction performance of the predictors.
- providing faster and more cost-effective predictors.
- providing a better understanding of the underlying process that generated the data.

1.5. Train – Test split

In an ideal world, you'd be able to run a test on data that your machine learning algorithm has never seen before. Waiting for new data, on the other hand, isn't always feasible in terms of time and money. You can divide the data into training and test sets at random as a first easy solution, the typical split is between 25 and 30% for testing and 75 to 70% for training.

When you need to fine-tune your learning algorithm, the second remedy comes into play. A third break, known as a validation set, is needed to prevent snooping. A suggested division is to divide the examples into thirds: 70% for training, 20% for validation, and 10% for testing.⁴⁴

⁴³ Guyon, I., Elisseeff, A., & Kaelbling, L. P. (Ed.) : An introduction to variable and feature selection. **Journal of Machine Learning Research**, Vol 3, Issue 7, 2003, pp. 1158-1159

⁴⁴ Ibid., pp. 191

2. Modeling Process

After raw data has been prepared and preprocessed, the data obtained is now ready for modeling.

But choosing the right algorithm or combination of algorithms for the job is a constant challenge for anyone working in the area of machine learning, which includes hundreds of statistical-based algorithms, it is essential to understand the three main categories of machine learning, these three categories are unsupervised, supervised, and reinforcement learning.⁴⁵ reinforcement learning is out of the scope of this study.

2.1. Unsupervised Learning:

Unsupervised learning occurs when an algorithm learns from plain examples without any associated response. This type of algorithm tends to restructure the data into something else, such as new features or a new series of uncorrelated values. Some recommendation systems that you find in the form of marketing automation are based on this type of learning.⁴⁶

The main technique that we will be using in this study is called clustering, more specifically K-means algorithm.

2.1.1. Clustering:

Clustering refers to a wide range of techniques for identifying subgroups or clustering clusters in a data set, when we cluster the observations in a data set, we try to divide them into discrete groups so that the observations within each group are relatively similar to each other, but the observations in other groups are considerably different, since clustering is popular across many disciplines, there are numerous clustering algorithms. We concentrate on the most well-known clustering algorithm: Clustering with K-means.⁴⁷

⁴⁵ Oliver Theobald, op. cit., pp. 15

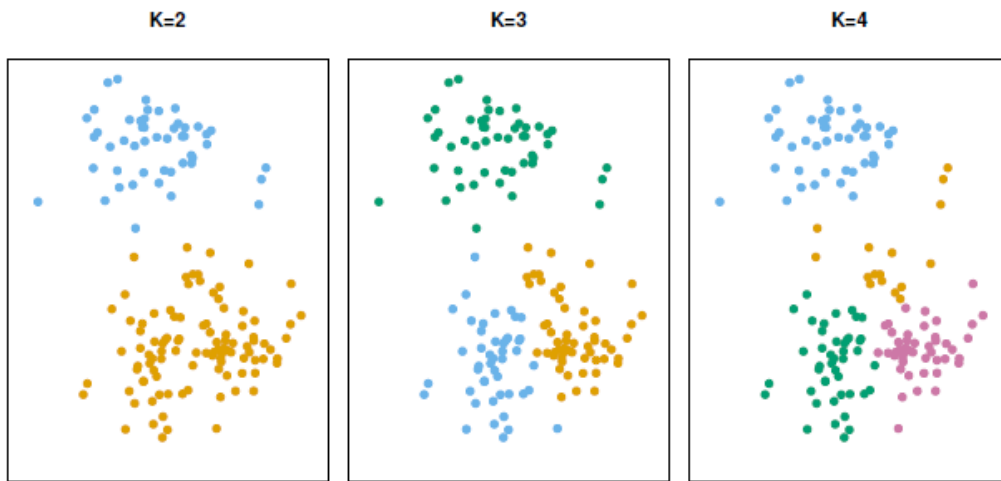
⁴⁶ John Paul Mueller, Luca Massaron : op. cit., pp. 169

⁴⁷ James, G., Witten, D., Hastie, T., & Tibshirani, R.: « **An introduction to statistical learning** », 2nd. ed, Springer, 2021, pp. 516-517.

2.1.2. K-means:

K-means clustering is a straightforward and elegant method for separating a data collection into K unique, non-overlapping clusters. To do K-means clustering, we must first define the number of clusters K ; the K-means algorithm will then assign each observation to exactly one of the K clusters. Figure 12 depicts the results of K-means clustering on a simulated example with 150 observations in two dimensions and three distinct values of K .⁴⁸

Figure 12. K-means clustering on 150 simulated point



Source : James, G., Witten, D., Hastie, T., & Tibshirani, R.: « An introduction to statistical learning », 2nd. ed, Springer, 2021, pp. 517.

The K-means clustering procedure results from a simple and intuitive mathematical problem. We begin by defining some notation. Let C_1, \dots, C_K denote sets containing the indices of the observations in each cluster. These sets satisfy two properties:⁴⁹

1. $C_1 \cup C_2 \cup \dots \cup C_K = \{1, \dots, n\}$. In other words, each observation belongs to at least one of the K clusters.
2. $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$. In other words, the clusters are nonoverlapping: no observation belongs to more than one cluster.

⁴⁸ James, G., Witten, D., Hastie, T., & Tibshirani, R.: op. cit., pp. 517

⁴⁹ Ibid. pp. 518

For instance, if the i^{th} observation is in the k^{th} cluster, then $i \in C_k$. The idea behind K-means clustering is that a good clustering is one for which the within-cluster variation is as small as possible. The within-cluster variation for cluster C_k is a measure $W(C_k)$ of the amount by which the observations within a cluster differ from each other. Hence, we want to solve the problem:

$$\text{minimize } C_1, \dots, C_K = \sum_{k=1}^K W(C_k)$$

In words, this formula says that we want to partition the observations into K clusters such that the total within-cluster variation, summed over all K clusters, is as small as possible. Solving the equation seems like a reasonable idea, but in order to make it actionable we need to define the within-cluster variation. There are many possible ways to define this concept, the most common choice involves squared Euclidean distance:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2$$

where $|C_k|$ denotes the number of observations in the k^{th} cluster. In other words, the within-cluster variation for the k^{th} cluster is the sum of all of the pairwise squared Euclidean distances between the observations in the k^{th} cluster, divided by the total number of observations in the k^{th} cluster. Combining the two previous equations gives the optimization problem that defines K-means clustering,

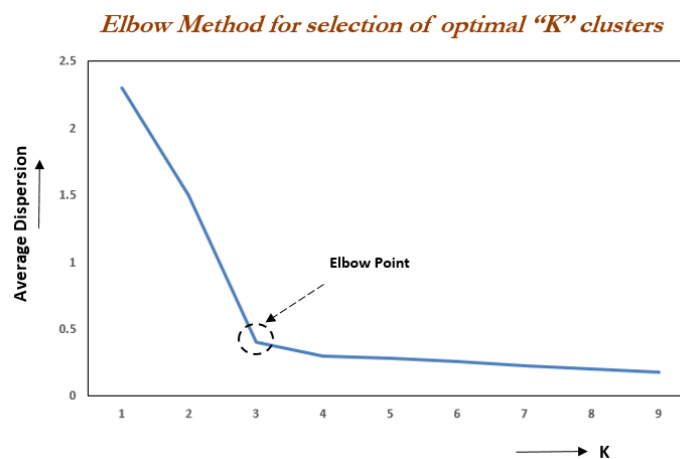
$$\text{minimize } C_1, \dots, C_K = \sum_{k=1}^K \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2$$

There are several methods or approaches to dealing with the problem of determining the best number of clusters, the most prominent one is called the Elbow Method.⁵⁰

⁵⁰ Nanjundan, S., Sankaran, S., Arjun, C.R., & Anand, G. « Identifying the number of clusters for K-Means: A hypersphere density based approach ». **ArXiv**, 2019, pp. 02.

In k-means clustering, the elbow method is used to estimate the ideal number of clusters. The elbow method depicts the value of the cost function as a function of k . When you may be aware, as k increases, the average distortion decreases, each cluster has fewer constituent instances, and the instances are closer to their respective centroids. However, as k grows, the improvements in average distortion diminish. The value of k at which the improvement in distortion decreases the most is known as the elbow, and it is the point at which we should cease dividing the data into further clusters⁵¹, as shown in Figure 13

Figure 13: The Elbow Method



Source: Pratap Dangeti: « Statistics for Machine Learning », Packt Publishing, 2017, pp. 313

2.2. Supervised Learning:

The learner receives a set of labeled examples as training data and makes predictions for all unseen points. This is the most prevalent case encountered while dealing with classification, regression, and ranking problems. The churn and spam detection problems are examples of supervised learning.⁵²

A supervised learning algorithm's purpose is to use the dataset to create a model that accepts a feature vector x as input and returns information that allows the label for this feature vector to be deduced. For example, a model built with the people dataset could take

⁵¹ Pratap Dangeti: « **Statistics for Machine Learning** », Packt Publishing, 2017, pp. 313

⁵² Mohri, M., Rostamizadeh, A, Talwalkar, A : « **Foundations of Machine Learning** », MIT Press, 2018, pp. 06.

as input a feature vector characterizing a person and output a probability that the individual has cancer.

Classification and regression are the two major categories of supervised machine learning issues. Since churn prediction is a classification problem, we will be focusing more on the classification algorithms.

2.2.1. Regression:

Regression has a long history in a variety of fields, including statistics, economics, psychology, social sciences, and political science. Linear regression, in addition to being capable of a wide range of predictions incorporating numeric values, binary and multiple classes, probabilities, and count data, may also help you analyze group differences, model consumer preferences, and quantify the importance of a feature in a model.⁵³

Regression analysis is concerned with the study of the dependence of one variable, the dependent variable, on one or more other variables, the explanatory variables, with a view to estimating and/or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the latter.⁵⁴

2.2.2. Classification:

The linear regression model is based on the assumption that the response Y is quantitative. However, in many cases, the response variable is qualitative. The techniques for predicting qualitative responses are known as classification. The methods for classifying observations first predict the likelihood that the observation belongs to each of the categories as the basis for making the classification.⁵⁵

There are many classification techniques, or classifiers, that one might use to predict a qualitative response, some widely-used classifiers are: Logistic Regression, Linear

⁵³ John Paul Mueller, Luca Massaron : op. cit., pp. 258

⁵⁴ Gujarati, Damodar N., Porter, Dawn C.: « **Basic econometrics** », McGraw Hill, 5th. Ed., New York, 2009, pp. 15

⁵⁵ James, G., Witten, D., Hastie, T., & Tibshirani, R.: op. cit., pp. 129

Discriminant Analysis, Decision Trees, Neural Networks, SVM, Naive Bayes, and K-nearest neighbors(kNN) ⁵⁶, some of these techniques are listed in Table 1.

Table 1: Comparing learning algorithms (**** stars for best, * star for worst)

	Decision Trees	Neural Networks	Naïve Bayes	kNN	SVM	Rulelearners
Accuracy in general	**	****	*	**	****	**
Speed of learning with respect to number of attributes and the number of instances	****	*	****	****	*	**
Speed of classification	****	****	****	*	****	****
Tolerance to missing values	****	*	****	*	**	**
Tolerance to irrelevant attributes	****	*	**	**	****	**
Tolerance to redundant attributes	**	**	*	**	***	**
Tolerance to highly interdependent attributes (e.g. parity problems)	**	****	*	*	***	**
Dealing with discrete/binary/continuous attributes	****	*** (not discrete)	*** (not continuous)	*** (not directly discrete)	** (not discrete)	*** (not directly continuous)
Tolerance to noise	**	**	***	*	**	*
Dealing with danger of overfitting	**	*	***	***	**	**
Attempts for incremental learning	**	****	****	****	**	*
Explanation ability/transparency of knowledge/classifications	****	*	****	**	*	****
Model parameter handling	****	*	****	****	*	***

Source : Kotsiantis, Sotiris. : Supervised Machine Learning: A Review of Classification Techniques.

Informatica, Vol. 31, 2007, pp. 263

2.2.2.1. Decision Trees:

➤ Definition:

Leo Breiman coined the phrase "classification and regression trees," or CART for short, to refer to decision tree methods that can be used for classification or regression prediction modeling tasks. Traditionally, this method is referred to as a decision tree, while in some platforms, it is referred to as CART. The CART algorithm serves as a foundation for many other algorithms (such as bagged decision trees and random forests).⁵⁷

⁵⁶ : James, G., Witten, D., Hastie, T., & Tibshirani, R.: op. cit., pp. 129

⁵⁷ Jason Brownlee : « **Master Machine Learning Algorithms Discover How They Work and Implement Them from Scratch** », Machine Learning Mastery, 2016, pp. 99

The CART model is represented as a binary tree. In terms of algorithms and data structures. Each node represents an input variable (x) and the variable's split point (assuming the variable is a number). The tree leaf node Contains the output variable (y) used to make predictions.⁵⁸

Trees that classify instances by sorting them based on feature values are known as decision trees. Each node in a decision tree represents a feature in a classification instance, and each branch indicates a value that the node can adopt. Instances are categorized and arranged according on their feature values, beginning with the root node⁵⁹,. Figure 14 is an example of a decision tree for the training set of Table 2.

Table 2: Training set

at1	at2	at3	at4	Class
a1	a2	a3	a4	Yes
a1	a2	a3	b4	Yes
a1	b2	a3	a4	Yes
a1	b2	b3	b4	No
a1	c2	a3	a4	Yes
a1	c2	a3	b4	No
b1	b2	b3	b4	No
c1	b2	b3	b4	No

Source: Kotsiantis, Sotiris.: Supervised Machine Learning: A Review of Classification Techniques.

