Masters of Engineering

Internship report

Machine Learning Approaches to Customer Churn in BtoB context : An internship at Equatorial Coca Cola Bottling Company

Author: KAMEL TAREK

Supervisor: PR. AHMED AMAMMOU

FEBRUARY 2024 - JULY 2024

2 - Acknowledgements

I would like to express my deepest gratitude to everyone who supported and guided me during my internship at Equatorial Coca-Cola Bottling Company. This enriching experience would not have been possible without the invaluable contributions and encouragement from many individuals.

First and foremost, I extend my heartfelt thanks to my supervisor, MOHAMMED BENNANI, whose expertise, patience, and insightful feedback were instrumental in shaping my learning journey. Your mentorship and support have been invaluable, and I am profoundly grateful for the knowledge and skills I have gained under his guidance.

I am also immensely grateful to the entire team at Equatorial Coca-Cola Bottling Company. Your warm welcome and willingness to share your knowledge made my integration into the company seamless and enjoyable. A special thank you to my colleagues in the BTS team, whose camaraderie and collaborative spirit made every day at work a rewarding experience. Your dedication and passion for excellence are truly inspiring.

I would also like to acknowledge the leadership of the company for providing a conducive and stimulating environment for professional growth. The culture of innovation and continuous improvement at Equatorial Coca-Cola Bottling Company has left a lasting impression on me.

A heartfelt thank you to my professor, AMMAMOU Ahmed for your unwavering support and guidance throughout this internship. Your belief in my abilities, along with your invaluable advice and insights, have significantly contributed to my personal and professional growth. Your encouragement and mentorship have been a constant source of inspiration.

To my family and friends, thank you for your unwavering support and encouragement throughout this internship. Your belief in my abilities has been a constant source of motivation.

This internship has been a pivotal chapter in my career, and I am sincerely thankful to everyone who played a part in making it a success.

3- Abstract :

In today's competitive business landscape, the focus on customer retention over customer acquisition has become a widely acknowledged strategy for sustainable growth. It is imperative for companies to identify customers at risk of defection and proactively engage them with personalized incentives and retention offers. This necessitates the development of predictive models capable of pinpointing customers with a higher likelihood of churning in the near future.

Moreover, recognizing that not all customers hold equal value for the business underscores the importance of implementing customer segmentation strategies. By segmenting customers based on their unique needs and preferences, companies can tailor their offerings to better meet individual requirements. The primary objective of this study is to outline the methodologies and key challenges encountered in constructing a predictive algorithm and proposing various segmentation techniques to mitigate churn rates and enhance commercial campaigns Coca Cola Bottling Company ECCBC.

In contemporary marketing, the concept of customer churn holds significant importance and should not be overlooked by B2B companies[1]. With enhanced access to information, customers are more inclined to switch between competitors, making it easier and less costly for them to do so. Recognizing this, firms aim to identify potential churners and prevent defection by offering incentives. The primary motivations behind the current study can be condensed into two key points:

The beverage market is highly competitive, currently experiencing a price war.

Competitors are reducing commissions to attract more business and safeguard their customer base.

This trend impacts ECCBC’s ability to maintain margins and retain its customer base. Some customers migrate to competitors offering more affordable options.

Leveraging the wealth of untapped user data stored in various systems, the study seeks to harness machine learning algorithms to predict and prevent customer churn. Additionally, segmentation techniques are explored to enhance cross-selling strategies and foster customer loyalty [6].

Dans le contexte concurrentiel actuel des affaires, l'importance de la fidélisation de la clientèle par rapport à son acquisition est devenue une stratégie largement reconnue pour assurer une croissance durable. Les entreprises doivent absolument repérer les clients qui risquent de partir et les encourager activement avec des incitations personnalisées et des offres de fidélisation. Pour ce faire, il est nécessaire de développer des modèles prédictifs capables d'identifier les clients ayant le plus de chances de partir prochainement.

De plus, la reconnaissance que tous les clients ne sont pas également précieux souligne l'importance de mettre en œuvre des stratégies de segmentation de la clientèle. En divisant les clients en fonction de leurs besoins et préférences uniques, les entreprises peuvent adapter leurs offres pour mieux répondre à chaque client. L'objectif principal de cette étude est de décrire les méthodes et les défis majeurs rencontrés dans la création d'un algorithme prédictif, ainsi que de proposer différentes techniques de segmentation pour réduire les taux de départ des clients et améliorer les campagnes commerciales de la Coca Cola Bottling Company ECCBC.