Informatica, Vol. 31, 2007, pp. 251

➤ Learn a CART Model from Data:

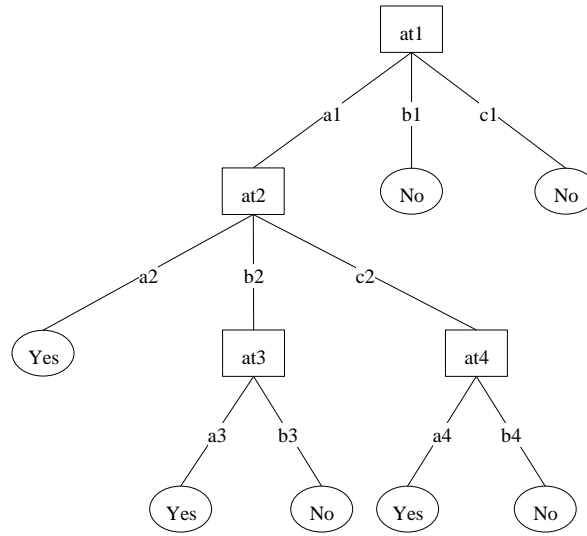
Making a binary decision tree is a technique of breaking up the input space. To divide the space, a greedy method known as recursive binary splitting is applied. This is a numerical technique in which all of the values are aligned and several split points are explored and tested using a cost function. The split with the lowest cost (since we decrease costs) is chosen. In a greedy fashion, all input variables and all feasible split points are assessed and chosen (e.g. the very best split point is chosen each time)⁶⁰.

⁵⁸ Jason Brownlee : op. cit., pp. 73

⁵⁹ Kotsiantis, Sotiris. : Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, Vol. 31, 2007, pp.251

⁶⁰ Jason Brownlee : op. cit., pp. 73

Figure 14: A Decision Tree



Source: Kotsiantis, Sotiris. : *Supervised Machine Learning: A Review of Classification Techniques*.
Informatica, Vol. 31, 2007, pp. 251

The Gini cost function is utilized for classification, which indicates how pure the leaf nodes are (how mixed the training data assigned to each node is):

$$Gini = \sum_{k=1}^n p_k \times (1 - p_k)$$

Where p_k is the number of training cases with class k in the rectangle of interest. A node with all classes of the same type will have $Gini = 0$, whereas a Gini with a 50-50 split of classes for a binary classification issue (worst purity) will have $G = 0.5$.⁶¹

➤ Stopping Criterion:

The most common stopping process is to use the least count of the number of training instances assigned to each leaf node. If the count is less than the minimum, the split is not accepted and the final leaf node is considered the final one.⁶²

⁶¹ Jason Brownlee : op. cit., pp. 73

⁶² Ibid., pp. 73

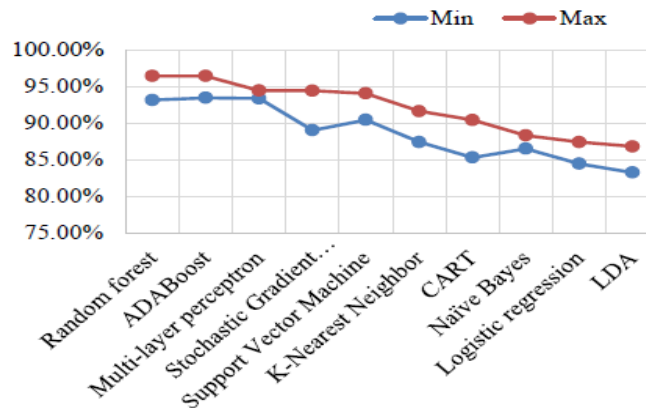
➤ Pruning The Tree:

Working through each leaf node in the tree and evaluating the effect of eliminating it using a hold-out test set is the quickest and simplest pruning strategy. More complicated methods, such as cost complexity pruning, can be used, in which a learning parameter is used to determine if nodes can be deleted based on the size of the sub-tree.

2.2.2.2. Ensemble Learning algorithms:

According to many studies, the best algorithms in terms of performance in customer churn prediction are the Ensemble Learning algorithms⁶³, as shown in Figure 15.

Figure 15: Accuracy of some models in customer churn prediction



Source: Sahar F. Sabbah :Machine-Learning Techniques for Customer Retention: A Comparative Study, *International Journal of Advanced Computer Science and Applications*, Vol.9, Issue.2, 2018, pp.279

➤ Definition:

An ensemble method is a technique that combines many simple "building block" models to create a single, potentially highly powerful model. These simple building block models are commonly referred to be weak learners because they can only make mediocre predictions on their own. Bagging, random forests and boosting are ensemble approaches with a regression or classification tree as the basic building component.⁶⁴

⁶³ Sahar F. Sabbah : Machine-Learning Techniques for Customer Retention: A Comparative Study, *International Journal of Advanced Computer Science and Applications*, Vol. 9, Issue. 2, 2018, pp. 279

⁶⁴ James, G., Witten, D., Hastie, T., & Tibshirani, R.: op. cit., pp. 340

➤ Bagging:

Decision trees suffer from a high variance, which means that if we randomly divide the training data into two parts, and we apply a decision tree to both halves, the outputs may be distinctly different, Bootstrap aggregation, or bagging, is a general-purpose procedure for reducing the variance of a statistical learning method.⁶⁵

Given a set of n independent observations Z_1, \dots, Z_n , each with variance σ^2 , the variance of the mean Z of the observations is given by σ^2/n . In other words, averaging a set of observations reduces variance, we could calculate $f1(x), f2(x), \dots, fB(x)$ using B separate training sets, and average them in order to obtain a single low-variance⁶⁶, given by:

$$f_{avg}(x) = \frac{1}{B} \sum_{b=1}^B f^b(x)$$

But this is not practical because we generally do not have access to multiple training sets, Instead, we can bootstrap, by taking repeated samples from the (single) training data set. In this approach we generate B different bootstrapped training data sets. We then train our method on the b^{th} bootstrapped training set in order to get $f^{*b}(x)$, and finally average all the predictions, to obtain the bagging equations as⁶⁷:

$$f_{bag}(x) = \frac{1}{B} \sum_{b=1}^B f^{*b}(x)$$

➤ Random Forests:

Random forests outperform bagged trees due to a random tiny modification that decorrelates the trees. On bootstrapped training samples, we construct a forest of decision trees, similar to bagging. However, when creating these decision trees, each time a split in a tree is examined, a random sample of m predictors from the whole set of p predictors is chosen as split candidates. Only one of the m predictors may be used in the split.⁶⁸

⁶⁵ James, G., Witten, D., Hastie, T., & Tibshirani, R.: op. cit., pp. 340

⁶⁶ Ibid., pp. 340

⁶⁷ Ibid., pp. 341

⁶⁸ Ibid., pp. 343

The number of features that can be looked for at each split point (m) must be specified as an algorithm parameter. You can experiment with different settings and fine-tune it through cross validation.⁶⁹

- A decent default for classification is: $m = \sqrt{p}$.
- A decent default for regression is $m = p/3$.

➤ Boosting:

Boosting, like bagging, is a general approach that may be used to a variety of statistical learning approaches for regression or classification. In this section, we will consider boosting in the context of decision trees⁷⁰.

Boosting is a broad ensemble method for constructing a strong classifier from a set of weak classifiers. This is accomplished by first developing a model from the training data, followed by the creation of a second model that seeks to rectify the faults in the first model. Models are added until the training set is properly predicted or until the maximum number of models is reached. AdaBoost was the first truly successful binary classification boosting technique. It is the ideal place to begin learning about boosting. Modern boosting methods, most notably stochastic gradient boosting machines, are based on AdaBoost.⁷¹

On a classification task, weak learners are models that achieve accuracy slightly above random chance.

Decision trees with one level are the best suited and hence most commonly used algorithm with AdaBoost. Because these trees are so short and only have one decision for classification, they are commonly referred to as decision stumps. Weights are assigned to each instance in the training dataset. The initial weight is determined as follows⁷²:

$$weight(x_i) = \frac{1}{n}$$

⁶⁹ Jason Brownlee : op. cit., pp. 128

⁷⁰ James, G., Witten, D., Hastie, T., & Tibshirani, R.: op. cit., pp. 345

⁷¹ Jason Brownlee : op. cit., pp. 136

⁷² Ibid., pp. 136-137

Where x_i is the i^{th} training instance and n is the number of training instances.

Using weighted samples, a weak classifier (decision stump) is created using the training data, because only binary (two-class) classification problems are permitted, each decision stump makes a single decision on a single input variable and returns a +1.0 or -1.0 value for the first or second class value.

The trained model's misclassification rate is computed. This is traditionally calculated as:

$$error = \frac{correct - N}{N}$$

Where error denotes the rate of misclassification, correct denotes the number of training instances correctly predicted by the model, and N denotes the total number of training instances.⁷³

A stage value is calculated for the trained model which provides a weighting for any predictions that the model makes. The stage value for a trained model is calculated as follows:

$$stage = \ln\left(\frac{1 - error}{error}\right)$$

Where stage is the stage value used to weight model predictions, $\ln()$ is the natural logarithm, and error is the model's misclassification error. The effect of the stage weight is that more accurate models carry more weight or contribute more to the overall prediction.

Weak models are added in a sequential fashion and trained using weighted training data. The process is repeated until a predetermined number of weak learners have been created (a user parameter) or no further improvement on the training dataset can be made. When you're finished, you'll have a pool of weak learners, each with a stage value.⁷⁴

⁷³ Jason Brownlee : op. cit., pp. 137

⁷⁴ Ibid., pp. 138

3. Model evaluation

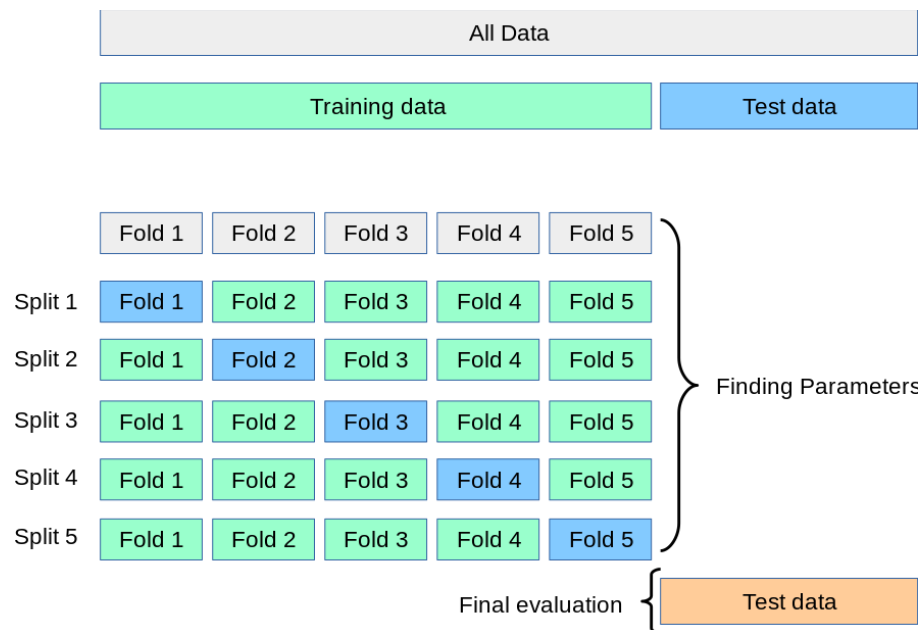
During machine learning and data mining modeling process, the evaluation metrics are crucial in achieving the best classifier.

3.1. Cross-validation

Cross-validation refers to a collection of techniques that separate data into training and test sets. The algorithm is fed the training set, along with the correct answers (in this case, prices), and it becomes the set used to make predictions. The algorithm is then asked to predict the outcome of each object in the test set. The algorithm's answers are compared to the correct answers, and an average score for how well the algorithm performed is computed.⁷⁵

Typically, the test set will be a small portion, maybe 5 to 10% of the data, this process is illustrated in Figure 16.

Figure 16: Cross Validation



Source: https://scikit-learn.org/stable/modules/cross_validation.html

⁷⁵ Toby Segaran. **Programming collective intelligence**, O'Reilly Media, USA, First. ed. 2007, pp. 176

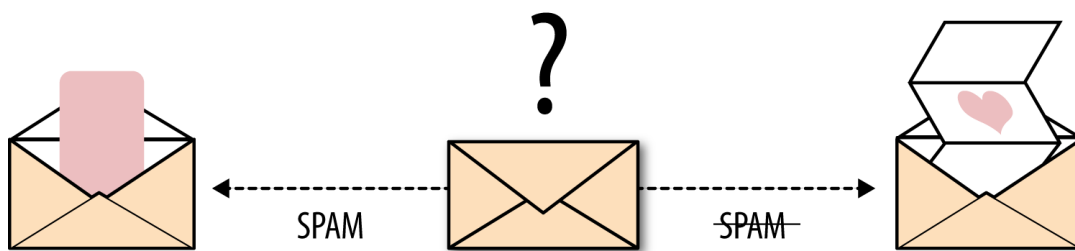
3.2. Hyperparameter optimization

Hyperparameters are parameters that the modeler specifies rather than the algorithm⁷⁶, their function is to define a hypothesis a priori, while other parameters specify it posteriori, after the algorithm interacts with the data and determines that certain parameter values work better in obtaining good predictions through an optimization process. Not all machine learning algorithms require hyper-parameter tuning but Some of the most advanced machine learning algorithms, however, still involve considerable hyper-parameter tuning. In the other hand, excessive hyper-parameter tinkering may cause the algorithm to detect false signals in the data.⁷⁷

3.3. Evaluation Metrics

Evaluation metrics are in the core of machine learning and predictive modeling, each ML task has different metrics, these tasks can vary from classification, regression or ranking in supervised learning, a good example for binary classification problems is spam detection⁷⁸, as shown in Figure 17.

Figure 17: Email spam detection



*Source: Alice Zheng, **Evaluating Machine Learning Models**, O'Reilly Media, USA, 1st ed, 2015, pp.08*

⁷⁶ Mohri, M., Rostamizadeh, A, Talwalkar, A : op. cit., pp. 04

⁷⁷ John Paul Mueller, Luca Massaron : op. cit., pp. 194-195

⁷⁸ Alice Zheng, **Evaluating Machine Learning Models**, O'Reilly Media, USA, 1st. ed, 2015, pp. 08

Since customer churn prediction is a classification problem, we will discuss the main classification metrics below:

3.3.1. Confusion Matrix

It is one of the most commonly used evaluation metrics in predictive modeling, owing to its simplicity and ability to be used to estimate other common metrics such as accuracy, precision, recall, and sensitivity.

For N output classification problems, NxN (N by N) matrix will be used, which means that in binary classification problems, the confusion matrix will be a 2 x 2 matrix, as shown in Figure 18 with two labels A and B.⁷⁹

Figure 18: Confusion Matrix

Actual	A	TP	FN
	B	FP	TN
		A	B
		Predicted	

Source: Hoss Belyadi, Alireza Haghighat : Machine Learning Guide for Oil and Gas Using Python, Gulf Professional Publishing, 2021, pp. 189

There are four possibilities for the instance to end up:

⁷⁹ Hoss Belyadi, Alireza Haghighat : **Machine Learning Guide for Oil and Gas Using Python**, Gulf Professional Publishing, 2021, pp. 189

- True Positives (TP): predicted positive, true value positive
- False Positives (FP): predicted positive, true value negative
- False Negatives (FN): predicted negative, true value positive
- True Negatives (TN): predicted negative, true value negative

3.3.2. Accuracy

Accuracy simply measures how much the classifier predicts correctly. It is the proportion of accurate predictions to overall predictions.⁸⁰

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Number of all predictions}}$$

3.3.3. Precision – Recall

Precision and recall are metrics used to evaluate a classification model.

Precision addresses the question, "How many of the items predicted to be relevant by the ranker/classifier are genuinely relevant?", whereas recall addresses the question, "How many of all the genuinely relevant items are identified by the ranker/classifier?"⁸¹

$$\text{Precision} = \frac{\text{True positive}}{\text{Predicted results}} = \frac{\text{True positive}}{\text{True positive} + \text{false positive}}$$

$$\text{Recall} = \frac{\text{True positive}}{\text{Actual results}} = \frac{\text{True positive}}{\text{True positive} + \text{false negative}}$$

3.3.4. F1 Score

The F1 Score, also known as the F score or F-measure, is the harmonic mean of recall and precision, it is a hybrid metric useful for unbalanced classes.

The harmonic mean, as opposed to the arithmetic mean, tends toward the lesser of the two elements as a result, the F1 score will be low if one of either precision or recall is low.⁸²

$$F1_{\text{Score}} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

⁸⁰ Alice Zheng, op. cit. pp. 08

⁸¹ Ibid. pp. 12

⁸² Ibid., pp. 12

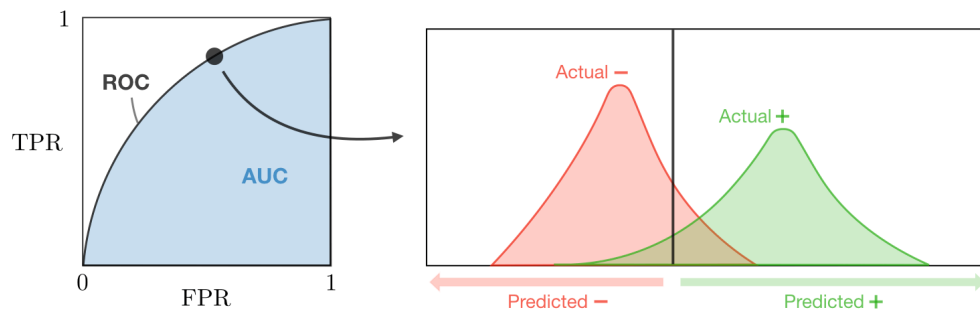
3.3.5. ROC curve

A receiver operating characteristics (ROC) graph is a technique for visually representing, sorting, and selecting classifiers based on their performance. An ROC graph illustrates the relative tradeoffs between benefits (true positives rate TPR) and costs (false positives rate FPR). The curve begins at 0.0, where there are no correct classifications but no false positives either, and ends at 1.1, where the model always predicts a positive classification result⁸³, as shown in Figure 19

3.3.6. AUC

The area under the receiving operating curve (AUC), also known as AUROC, is the area below the ROC, which is shown in Figure 19, which reduces the zone under the curve to a single number that is easier to interpret and compare⁸⁴.

Figure 19: ROC and AUC



Source: <https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-machine-learning-tips-and-tricks>

⁸³ Tom Fawcett: « An introduction to ROC analysis », **Pattern Recognition Letters**, Vol. 27, 2006, pp. 861-862

⁸⁴ Andrew P. Bradley : « The use of the area under the ROC curve in the evaluation of machine learning algorithms », **Pattern Recognition**, Vol. 30, Issue 7, 1997, pp. 1146

3.4. Model Interpretation

3.4.1. Definition of interpretability

In the context of ML systems, interpretability means providing explanation to humans, that is, explaining or presenting to a person in understandable terms.⁸⁵

Complex, difficult-to-interpret machine learning models such as deep neural networks, random forests, and gradient boosting machines are currently outperforming standard, and to some degree interpretable, linear/logistic regression models in many applications. However, there is often an obvious trade-off between model complexity and model interpretability, as a result, it is often difficult to understand why these complex models work so well.⁸⁶

There are numerous methods for interpreting ML algorithms, the most common ones are: Explanation Vectors, LIME (Local Interpretable Model-agnostic Explanations) and Shapley values. we will focus on the latter approach because it has a series of desirable theoretical properties

3.4.2. Shapley values

Shapley value, coined by Shapley (1953), is a model-agnostic method for explaining individual predictions with a solid theoretical foundation, it is a method originally invented for assigning payouts to players, this method builds on concepts from cooperative game theory.⁸⁷ The Shapley value necessitates a significant amount of computation time. Only the approximate solution is feasible in 99.9% of real-world problems. A precise calculation of the Shapley value is practically infeasible.⁸⁸

⁸⁵ Hugo Jair Escalante, Sergio Escalera, Isabelle Guyon, Xavier Baro, Yagmur Gucluturk, Umut Guclu, and Marcel van Gerven : **Explainable and Interpretable Models in Computer Vision and Machine Learning**, 1st. ed. Springer Publishing Company, 2018, pp. 05

⁸⁶ Kjersti Aas, Martin Jullum, Anders Løland: Explaining individual predictions when features are dependent: More accurate approximations to Shapley values, **Artificial Intelligence**, Vol. 298, 2021, pp. 01

⁸⁷ Ibid. pp. 02

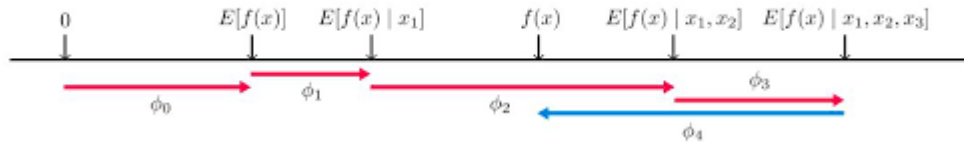
⁸⁸ Christoph Molnar : **Interpretable Machine Learning**, lulu.com, 2019

3.4.3. SHAP Value

SHAP (SHapley Additive exPlanations) is a tool developed by Lundberg and Lee (2016) to interpret individual predictions. SHAP is based on the ideal game-theoretic Shapley Values.⁸⁹

The goal of SHAP is to explain the prediction of an instance x by computing the contribution of each feature to the prediction, as shown in Figure 20, where ϕ is the shift amount that each variable contributes to the final prediction $f(x)$.

Figure 20: SHAP (SHapley Additive exPlanation)



Source: Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in Neural Information Processing Systems*. 2017, pp. 05

SAHP values attribute to each feature the shift in expected model prediction when conditioning on that feature. They demonstrate how to get from the base value $E[f(z)]$ that would be expected if we didn't know any features to the current output $f(x)$.

⁸⁹ Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in Neural Information Processing Systems*. 2017, pp. 04-05

Conclusion

Over the last two decades, **Machine Learning** has gained importance as a field that benefits various commercial and academic domains using many **supervised** and **unsupervised learning** algorithms. These algorithms can be assessed and tested using a variety of model evaluation metrics in order to choose the optimal machine learning model for a particular problem.

Through this chapter, we presented the most commonly used methods to treat our problem, which we discovered after conducting extensive documentary research on the subject. We have devoted it to the theoretical understanding of the methods we will use in our practical case.

In the following chapter, nine **Machine Learning** techniques will be applied on the telecommunication firm dataset to determine the optimal algorithm for forecasting future customers who may switch service providers.

CHAPTER III: ANALYSIS & RESULTS

CHAPTER III

ANALYSIS & RESULTS

Introduction

This chapter will summarize the findings of research projects dealing with the Churn problem to identify the most commonly used methods for developing our case study.

We will then provide the **Supervised** and **Unsupervised** learning methods we have chosen to carry out this research, specifically the **Ensemble** Learning models.

Lastly, we will discuss the model evaluation and validation techniques.

1. Exploratory Data Analysis

1.1. Dataset overview

1.1.1. Dataset description

The business dataset contains 66,469 customers information with 66 features. The names of the features are presented in Figure 21.

Figure 21: Feature names

```
Feature names:
Index(['year', 'month', 'user_account_id', 'user_lifetime', 'user_intake',
      'user_no_outgoing_activity_in_days', 'user_account_balance_last',
      'user_spendings', 'user_has_outgoing_calls', 'user_has_outgoing_sms',
      'user_use_gprs', 'user_does_reload', 'reloads_inactive_days',
      'reloads_count', 'reloads_sum', 'calls_outgoing_count',
      'calls_outgoing_spendings', 'calls_outgoing_duration',
      'calls_outgoing_spendings_max', 'calls_outgoing_duration_max',
      'calls_outgoing_inactive_days', 'calls_outgoing_to_onnet_count',
      'calls_outgoing_to_onnet_spendings', 'calls_outgoing_to_onnet_duration',
      'calls_outgoing_to_onnet_inactive_days',
      'calls_outgoing_to_offnet_count', 'calls_outgoing_to_offnet_spendings',
      'calls_outgoing_to_offnet_duration',
      'calls_outgoing_to_offnet_inactive_days',
      'calls_outgoing_to_abroad_count', 'calls_outgoing_to_abroad_spendings',
      'calls_outgoing_to_abroad_duration',
      'calls_outgoing_to_abroad_inactive_days', 'sms_outgoing_count',
      'sms_outgoing_spendings', 'sms_outgoing_spendings_max',
      'sms_outgoing_inactive_days', 'sms_outgoing_to_onnet_count',
      'sms_outgoing_to_onnet_spendings',
      'sms_outgoing_to_onnet_inactive_days', 'sms_outgoing_to_offnet_count',
      'sms_outgoing_to_offnet_spendings',
      'sms_outgoing_to_offnet_inactive_days', 'sms_outgoing_to_abroad_count',
      'sms_outgoing_to_abroad_spendings',
      'sms_outgoing_to_abroad_inactive_days', 'sms_incoming_count',
      'sms_incoming_spendings', 'sms_incoming_from_abroad_count',
      'sms_incoming_from_abroad_spendings', 'gprs_session_count',
      'gprs_usage', 'gprs_spendings', 'gprs_inactive_days',
      'last_100_reloads_count', 'last_100_reloads_sum',
      'last_100_calls_outgoing_duration',
      'last_100_calls_outgoing_to_onnet_duration',
      'last_100_calls_outgoing_to_offnet_duration',
      'last_100_calls_outgoing_to_abroad_duration',
      'last_100_sms_outgoing_count', 'last_100_sms_outgoing_to_onnet_count',
      'last_100_sms_outgoing_to_offnet_count',
      'last_100_sms_outgoing_to_abroad_count', 'last_100_gprs_usage',
      'churn'],
```

Source: Prepared by the student using Pandas package in python

A detailed description of the features is given in Appendix 03.

1.1.2. Source:

The dataset we used in this study is a public dataset about customers of an anonymous telecommunications company; the primary goal of this study is to perform analysis and prediction on this dataset to determine whether a customer will migrate to another company or not in order to increase profitability for the company providing the dataset.

This dataset was collected from Kaggle, an open-source platform that hosts datasets and data science projects.⁹⁰

1.2. Checking for missing values

There are no missing values in the dataset, as appears in Figure 22.

Figure 22: Missing values check

```
-----  
Null values check  
-----  
There is 0 null values in the Data Frame
```

Source: Prepared by the student using Pandas package in python

1.3. Correlation inspection

From the feature names explained in the previous section, we can expect a high correlation between many variables, and because of the large number of features in the dataset, a correlation filter is necessary to reduce the number of features.

➤ Correlation matrix:

The Figure 23. represents the correlations between all the features of the data frame.

⁹⁰ Kaggle, <https://www.kaggle.com/dimitaryanev/mobilechurndataxlsx> , [viewd: 27 August 2021]

Figure 23: Correlation Matrix of all features



Source: Prepared by the student using Matplotlib and Pandas packages in python,

we can see a high correlation between many variables. This means that all those variables explain the same thing. That is why removing them is the best option to reduce the complexity of the models. A correlation filter is applied in Table 3.

There are 17 features with a correlation higher than 0.85 with other features. These features will increase the complexity of the models without adding much value to the predictive power of models.

All of the features in the '**Feature Name 1**' column will be dropped to reduce the complexity of future models.