Dans le domaine du marketing moderne, la notion de départ des clients est extrêmement importante et ne doit pas être négligée par les entreprises B2B. Avec un accès accru à l'information, les clients sont plus enclins à changer de fournisseur, ce qui rend cela plus facile et moins coûteux pour eux. Ainsi, les entreprises s'efforcent d'identifier les clients à risque de départ et de les retenir en leur proposant des incitations. Les motivations principales de cette étude peuvent être résumées en quatre points clés : le marché des boissons est fortement concurrentiel, avec une guerre des prix en cours. Les concurrents réduisent leurs commissions pour attirer davantage de clients et protéger leur base existante.

Cette tendance affecte la capacité de l'ECCBC à maintenir ses marges et à retenir ses clients. Certains clients se tournent vers des concurrents offrant des options plus abordables. En exploitant la richesse des données utilisateurs inexploitées stockées dans divers systèmes, l'étude cherche à utiliser des algorithmes d'apprentissage automatique pour prédire et prévenir le départ des clients. De plus, des techniques de segmentation sont étudiées pour améliorer les stratégies de vente croisée et encourager la fidélité des clients.

4- List of figures :

5. List of Tables

6.Abrevieations and acronymes

B2B : Business To Business

RFM :Recency, Frequency, Monetary

CLTV : Customer LifeTime Value

ICP: Ideal Customer Profile

ML : Machine Learning

EDA : Exploratory Data Analysis

ECCBC : Equatorial Coca Cola Bottling Company

DT: Decision Tree

RF: Random Forest

LR: Logistic Regression

XGBoost: Gradient Boost

KNN: K nearest Neighbour

SVM: Support Vector Machine

SAP: System Analysis Program Development (Systemanalyse Programmentwicklung)

7. Table of content :

8. Introduction :

In today's competitive business world, keeping existing customers has become just as important as attracting new ones. Companies have realized that building long-term relationships with their customers is key to sustainable growth. This shift means businesses need to spot customers who might leave and reach out to them with personalized offers and incentives. Predictive models are crucial in this effort, helping companies identify which customers are likely to churn, or stop buying from them, soon.

Not all customers are the same, though. Some are more valuable to the business than others. That's why customer segmentation is so important. By understanding the unique needs and preferences of different customer groups, companies can tailor their products and services to better meet individual expectations. This study focuses on developing a predictive algorithm to identify potential churners and using segmentation techniques to help Coca Cola Bottling Company ECCBC retain their customers and improve their marketing efforts.

This report is organized to provide a clear understanding of the project's context, significance, and objectives. We start with an overview of the host organization, Equatorial Coca Cola Bottling Company, discussing the motivations behind the internship and the steps taken during the project. This sets the stage for understanding the business environment and the specific challenges we aimed to address.

Next, we dive into the state-of-the-art methods and tools we used. We describe the dataset, the technical environment, and the libraries we worked with, along with current techniques in RFM churn analysis on SAP transactional data. This section also covers the methodologies we adopted and the initial results from our RFM analysis, laying the groundwork for our predictive modeling.

In the third part of the report, we detail our methodologies, including how we predicted churn, segmented customers, assessed data quality, and prepared our data. We explain the process of defining our target variable and conducting exploratory data analysis (EDA) on different datasets. We also evaluate the performance of various machine learning models, finding that Support Vector Machines (SVM) were the most effective for predicting churn. We introduce new key performance indicators (KPIs) and discuss the deployment of our model and the creation of a user-friendly dashboard interface.

The main goal of this study is to use machine learning and customer segmentation to predict and reduce customer churn, enhancing ECCBC's marketing strategies and operational efficiency. Through detailed experimentation and analysis, we developed effective methods and overcame challenges, ultimately providing valuable insights for the sales teams. In the following chapters, we will explore these methodologies, results, and their implications in detail, giving you a thorough understanding of the project's impact on ECCBC’s strategy and performance.

9. Main chapters:

- **Chapter 1: Overview of the Host Organization, Context, Objectives, and Project Execution**

1. Host organization :

1.1. General overview of Equatorial Coca Cola Bottling Company:

Equatorial Coca-Cola Bottling Company (ECCBC) stands as the esteemed bottling partner of The Coca-Cola Company across North and West Africa. With a rich legacy spanning decades, ECCBC's operations encompass the production, commercialization, and distribution of an array of beverages, including iconic brands such as Coca-Cola, Fanta, and Sprite [4].

Originating in 1989 in Equatorial Guinea, ECCBC's influence has extended to various countries across the region, including Guinea Conakry, Mauritania, Cape Verde, Guinea Bissau, and The Gambia. Over time, ECCBC has evolved, undergoing a re-foundation in 1997 to fuel expansion into new territories like Ghana, Morocco, and Algeria [4].