Table 3: Correlation of filtered features

Feature Name 1	Feature Name 2	Correlation
calls_outgoing_to_abroad_inactive_days	calls_outgoing_to_offnet_inactive_days	1
calls_outgoing_to_abroad_inactive_days	calls_outgoing_to_onnet_inactive_days	1
sms_outgoing_to_offnet_inactive_days	sms_outgoing_inactive_days	1
sms_outgoing_to_offnet_inactive_days	sms_outgoing_to_onnet_inactive_days	1
calls_outgoing_to_onnet_inactive_days	calls_outgoing_inactive_days	1
sms_outgoing_to_onnet_inactive_days	sms_outgoing_inactive_days	1
sms_outgoing_to_abroad_inactive_days	sms_outgoing_to_onnet_inactive_days	1
sms_outgoing_to_abroad_inactive_days	sms_outgoing_to_offnet_inactive_days	1
sms_outgoing_to_abroad_inactive_days	sms_outgoing_inactive_days	1
calls_outgoing_to_offnet_inactive_days	calls_outgoing_to_onnet_inactive_days	1
calls_outgoing_to_abroad_inactive_days	calls_outgoing_inactive_days	1
calls_outgoing_to_offnet_inactive_days	calls_outgoing_inactive_days	1
sms_outgoing_to_offnet_spendings	sms_outgoing_to_offnet_count	0.991952471
last_100_calls_outgoing_to_offnet_duration	last_100_calls_outgoing_duration	0.917000884
calls_outgoing_to_offnet_duration	calls_outgoing_duration	0.899741849
sms_outgoing_to_onnet_spendings	sms_outgoing_to_onnet_count	0.894003649
gprs_usage	gprs_session_count	0.893399068
calls_outgoing_to_offnet_duration	calls_outgoing_to_offnet_spendings	0.890645874
last_100_sms_outgoing_to_offnet_count	last_100_sms_outgoing_count	0.887187198
last_100_calls_outgoing_duration	calls_outgoing_duration	0.882746552
sms_outgoing_to_offnet_count	sms_outgoing_count	0.881271847
calls_outgoing_to_onnet_duration	calls_outgoing_to_onnet_spendings	0.879409798
sms_outgoing_to_offnet_spendings	sms_outgoing_count	0.873579731
last_100_sms_outgoing_count	sms_outgoing_count	0.873076028
last_100_calls_outgoing_to_offnet_duration	calls_outgoing_to_offnet_duration	0.872431111
last_100_sms_outgoing_to_onnet_count	sms_outgoing_to_onnet_count	0.870395979

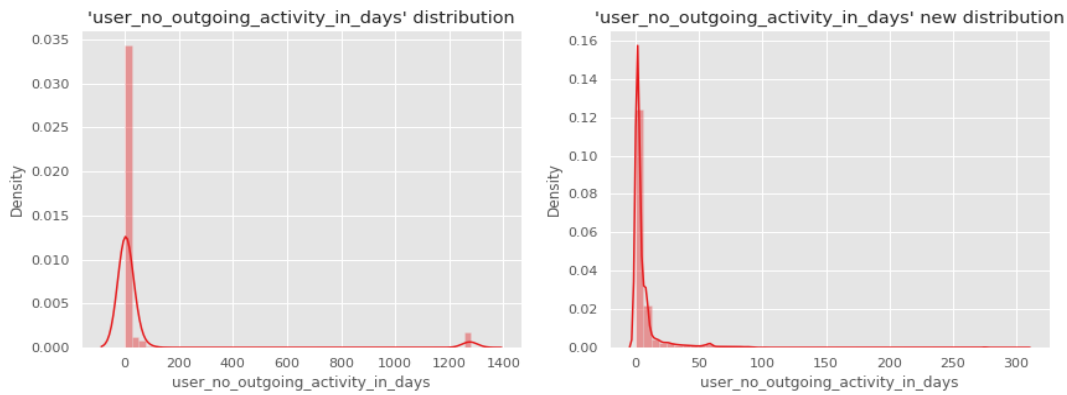
Source: Prepared by the student using Pandas package and formatted with Excel

1.4. Univariate analysis:

Due to a large number of features (47), the visual exploratory data analysis will focus on some of the most prominent features.

➤ **user_no_outgoing_activity_in_days:**

Figure 24: Distribution of non-activity of users in days



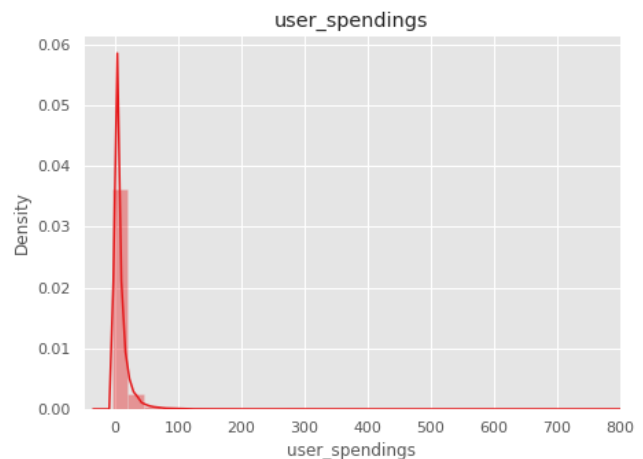
Source: Prepared by the student using Seaborn and Matplotlib packages

- The distribution of `user_no_outgoing_activity_in_days` is multimodal (at the left of Figure 24), which means some clients have no activity for more than three years (1200 days).
- Customers who do not show any activity whatsoever for more than one year are considered non-active customers and, therefore, are excluded from the study.
- The new distribution of the feature (at the right of Figure 24) after dropping 3,004 clients with more than one year of inactivity.

➤ **user_spending:**

Figure 25: Distribution of user_spending

Figure 25 shows that most of the users spend from 0 to 100. The distribution is highly skewed because some users spend until 800.

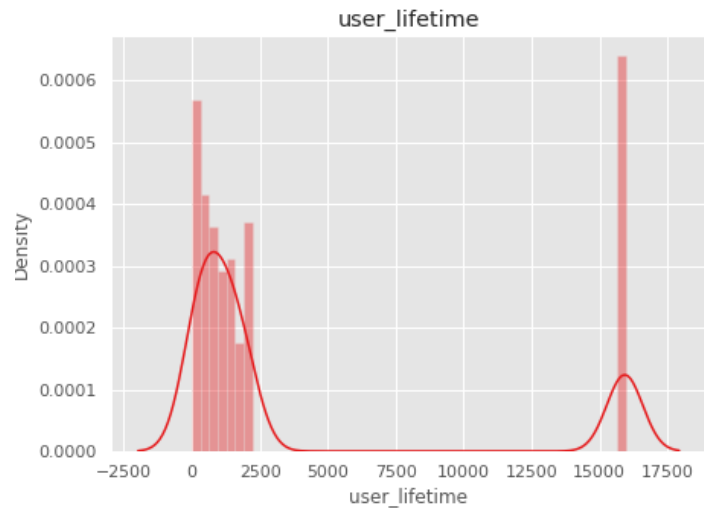


Source: Prepared by the student using Seaborn and Matplotlib packages

➤ **user_lifetime:**

Figure 26: Distribution of user_lifetime

Figure 26 shows that most customers are between 0 and 4,000, but some customers with more than 15,000, which suggests that those customers are very loyal to the company and should be kept in the study.

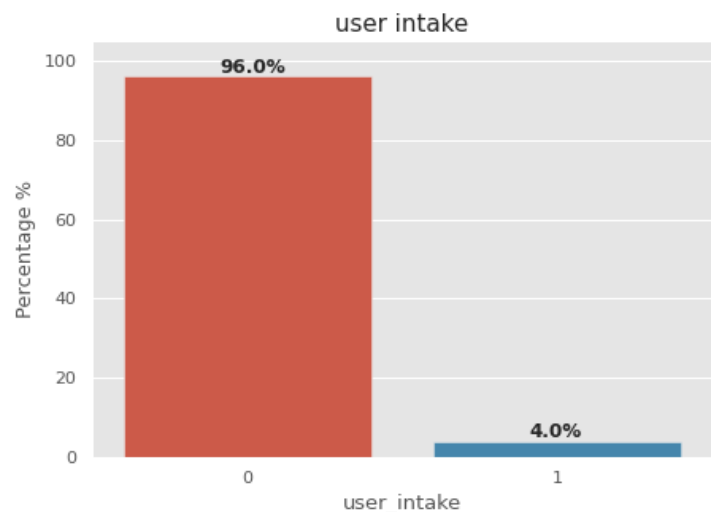


Source: Prepared by the student using Seaborn and Matplotlib packages

➤ **user_intake:**

Figure 27: Distribution of user_intake

Figure 27 shows that only 4% of customers are new customers, which is understandable for large telecom companies.



Source: Prepared by the student using Seaborn and Matplotlib packages

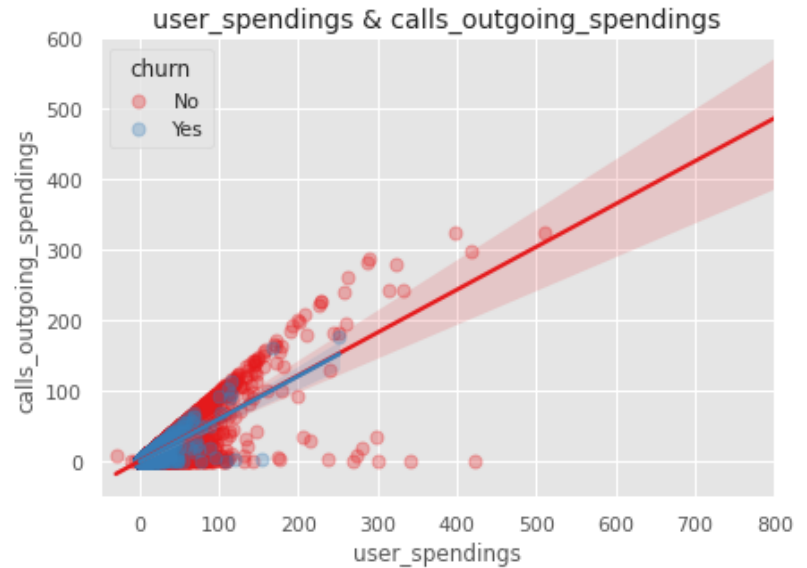
1.5. Multivariate analysis:

➤ **user_spending & calls_outgoing_spending:**

Figure 28: Customers spending habits

Figure 28 shows a positive correlation between user_spending and calls_outgoing_spending, suggesting that outgoing calls are one of the primary services driving the customers' spending habits.

The more the user spends, the less probable that he will churn and vice-versa.



Source: Prepared by the student using Seaborn and Matplotlib packages

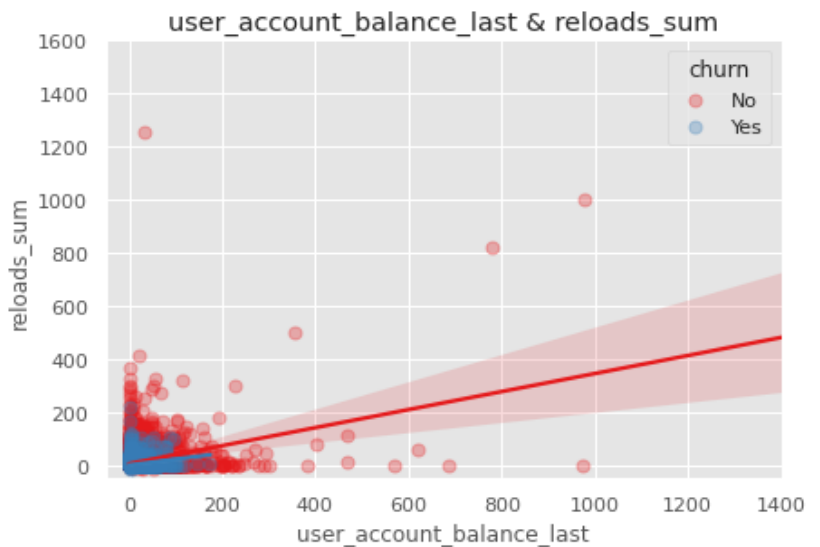
➤ **user_account_balance_last & reloads_sum:**

Figure 29: Spending frequency

Figure 29 shows that if the user has a relatively high balance in his account, he is less likely to churn.

So, the lesser the balance account and reloads sum, the more likely the user will churn.

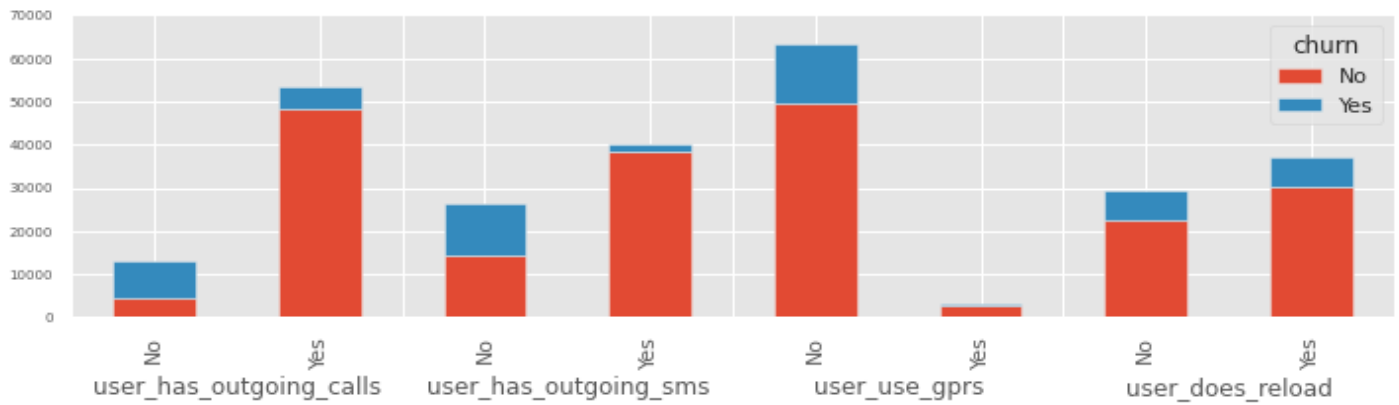
Some clients have 0 reload sum, but they have a good last balance, which means they are not spending much.



Source: Prepared by the student using Seaborn and Matplotlib packages

➤ **Churn by service:**

Figure 30: Churn by different services



Source: Prepared by the student using Seaborn and Matplotlib packages

From Figure 30, we can note that:

- If the user has NO outgoing calls, he is more likely to churn, which means that 'user_has_outgoing_calls' is suitable for separating churners from no-churners.
- If the user has outgoing SMS, he is not likely to churn.
- If the user uses GPRS, he is not likely to churn.
- The 'user_does_reload' variable does not separate well from the target variable 'churn'.

2. RFM Segmentation

The goal of the RFM analysis in our study is to determine and keep the clients that are most profitable to the company and exclude the occasional clients that barely add value to the company from the study because the cost of retention of these customers may be greater than the profit generated in the future.

2.1. Creating the RFM table

- **reloads_inactive_days** will represent Recency.
- **reloads_count** will represent Frequency.
- **reloads_sum** will represent Monetary.

Table 4 shows the first 5 observation of the new RFM table

Table 4: Sample from RFM table

Recency	Frequency	Monetary
54	0	0
3	1	12
7	6	49.53125
275	1	24.015625
15	2	8

Source: Prepared by the student using Pandas package

2.2. Standardizing the RFM table

Standardizing is applied on the RFM table so that the difference of measurement units does not affect the clustering algorithm. A sample of 5 customers is presented in Table 5.

Table 5: Scaled RFM table sample

Recency	Frequency	Monetary
-0.55197	-0.69554	-0.44180
-0.64357	0.04842	0.23866
-0.63639	3.76821	2.36689
-0.15503	0.04842	0.92001
-0.62202	0.79238	0.01184

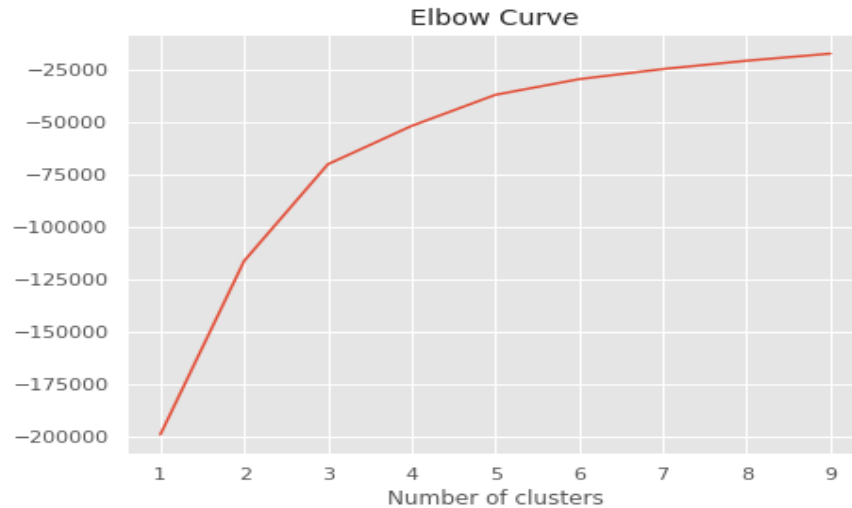
Source: Prepared by the student using Pandas package

2.3. Unsupervised learning with K-means

2.3.1. Finding the ideal number of clusters (Elbow method)

From Figure 31, we note that $k = 3$ is the ideal number of clusters.

Figure 31: Elbow method

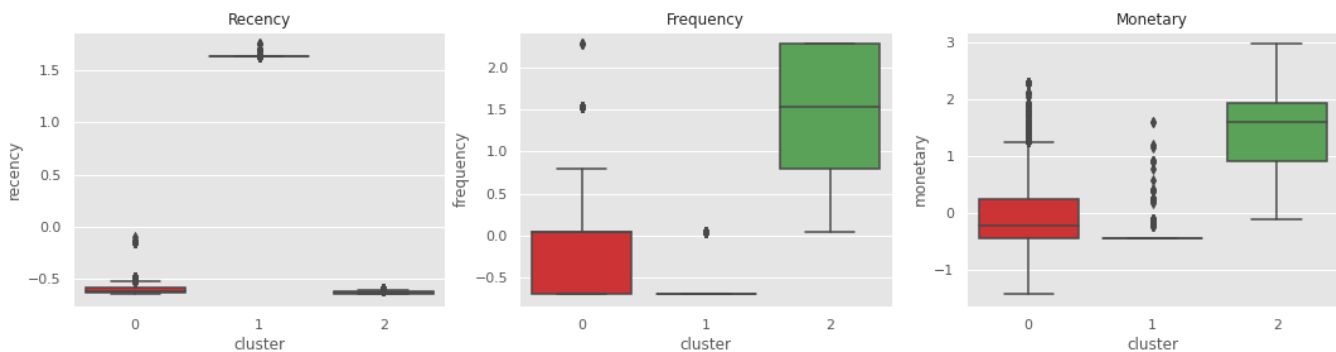


Source: Prepared by the student using Matplotlib package

2.3.2. Fitting K-means and interpreting the results

After fitting the RFM table by the K-means algorithm with $k = 3$, we would like to interpret the results and choose the clusters to keep and the ones to drop. The results are given in Figure 32.

Figure 32: Recency, Frequency & Monetary Box-plots



Source: Prepared by the student using Seaborn and Matplotlib packages

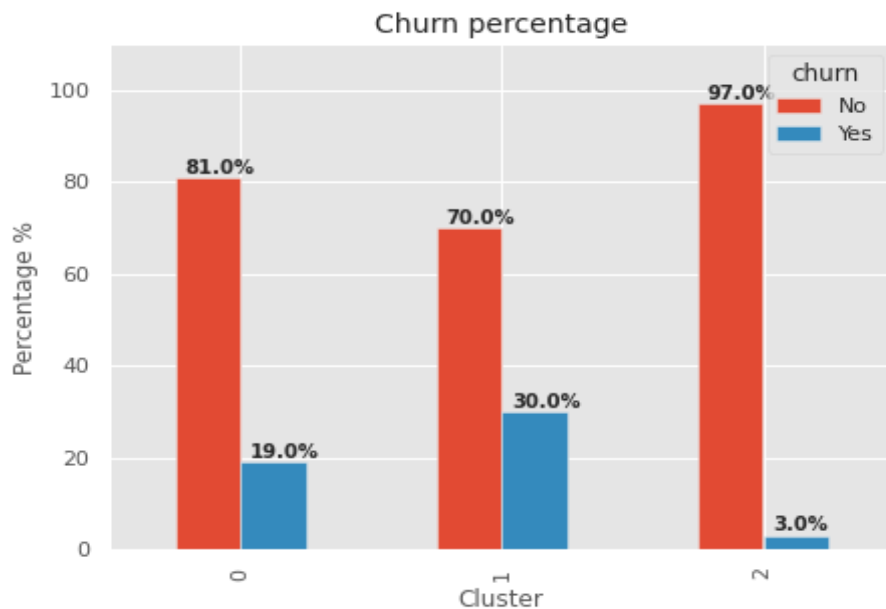
- In the Recency Box-Plot, cluster 1 has the highest value, which means those clients did not reload for a long time compared to the other two clusters.
- In the Frequency Box-Plot, cluster 1 has the lowest frequency value.
- In the Monetary Box-Plot, cluster 2 has the highest value between the three, which means that this group is the most profitable between the three groups, cluster 0 comes next, and cluster 1 is the last.

2.3.3. Cluster comparison by churn:

After analyzing Figure 33 below, we conclude that:

- Churn happens more often in clusters 1, then 0 and 2, respectively. These results match the previous results.
- Cluster 2 is the most profitable, and it includes the most loyal clients.
- Cluster 1 is the least profitable, and it contains the occasional clients.

Figure 33: Comparing churn percentage by cluster



Source: Prepared by the student using Seaborn and Matplotlib packages

Thus, we eliminated 17,949 customers belonging to cluster 1 because they are not very profitable, and the expected retention costs may exceed the expected profit.

3. Preprocessing Data

In this section, we will apply various data preprocessing techniques to prepare data for the modelling stage.

3.1. Standardization

Standardization is a critical approach usually conducted as a pre-processing step before many Machine Learning models to standardize the range of features in the input data set.

Here, we are standardizing the independent features by removing the mean and scaling to unit variance. A sample of 5 rows and 5 columns is presented in Table 6.

Table 6: Standardized features sample

gprs_inactive_days	calls_outgoing_to_onnet_count	user_has_outgoing_calls	month	last_100_reloads_sum
0.239888386	-0.137279615	0.516855517	-0.400824715	0.029442737
0.239888386	-0.137279615	-1.934776677	-0.400824715	-0.610330928
0.437388907	-0.137279615	0.516855517	2.539033617	-0.150493606
0.239888386	-0.137279615	0.516855517	-0.400824715	0.34932957
0.239888386	-0.137279615	0.516855517	-0.400824715	-0.610330928

Source: Prepared by the student using Pandas package

3.2. Train-Test Splitting

Here, we are splitting the dataset into training and testing data. We used 80% of the data for training and 20% for testing. The resulting data is presented in Table 7.

Table 7: Split data sizes

Set	Size
X_train	49,296
X_test	12,325
y_train	49,296
y_test	12,325

Source: Prepared by the student using Pandas package

3.3. Class imbalance

Imbalanced data refers to classification problems where the classes are not represented equally in the training set.

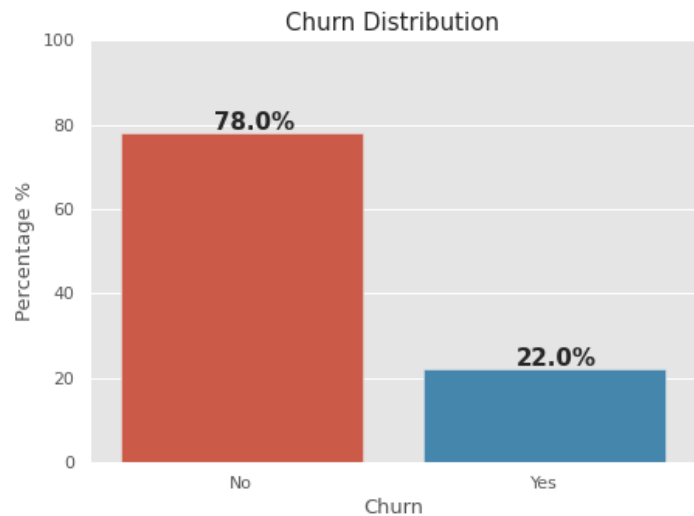
3.3.1. Checking for class imbalances:

The class imbalance in our training set is presented in Figure 34.

Figure 34: Class imbalance in the training set

We note that there is a severe class imbalance in the training set between churners and no churners. This will affect the ability of the models to learn from the minority class (churners).

Resampling techniques have to be applied to remove the class imbalance.

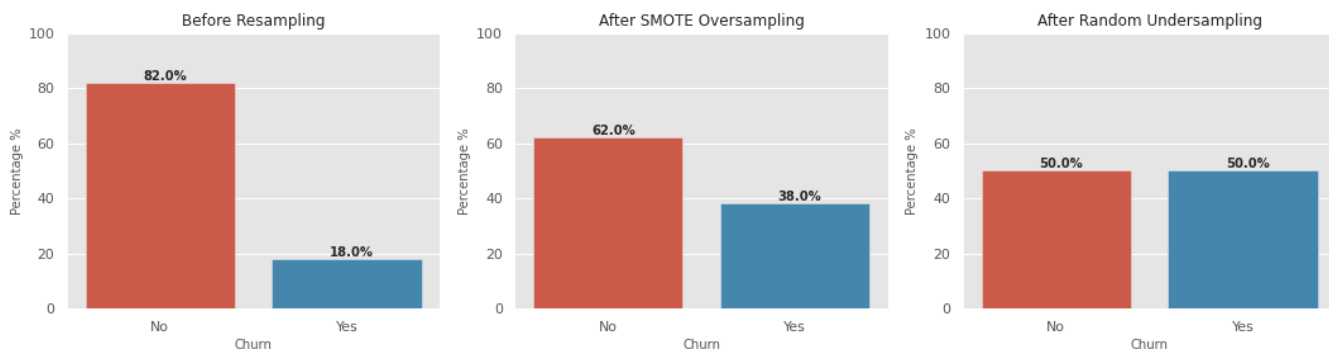


Source: Produced by the student using Seaborn and Matplotlib packages

3.3.2. Resampling:

To address the class imbalance problem, two resampling techniques will be used: SMOTE and Random Undersampling. The results of using these techniques are depicted in Figure 35.

Figure 35: Phases of resampling



Source: Prepared by the student using Sklearn, Seaborn and Matplotlib packages

The final result appears to be perfectly balanced with 50% in each class.

3.4. Feature selection

Due to a large number of features 47, we will apply a feature selection method to reduce the number of features and reduce the models' complexity.

We are using the REFCV method that uses five cross-validation folds to determine the optimal number of features and their names. The results are given in Figure 36.

Figure 36: Feature selection's best features

```
Optimal number of features : 37
Best features : Index(['calls_outgoing_count', 'calls_outgoing_duration',
'calls_outgoing_duration_max', 'calls_outgoing_inactive_days',
'calls_outgoing_spendings', 'calls_outgoing_spendings_max',
'calls_outgoing_to_abroad_count', 'calls_outgoing_to_abroad_duration',
'calls_outgoing_to_abroad_spendings', 'calls_outgoing_to_offnet_count',
'calls_outgoing_to_offnet_spendings', 'gprs_inactive_days',
'last_100_calls_outgoing_to_abroad_duration',
'last_100_calls_outgoing_to_onnet_duration', 'last_100_gprs_usage',
'last_100_reloads_count', 'last_100_reloads_sum',
'last_100_sms_outgoing_to_abroad_count', 'month', 'reloads_count',
'reloads_inactive_days', 'reloads_sum', 'sms_incoming_count',
'sms_incoming_from_abroad_count', 'sms_outgoing_count',
'sms_outgoing_inactive_days', 'sms_outgoing_spendings',
'sms_outgoing_spendings_max', 'sms_outgoing_to_abroad_count',
'sms_outgoing_to_abroad_spendings', 'sms_outgoing_to_onnet_count',
'user_account_balance_last', 'user_has_outgoing_sms', 'user_intake',
'user_lifetime', 'user_no_outgoing_activity_in_days', 'user_spendings'],
dtype='object')
CPU times: user 2min 26s, sys: 568 ms, total: 2min 27s
Wall time: 9min 57s
```

Source: Prepared by the student using Sklearn

According to the RFECV method, the optimal number of features for our dataset is 37. Thus, we will only be keeping these features.