A significant milestone was reached in 2021 when ECCBC completed the transfer of most of its functions and management team from Barcelona, Spain, to its new headquarters in Casablanca, Morocco. This strategic move reflects ECCBC's unwavering confidence in Africa's potential, its commitment to leadership in the bottling industry, and its dedication to fostering the continent's social and economic progress [4].

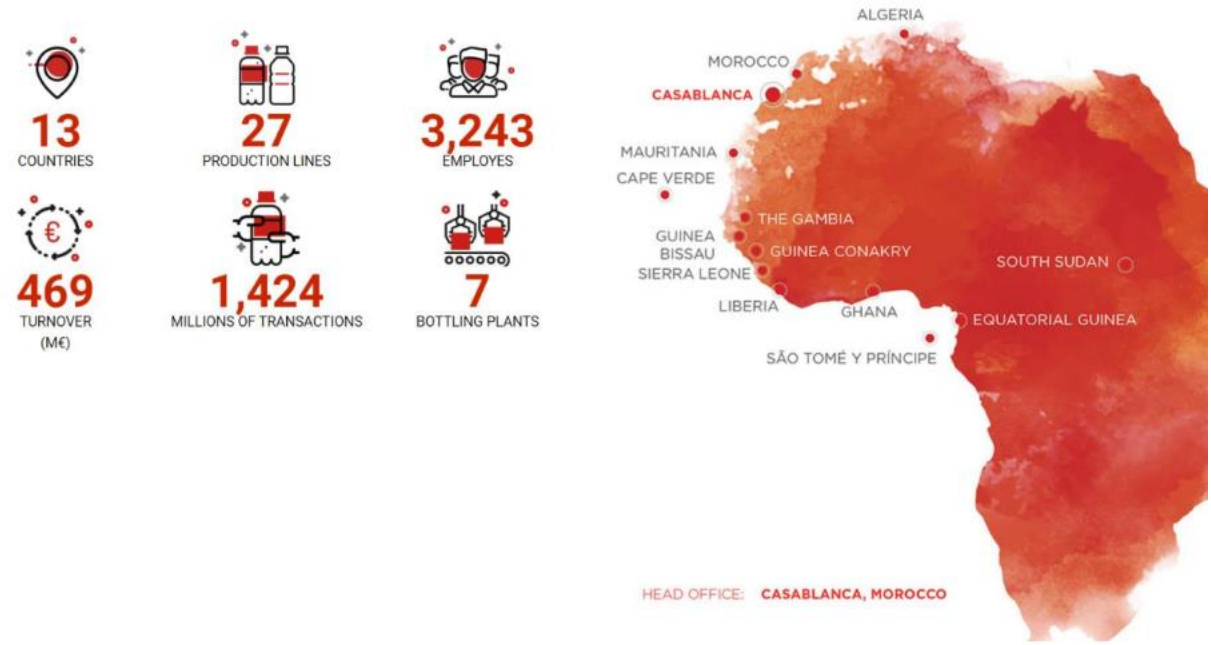


Figure 1: ECCBC IN NUMBERS.

2. Motivations and context of the internship

In recent decades, the proliferation of marketing-related data, including scanner data and internet data, coupled with organizations' increasing demand for novel analytical approaches, has spurred growing interest in Artificial Intelligence (AI)-based solutions for marketing challenges. Among the various roles that AI systems can fulfill in marketing, predictive modeling, particularly churn modeling, stands out as highly promising [3].

Customer churn prediction entails assessing the likelihood of future churn behavior for each customer in the database using predictive models based on past information and behavior. Therefore, to develop effective customer retention strategies, the models employed must strive for maximum accuracy to avoid wasteful spending on customers unlikely to churn. Consequently, data mining techniques rooted in AI have become popular for modeling churn due to their emphasis on predictive capability [3].

However, research directions advocating for the application of strategic intelligence in industrial marketing, such as business intelligence, competitive intelligence, and knowledge management, have received inadequate attention from both academia and industry. This gap is attributed to limited availability of B2B 'big data' compared to B2C contexts, where mining large datasets to extract customer insights is less common. Moreover, while there is existing literature on churn prediction across various sectors, the majority focuses on B2C contexts, leaving B2B churn prediction underexplored. Nonetheless, many supervised learning techniques utilized in B2C markets can also be applied to B2B contexts [1] [3].