Now, we are ready to begin modelling.

4. Modelling

After all the preprocessing tasks are done, we move to the modelling phase, where we will be trying several algorithms and choose the best performing one for our problem.

4.1. Model selection

4.1.1. with cross-validation



We will fit nine ML algorithms on 10-fold cross-validation to our training data, and we will compare these algorithms' results on the F1_score. The results of the algorithms are given in Table 8, where cells in  represent the highest value in the column, and the cells in  represent the lowest value in the same column. The code for generating this table is available in Appendix 02.

Table 8: ML algorithms comparison

Name	Accuracy	Recall	Precision	F1_score	Fit Time	Recall_STD
XGBClassifier	0.921545139	0.916648253	0.926017856	0.921301316	105.9989553	0.003033002
LGBMClassifier	0.920017361	0.913783379	0.925630685	0.919660941	236.4581318	0.003407604
RandomForestClassifier	0.912256944	0.895884475	0.92655831	0.910953325	7.872073913	0.004724015
ExtraTreesClassifier	0.911779514	0.894262908	0.927082249	0.910368738	4.382791567	0.004060684
GradientBoostingClassifier	0.907057292	0.894370763	0.918025917	0.90603998	12.28505998	0.002095277
BaggingClassifier	0.905269097	0.888522407	0.919679955	0.903827856	3.863122225	0.003542791
DecisionTreeClassifier	0.873993056	0.89499137	0.859343922	0.87679649	0.608851099	0.004530147
AdaBoostClassifier	0.873333333	0.843258756	0.897752706	0.869633136	2.965610552	0.005032679
LogisticRegressionCV	0.857395833	0.803071387	0.901535424	0.849456436	18.0352833	0.003185389

Source: Prepared by the student using Pandas and Sklearn packages and formatted with Excel

NOTE: LogisticRegressionCV is out of the scope of this study, it was added just for the sake of comparison with ensemble models.

From the results in Table 8, we note:

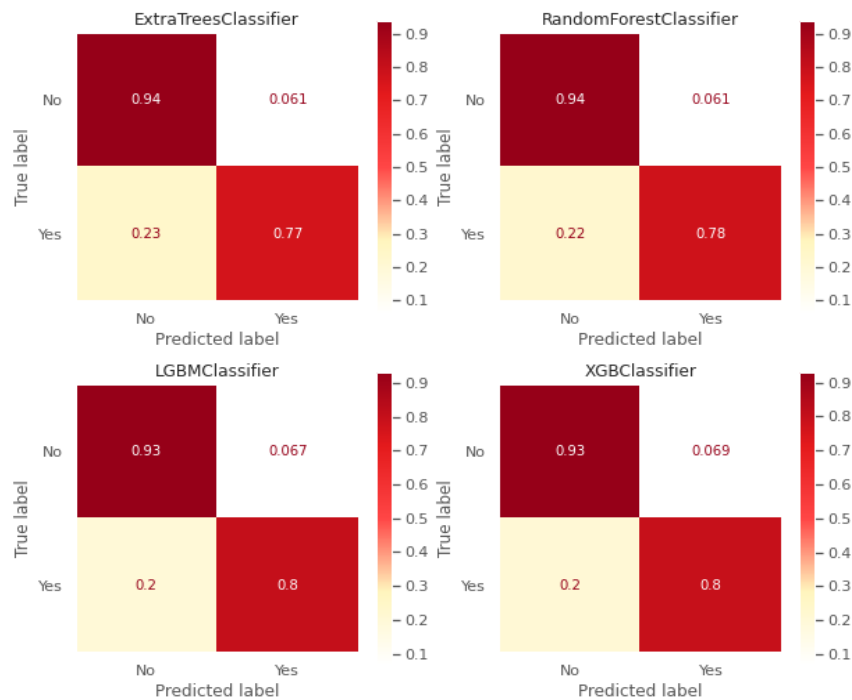
- Ensemble learning algorithms in the likes of RandomForestClassifier and XGBClassifier outperform simpler ones in the likes LogisticRegressionCV and DecisionTreeClassifier.

- LogisticRegressionCV has the lowest Accuracy, Recall and F1_score. Therefore it is the worst algorithm for our problem.
- LGBMClassifier and XGBClassifier took a way longer fitting time than the other algorithms, and this may be because they are boosting algorithms.
- XGBClassifier has the highest Accuracy, Recall, F1_score, and the second-lowest Recall_STD (recall standard deviation). Therefore it is the best model for our problem.

4.1.2. with test data

We will fit the top 4 performing algorithms with cross-validation on our test data to see the performance of every model on unseen data. The results are presented in confusion matrices of every model in Figure 37.

Figure 37: Confusion matrices of top-performing models



Source: Prepared by the student using Sklearn and Matplotlib packages

LGBMClassifier and XGBClassifier are outperforming RandomForestClassifier and ExtraTreesClassifier.

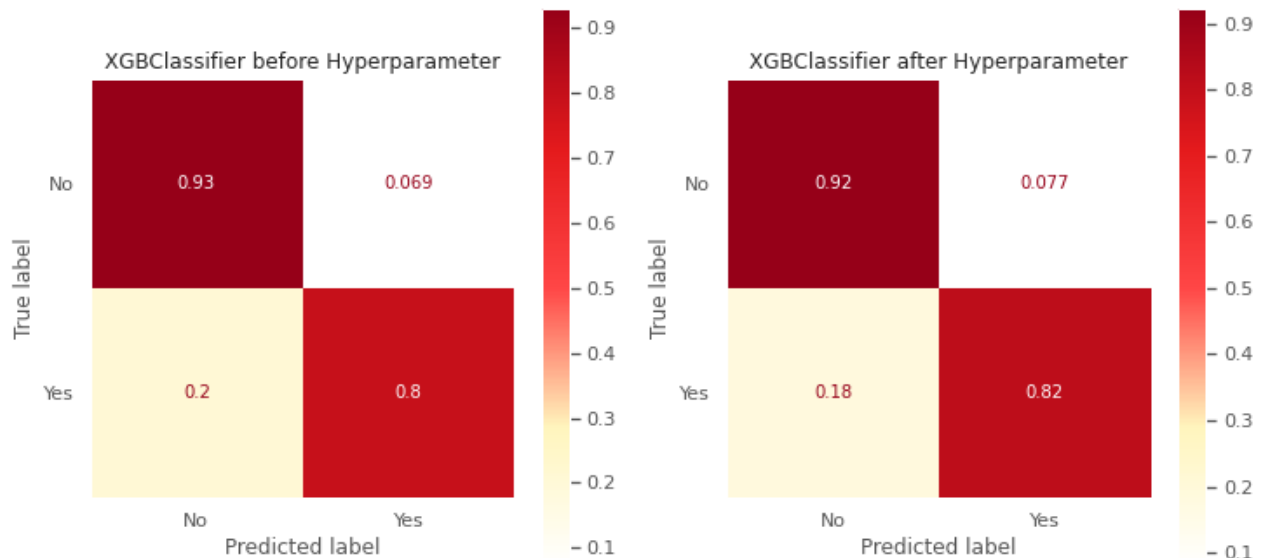
The results from Figure 37 match the ones from Table 8, the best algorithm that applies for the churn problem is XGBClassifier.

4.2. Hyperparameter optimization

Once our model has been chosen, we could optimize its parameters to meet the business requirement. In our case, we want the model to capture as many churners as possible. Therefore, we will try and optimize the model parameters using the maximization of the recall score as an objective.

We used the library Optuna for python to hyperparameter the XGBClassifier using the Bayesian Optimization algorithm and set the maximization of the recall score as objective. The results on the test set are presented in the form of a confusion matrix in Figure 38.

Figure 38: Hyperparameter results



. **Source:** Prepared by the student using Optuna and Matplotlib packages

The new model has improved in the Recall score of the positive class 'Yes', Although it has lost some Accuracy and Precision, which is understandable and does not affect much our study.

5. Model Evaluation and Interpretation

In this section, we will evaluate and interpret the model chosen before.

5.1. Classification report

The classification report summarises the performance of the model. The classification report of our model is presented in Figure 39.

Figure 39: Classification report

	precision	recall	f1-score	support
No	0.96	0.92	0.94	8000
Yes	0.69	0.82	0.75	1704
accuracy			0.90	9704
macro avg	0.83	0.87	0.85	9704
weighted avg	0.91	0.90	0.91	9704

Source: Prepared by the student using Sklearn package

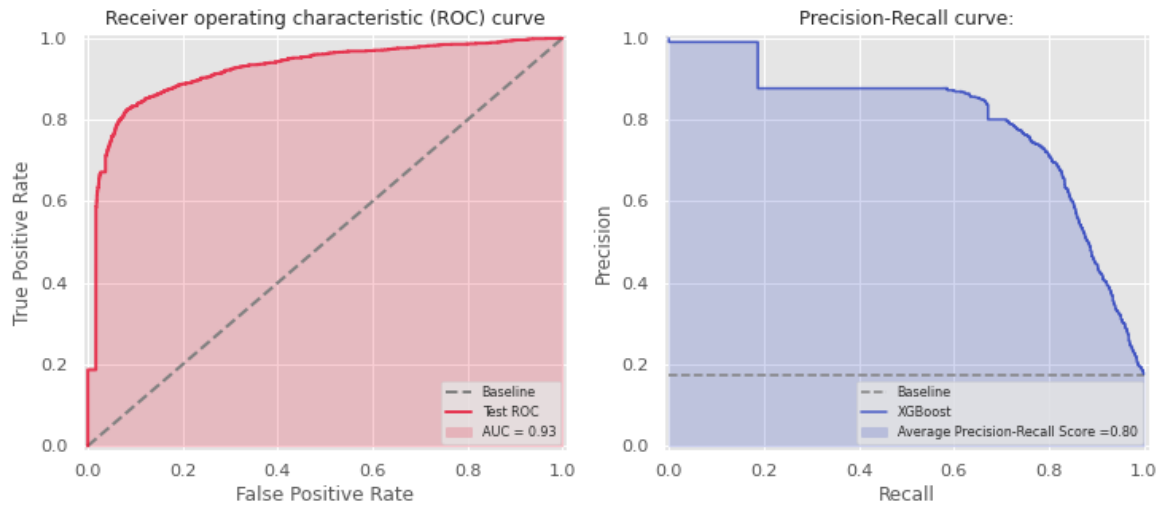
- The model performs well according to the classification report. The accuracy is 90% which is good.
- The recall of the 'Yes' class is 82% which means that the model has captured 82% of real churners, while the overall weighted recall is 90% which is good.
- The precision on the 'Yes' class is 69% which means that the model identified 31% of non-churners as churners, this doesn't affect much the business objective because the objective was to capture as much of the churners as possible.
- The overall weighted f1-score is 91% which is good.

The overall performance of the model looks satisfying.

5.2. ROC & Precision/Recall curves

Analyzing the ROC and Precision/Recall curves should give an overview of the performance of the model. The ROC and Precision/Recall curves of our chosen model are presented in Figure 40.

Figure 40: ROC & Precision/Recall curves



Source: Prepared by the student using Sklearn and Matplotlib packages

- The trade-off between sensitivity (or TPR) and specificity ($1 - \text{FPR}$) in the ROC curve of our model is good. In addition, the AUC is 0.93, which is close to 1, and that is very satisfying.
- However, the ROC curve is misleading when evaluating an imbalanced classification problem, we use the Precision/Recall curve to assess the model. It appears that our model's curve satisfies the requirement with an area under the curve of 0.80.

5.3. Model interpretation with SHAP values

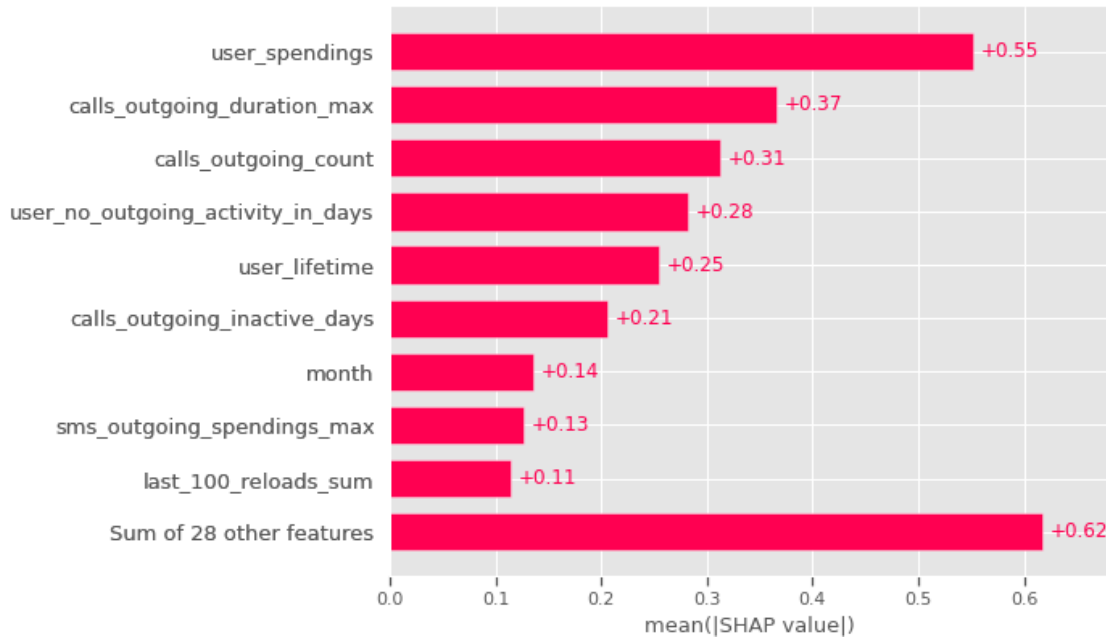
A sophisticated machine learning algorithm usually can produce accurate predictions. However, it is not easy to interpret the results. SHAP values allow the interpretation of the model on the prediction level and even globally on the global model level.

we use the SHAP library in python to produce the desired plots in order to interpret the results of the model

5.3.1. SHAP Feature importance:

The feature importance plot allows determining the most critical features in predicting churn. The feature importance plot of our model is presented in Figure 41.

Figure 41: SHAP Feature importance



Source: Prepared by the student using SHAP package

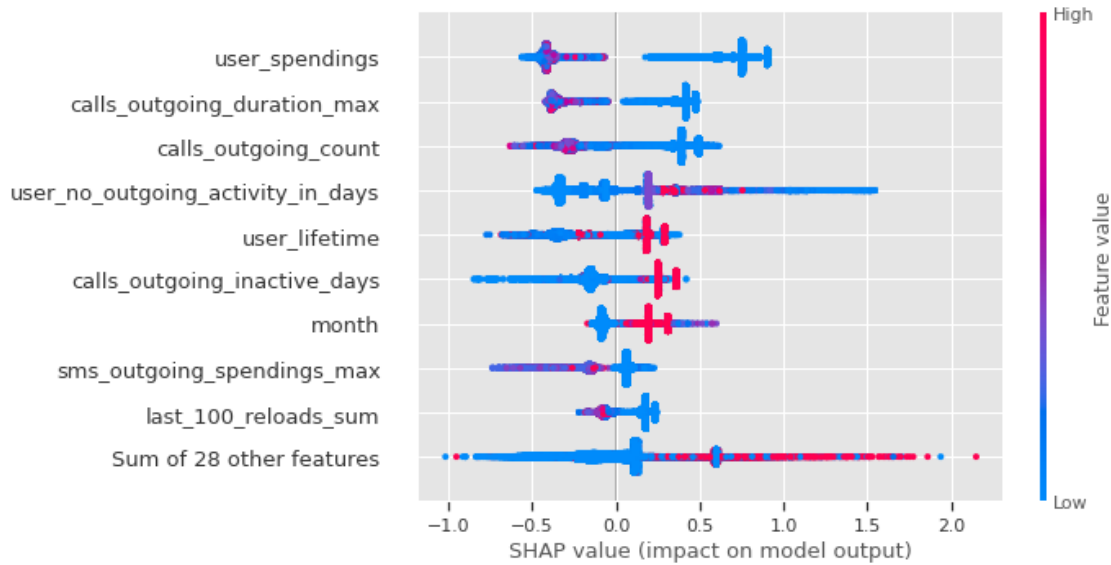
We can see that the `user_spendings` feature is the most important, followed by the `calls_outgoing_duration_max` and `user_no_outgoing_activity_in_days` features, respectively.

5.3.2. SHAP Summary plot:

The SHAP summary plot can further show the positive and negative relationships of the predictors with the target variable. It is constructed using all of the dots in the train data. It demonstrates the following:

- **Feature importance:** Variable importance is listed in descending order.
- **Feature value:** The colour indicates whether the variable for that observation is high (in red) or low (in blue).
- **Impact:** The horizontal placement indicates whether the value's effect is related to churn positively or negatively. The right is for the 'Yes' class, and the left is for the 'No' class.

Figure 42: SHAP Summary plot



Source: Prepared by the student using SHAP package

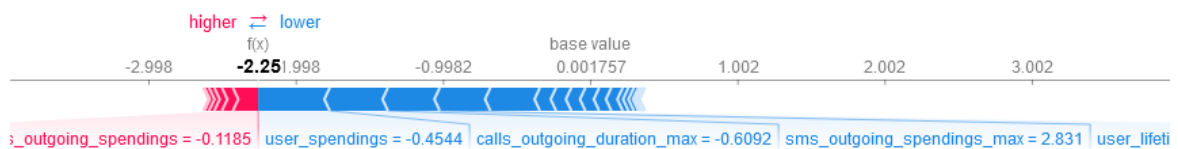
From the summary plot, we can note:

- **user_spending:** the lower the value, the more likely it is that the user will churn.
- **calls_outgoing_duration_max:** the lower the value, the more likely it is that the user will churn.
- **calls_outgoing_count:** the lower the value, the more likely for churn.
- **user_no_outgoing_activity_in_days:** the lower the value, the less likely for churn.
- **user_lifetime:** the higher the value, the more likely for churn and vice-versa.

5.3.3. Individual SHAP Value:

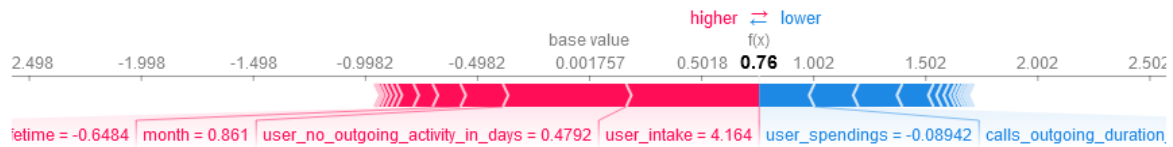
The SHAP values can also be done on individual cases like Figure 43 and Figure 44.

Figure 43: SHAP individual 1



Source: Prepared by the student using SHAP package

Figure 44: SHAP individual 2



Source: Prepared by the student using SHAP package

Unfortunately, since the data was standardized previously, we cannot interpret the results by their values, but we still can see the variables that influence the most on individual predictions prediction:

- **Individual 1:**

The prediction for this individual is 'NO', which means no_churner, the features that pushed the most to this prediction are: sms_outgoing_spending_max, calls_outgoing_duration_max, and user_spendings.

- **Individual 2:**

The prediction for this individual is 'YES', which means churner, the features that pushed the most to his churn are: user_intake, user_no_outgoing_avtivity_in_days, and month, while user_spendings and calls_outgoing_duration pushed for 'NO', but it was not enough.

We can apply this method to every prediction and explain the reasons for the outcome of the prediction.

Conclusion

In this chapter, we were able to build, first and foremost, our methodological approach, which enabled us to obtain significant results by relying on a thorough analysis using various tools and a thorough understanding of the industry.

The third chapter began by providing an overview of our dataset to identify our functional features and the target variable.

Following that, to learn more about our dataset, we used both **univariate** and **multivariate** analysis to extract critical insights based on visualizations and correlation matrices.

Afterwards, a **data preprocessing** stage was required before the modelling phase could begin, where we used various preprocessing techniques to prepare for the modelling stage.

Then, we went through the **modelling stage**, where we applied various classification models to determine which one performed the best in identifying churners.

Finally, we evaluated the **XGBoost** model that we have chosen and interpreted the results using the **SHAP** package.

In the General Conclusion, we will outline the results obtained from this study and confirm the mentioned hypothesis in the introduction.

GENERAL CONCLUSION

GENERAL CONCLUSION

Overview:

Predicting attrition among telecom operators requires extensive data analysis. Churn prediction is quickly becoming one of the most important sources of revenue for companies. Moreover, it helps to prevent the departure of customers who are about to discontinue their contract with the company, which opens the door to renegotiate the contract to maintain customer loyalty.

The first part of our work was devoted to a detailed study of CRM, the churn phenomenon, and the RFM analysis process. We have also been interested in Data mining and Machine Learning techniques used for prediction. We invested most of our time in assimilating these new concepts while keenly interested in the case studies previously conducted.

The second part of the study was dedicated to highlighting the data mining and machine learning methodology. We have studied the three main phases for building an accurate ML model: data preprocessing, modelling and model evaluation. We were also interested in studying the state-of-the-art algorithms in the CCP industry. We have also been interested in the state of the current literature on customer churn problems.

In the last part, we conducted the empirical study on an anonymous telecom dataset. We applied all the knowledge gained from the previous parts on this dataset. The objective of this study was to predict in advance as much as possible the users that are going to churn in the company's dataset with a quite good accuracy.

Throughout this classification problem, we aimed at dividing users into two groups of churners and non-churners. There are various data mining techniques that specialists have developed to create the predictive model for CCP problems, most notably ensemble learning methods, that have shown promising results for these cases.

The working dataset for this project was collected from a telecom company that each user indexed, and after going through the necessary data preprocessing and cleaning phases, the final dataset generated included 66,469 observations with up to 66 features.

We also used RFM analysis and the K-means clustering algorithm to partition our dataset to exclude clients who do not contribute significantly to the company's earnings from the study.

To achieve the best results possible, we split our dataset into two portions, with 80% of the dataset destined for training the model and the remaining for testing the model's reliability.

Analysis of Results:

The results were consistent with the current literature. It was discovered that ensemble learning algorithms such as Extreme Gradient Boosting (XGBoost) performed the best on the datasets used in this study and are also top performers in other studies. Furthermore, the performance of ML models compared to statistical models may be at least marginally better.

This study has led us to discover that XGBoost was performing better than other models, with an accuracy of 90%. Therefore, we decided to finalize our study with this model. We then applied hyperparameter optimization to this model to improve its recall as much as possible to meet our business needs.

In terms of identifying churners and non-churners, the resulting model's recall and precision values were both excellent. As a result, we could conclude that the XGBClassifier model is capable of handling both classes effectively.

Finally, the interpretation of the model using SHAP values revealed that the most relevant feature in predicting customer churn is user_spendings. Furthermore, we were able to examine and determine the influence of each feature on churn prediction both globally on the whole dataset and individually on two random customers.

lastly, we can confirm our first and second hypotheses as to the following:

- Data mining techniques are accurate in predicting client behaviour.
- As demonstrated in our practical case, ensemble techniques, such as XGBoost, can outperform other simpler classification models.

In addition, we reject the third hypothesis, which suggested that the most relevant feature in predicting customer churn is the user's last account balance.

Research Recommendations:

For the success of churn management at the company, we would like to suggest several recommendations that we think will have a dramatic impact on the business in the short and long term:

- Integrate churn prediction results into the customer relations and marketing departments' IT tools.
- Bring the marketing department closer to the CRM department and synchronize future actions (retention, promotion, . . .).
- The implementation of a strategy aiming at valorising projects based on datamining will allow the company to align itself with the other competitors.

Research Limitations:

The main restriction of our research was the inability to find a host company where we could conduct this research owing to the present pandemic; also, the problems of confidentiality in the different telecom operators in Algeria forced us to work with a dataset from a different source.

Future Research:

Future research could include other, more complex models to predict the churn. Some examples could include Weighted Random Forests, Artificial Neural Networks, or dimensionality reduction methods such as PCA and t-distributed Stochastic Neighbor Embedding (t-SNE).

Other ways the models could be improved would be using hybrid models that were achieving significant performance gains.

Finally, intensive hyperparameter and other optimizations on a big dataset are not possible on a home computer since they take numerous iterations to identify ideal parameters, which can significantly impact model performance. As a result, in future studies, either more powerful computers or cloud computing, for example, are suggested.

BIBLIOGRAPHY

BIBLIOGRAPHY

Books:

- Alexander H. Kracklauer, D. Quinn Mills, Dirk Seifert: « **Collaborative Customer Relationship Management: Taking CRM to the Next Level** », 1st. ed, Springer-Verlag Berlin Heidelberg, 2004
- Alice Zheng, « **Evaluating Machine Learning Models** », O'Reilly Media, USA, 1st ed, 2015
- Bhavani, Thuraisingham, : « **Data Mining: Technologies, Techniques, Tools and trends** », 1st. ed., CRC Press, 1998
- Christoph Molnar : « **Interpretable Machine Learning** », lulu.com, 2019.
- Deisenroth, Marc Peter, Faisal, A. Aldo, Ong, Cheng Soon: « **Mathematics for Machine Learning** », Cambridge University Press, 2020
- Francis Buttle, Stan Maklan : « **Customer Relationship Management: Concepts and Technologies** », 4th. ed, Routledge, 2019
- Gujarati, Damodar N., Porter, Dawn C.: « **Basic econometrics** », McGraw Hill, 5th. Ed., New York, 2009
- Han, Jiawei, Micheline Kamber, and Jian Pei: « **Data Mining: Concepts and Techniques** », 3rd. ed., Morgan Kaufmann Publishers, 2012,
- Hoss Belyadi, Alireza Haghighat : « **Machine Learning Guide for Oil and Gas Using Python** », Gulf Professional Publishing, USA, 2021
- Hugo Jair Escalante, Sergio Escalera, Isabelle Guyon, Xavier Baro, Yagmur Gucluturk, Umut Guclu, and Marcel van Gerven : « **Explainable and Interpretable Models in Computer Vision and Machine Learning** », 1st. ed., Springer Publishing Company, 2018

- James, G., Witten, D., Hastie, T., & Tibshirani, R.: « **An introduction to statistical learning** », 2nd. ed, Springer, 2021.
- Jason Brownlee : « **Master Machine Learning Algorithms Discover How They Work and Implement Them from Scratch** », Machine Learning Mastery, 2016
- John Paul Mueller, Luca Massaron : « **Machine Learning For Dummies** », New Jersey, John Wiley & Sons Inc, 2016
- Konstantinos Tsipis, Antonios Chorianopoulos: « **Data Mining Techniques in CRM: Inside Customer Segmentation** », 1st. ed, Wiley, 2010
- Mohri, M., Rostamizadeh, A, Talwalkar, A : « **Foundations of Machine Learning** », MIT Press, 2018
- Oliver Theobald : « **Machine learning for absolute beginners : a plain English introduction** », 2nd. ed., Scatterplot Press, 2017
- Pratap Dangeti: « **Statistics for Machine Learning** », Packt Publishing, 2017
- Rob Mattison: « **The Telco Churn Management Handbook** », XiT Press, Illinois, USA, 2005
- Toby Segaran. « **Programming collective intelligence** », O'Reilly Media, USA, 1st. ed., 2007

Articles & Journals:

- A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar: Customer churn prediction in telecommunication industry using data certainty, **Journal of Business Research**, Vol. 94, 2019.
- Andrew P. Bradley : « The use of the area under the ROC curve in the evaluation of machine learning algorithms,», **Pattern Recognition**, Vol. 30, Issue 7, 1997.
- Arthur Samuel, Some Studies in Machine Learning Using the Game of Checkers, **IBM Journal of Research and Development**, Vol. 3, Issue. 3, 1959.

- Burez, J. and Van den Poel, D. :Handling class imbalance in customer churn prediction, **Expert Systems with Applications.**, Vol. 36, , 2009 .
- García, S., Ramírez-Gallego, S., Luengo, J. et al. :Big data preprocessing: methods and prospects, **Big Data Analytics**, Vol. 1, Issue. 9, 2016.
- Guyon, I., Elisseeff, A., & Kaelbling, L. P. (Ed.) : An introduction to variable and feature selection. **Journal of Machine Learning Research**, Vol 3, Issue 7, 2003.
- J R Miglautsch: Thoughts on RFM scoring, **Journal of Database Marketing & Customer Strategy Management**, Vol. 08, Issue. 1, 2000
- J R Miglautsch: Thoughts on RFM scoring, **Journal of Database Marketing & Customer Strategy Management**, Vol. 08, Issue. 1, 2000
- Kjersti Aas, Martin Jullum, Anders Løland: Explaining individual predictions when features are dependent: More accurate approximations to Shapley values, **Artificial Intelligence**, Vol. 298, 2021.
- Kotsiantis, Sotiris. : Supervised Machine Learning: A Review of Classification Techniques. **Informatica**, Vol. 31, 2007
- Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions.", **Advances in Neural Information Processing Systems**. 2017.
- Msweli, Pumela.: «Modeling Switching Behaviour of Direct Selling Customers.», **South African Journal of Economic and Management Sciences**. Vol. 7, Issue. 2, 2004.
- Nanjundan, S., Sankaran, S., Arjun, C.R., & Anand, G. « Identifying the number of clusters for K-Means: A hypersphere density based approach ». **ArXiv**, 2019
- Nicolas Glady, Bart Baesens, Christophe Croux : « Modeling churn using customer lifetime value», **European Journal of Operational Research**, Vol. 197, Issue. 1, 2009 .
- Sahar F. Sabbeh : Machine-Learning Techniques for Customer Retention: A Comparative Study, **International Journal of Advanced Computer Science and Applications**, Vol. 9, Issue. 2, 2018

- Shaaban, E., Helmy, Y., Khder, A.E., & Nasr, M.M.: «A Proposed Churn Prediction Model», **International Journal of Engineering Research and Applications**, Vol. 02, Issue. 04, 2012.
- Sotiris Kotsiantis, Dimitris Kanellopoulos, P. E. Pintelas: Data Preprocessing for Supervised Learning. **International Journal of Computer Science**. Vol. 1, Issue. 1, 2006.
- Tom Fawcett: « An introduction to ROC analysis », **Pattern Recognition Letters**, Vol. 27, 2006.
- Torsten J Gerpott, Wolfgang Rams, Andreas Schindler: « Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market», **Telecommunications Policy**, Vol. 25, Issue. 4, 2001.

Thesis:

- Bochra CHEMAM: « *Utilisation d'une démarche de Datamining pour la prédiction du CHURN en téléphonie mobile post payé, Cas : ATM Mobilis* », Master's Thesis, ENSSEA, 2019.
- Oskar Sucki: « *Predicting the customer churn with machine learning methods - CASE: private insurance customer data* », Master's Thesis, Lappeenranta-Lahti University of Technology LUT, 2019.
- Yasser Houssein Eddine KEHAL: « *Data Mining for Customer Behavior Modeling in E-commerce, Case study: Fin-Tech Company (2012-2013)* », Master's Thesis, ENSSEA, 2020.

Conferences:

- Aggelis, V., & Christodoulakis, D.: Customer clustering using rfm analysis, **In Proceedings of the 9th WSEAS International Conference on Computers**, 2005

- Maryani and D. Riana, "Clustering and profiling of customers using RFM for customer relationship management recommendations," **5th International Conference on Cyber and IT Service Management**, 2017

Websites:

- Accenture, <https://newsroom.accenture.com/news/us-companies-losing-customers-as-consumers-demand-more-human-interaction-accenture-strategy-study-finds.htm>
- Achieveforum, <https://www.achievetheforum.com/resources-research>
- Algorithmia, <https://algorithmia.com/blog/machine-learning-use-cases>
- Forbes, <https://www.forbes.com/sites/jiawertz/2018/09/12/dont-spend-5-times-more-attracting-new-customers-nurture-the-existing-ones/?sh=316cb5035a8e>
- Harvard Business School, <https://hbswk.hbs.edu/archive/the-economics-of-e-loyalty>
- KPMG, <https://home.kpmg/xx/en/home/insights/2020/01/home.html>
- Sklearn, https://scikit-learn.org/stable/modules/cross_validation.html
- Stanford Education, <https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-machine-learning-tips-and-tricks>

APPENDENCES

APPENDENCES

Appendix 01: Code for importing used libraries and packages

```
#General
import pandas as pd
import numpy as np
import sklearn
#Preprocessing
from itertools import cycle
from sklearn import feature_selection
from sklearn import model_selection
from sklearn import metrics
from scipy import stats, interp
from sklearn.preprocessing import StandardScaler,
from sklearn.model_selection import train_test_split, KFold, from
sklearn.model_selection import cross_val_predict
from sklearn.feature_selection import RFECV
from imblearn.under_sampling import RandomUnderSampler
from imblearn.over_sampling import SMOTE
#Models
from sklearn import tree, linear_model, ensemble,
from xgboost import XGBClassifier
from lightgbm import LGBMClassifier
from sklearn.cluster import KMeans
#Model_Evaluation
import shap
import optuna
from sklearn.metrics import roc_curve, auc, make_scorer, accuracy_score
from sklearn.metrics import plot_confusion_matrix, classification_report
from sklearn.metrics import roc_auc_score, precision_recall_curve
from sklearn.metrics import recall_score, precision_score
from sklearn.metrics import f1_score, average_precision_score
#Visualization
import matplotlib.pyplot as plt
import matplotlib.style as style
from matplotlib.colors import LinearSegmentedColormap
import seaborn as sns
#Setting_Parameters
import warnings
warnings.filterwarnings('ignore')
sns.set()
style.use('ggplot')
pd.set_option("display.max_rows", 100)
%matplotlib inline
plt.rcParams["figure.figsize"] = (12, 8)
```

Appendix 02: Code used for model selection

```
#Machine Learning Algorithm (MLA) Selection and Initialization
MLA = [
    #Ensemble Methods
    ensemble.AdaBoostClassifier(),
    ensemble.BaggingClassifier(),
    ensemble.ExtraTreesClassifier(),
    ensemble.GradientBoostingClassifier(),
    ensemble.RandomForestClassifier(),
    #xgboost: http://xgboost.readthedocs.io/en/latest/model.html
    XGBClassifier(),

    #LightGBM
    LGBMClassifier(n_jobs=-1),

    #GLM
    linear_model.LogisticRegressionCV(solver='liblinear'),

    #Trees
    tree.DecisionTreeClassifier(),]

#split dataset in cross-validation with this splitter class
cv_split = model_selection.ShuffleSplit(n_splits = 10, test_size = .3,
                                       train_size = .6, random_state = 0 )

#create table to compare MLA metrics
MLA_columns = ['Name', 'Accuracy', 'Recall', 'Precision', 'F1_score', 'Fit Time']
MLA_compare = pd.DataFrame(columns = MLA_columns)

#index through MLA and save performance to table
row_index = 0
for alg in MLA:

    #set name and parameters
    MLA_name = alg.__class__.__name__
    MLA_compare.loc[row_index, 'Name'] = MLA_name
    #MLA_compare.loc[row_index, 'MLA Parameters'] = str(alg.get_params())

    #score model with cross validation:
    scoring = {'accuracy' : make_scorer(accuracy_score),
              'precision' : make_scorer(precision_score),
              'recall' : make_scorer(recall_score),
              'f1_score' : make_scorer(f1_score)}

    cv_results = model_selection.cross_validate(alg, X, y, cv = cv_split,
                                              scoring=scoring)
    MLA_compare.loc[row_index, 'Fit Time'] = cv_results['fit_time'].mean()
    MLA_compare.loc[row_index, 'Accuracy'] = cv_results['test_accuracy'].mean()
    #print(cv_results['test_accuracy'].mean())
    MLA_compare.loc[row_index, 'Recall'] = cv_results['test_recall'].mean()
    MLA_compare.loc[row_index, 'Recall_STD'] = cv_results['test_recall'].std()
    MLA_compare.loc[row_index, 'Precision'] = cv_results['test_precision'].mean()
    MLA_compare.loc[row_index, 'F1_score'] = cv_results['test_f1_score'].mean()
    print(cv_results['fit_time'].mean(), '==> ', MLA_name, 'is done!')
    row_index+=1

#print and sort table:
MLA_compare.sort_values(by = ['F1_score'], ascending = False, inplace = True)
```

Appendix 03: Features description

Feature	Description
year	Year
month	Month
user_account_id	Unique customer identifier
user_lifetime	Customer aging in months
user_intake	New customer identifier
user_no_outgoing_activity_in_days	Number of days when customer did not do any action
user_account_balance_last	Customer account balance at the end of the period
user_spending	Revenue spend in the period
user_has_outgoing_calls	Customer made at least 1 call
user_has_outgoing_sms	Customer made at least 1 sms
user_use_gprs	Customer used data at least once
user_does_reload	Customer has done at least 1 recharge
reloads_inactive_days	Number of days without recharge
reloads_count	Number of recharges
reloads_sum	Amount of recharges
calls_outgoing_count	Number of outgoing calls
calls_outgoing_spending	Amount spent on outgoing calls
calls_outgoing_duration	Duration of all outgoing calls
calls_outgoing_spending_max	The most expensive call per period
calls_outgoing_duration_max	The longest call per period
calls_outgoing_inactive_days	Number of days without outgoing calls
calls_outgoing_to_onnet_count	Number of calls to on-net
calls_outgoing_to_onnet_spending	Amount spent on outgoing calls to on-net
calls_outgoing_to_onnet_duration	Duration of all outgoing calls to on-net
calls_outgoing_to_onnet_inactive_days	Number of days without outgoing call to on-net
calls_outgoing_to_offnet_count	Number of calls to off-net
calls_outgoing_to_offnet_spending	Amount spent on outgoing calls to off-net
calls_outgoing_to_offnet_duration	Duration of all outgoing calls to off-net
calls_outgoing_to_offnet_inactive_days	Number of days without outgoing call to off-net
calls_outgoing_to_abroad_count	Number of calls to other countries
calls_outgoing_to_abroad_spending	Amount spent on outgoing calls to other countries
calls_outgoing_to_abroad_duration	Duration of all outgoing calls to other countries
calls_outgoing_to_abroad_inactive_days	Number of days without outgoing call to other countries
sms_outgoing_count	Number of outgoing sms messages
sms_outgoing_spending	Amount spend on outgoing sms messages

sms_outgoing_spending_max	The most expensive sms message
sms_outgoing_inactive_days	Number of days without outgoing sms message
sms_outgoing_to_onnet_count	Number of outgoing sms messages to on-net
sms_outgoing_to_onnet_spending	Amount spend on outgoing sms messages to on-net
sms_outgoing_to_onnet_inactive_days	Number of days without outgoing sms message to on-net
sms_outgoing_to_offnet_count	Number of outgoing sms messages to off-net
sms_outgoing_to_offnet_spending	Amount spend on outgoing sms messages to off-net
sms_outgoing_to_offnet_inactive_days	Number of days without outgoing sms message to off-net
sms_outgoing_to_abroad_count	Number of outgoing sms messages to other countries
sms_outgoing_to_abroad_spending	Amount spend on outgoing sms messages to other countries
sms_outgoing_to_abroad_inactive_days	Number of days without outgoing sms message to other countries
sms_incoming_count	Number of incoming sms messages
sms_incoming_spending	Amount spent on incoming sms messages
sms_incoming_from_abroad_count	Number of incoming sms messages from other countries
sms_incoming_from_abroad_spending	Amount spend on incoming sms messages from other countries
gprs_session_count	Number of data connections
gprs_usage	Number of kb used
gprs_spending	Money amount spent on data
gprs_inactive_days	Number of days without data usage
last_100_reloads_count	Number of recharges over the last 100 days
last_100_reloads_sum	Amount of recharges over the last 100 days
last_100_calls_outgoing_duration	Calls outgoing duration over the last 100 days
last_100_calls_outgoing_to_onnet_duration	Calls outgoing to on-net duration over last 100 days
last_100_calls_outgoing_to_offnet_duration	Calls outgoing to off-net duration over last 100 days
last_100_calls_outgoing_to_abroad_duration	Calls outgoing to other countries duration over last 100 days
last_100_sms_outgoing_count	Number of SMS messages over 100 days
last_100_sms_outgoing_to_onnet_count	Number of SMS messages to on-net over 100 days
last_100_sms_outgoing_to_offnet_count	Number of SMS messages to off-net over 100 days
last_100_sms_outgoing_to_abroad_count	Number of SMS messages to other countries over 100 days
last_100_gprs_usage	Number of kb used over last 100 days

Source : <https://www.googleapis.com/download/storage/v1/b/kaggle-user-content/o/inbox%2F1335623%2F7777aa5eb46ee201808eed17aa82a866%2FData%20dictionary.xlsx?generation=1574255612876504&alt=me>