This research investigates the effectiveness of various machine learning algorithms in predicting customer churn within. We will explore a range of algorithms, each with distinct strengths and weaknesses [6]:

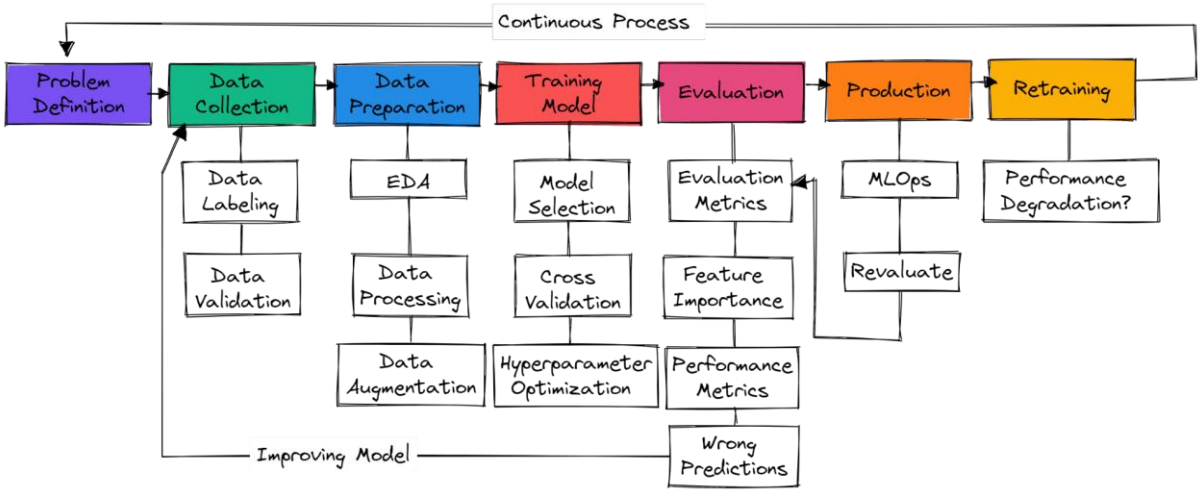


Figure 2: CHURN PREDICTION WORKFLOW

3. Chronological Description of Execution Steps:

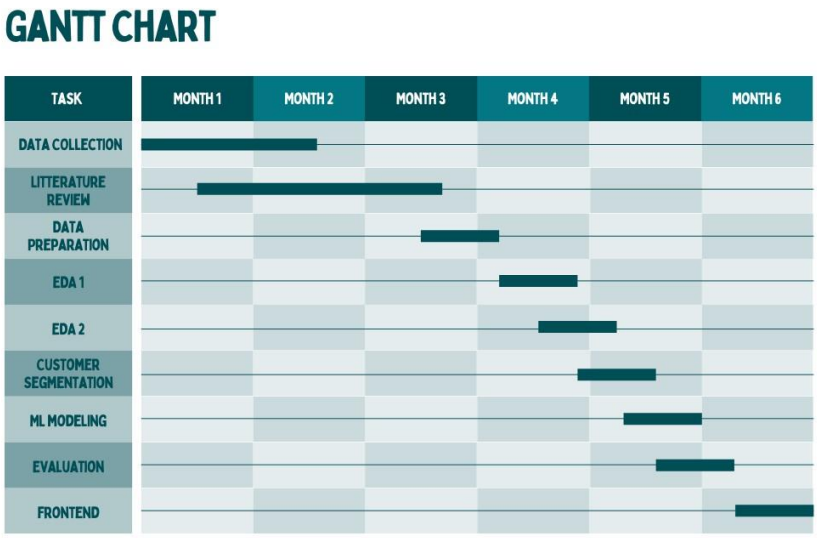


Figure 3:PROJECT GANT CHART

- **Chapter 2: State of the Art and Methods Used**

1. Provided DataSet Description :

The dataset provided during the internship at ECCBC consists of sales data originating from SAP spanning the year 2023 (Approx. 6.5 Million rows after cleaning and preprocessing). It encompasses various dimensions essential for analyzing sales performance and customer behavior. The primary dimensions included in the dataset are as follows:

Table 1: SAP Data table

	DATE_ISO	SALES_HEADER_ID	CUSTOMER_ID	PHYSICAL_CASES	UNIT_CASES	AMOUNT_EUR	AMOUNT_PROMO_EUR	AMOUNT_PERMA_EUR	PRIME_LOYALTY_EUR	SALES_HEADER_COD	...	ARTICLE_COD	CATEGORY_ID	CUSTOMER_TYPE_ID	TERRITORY_ID	CHANNI
323	2023-10-06	73-11-20231006-865399879-3860934	3860934	4,75521	27	0.0	-6,28	0	0	8.654e+08	...	7316	135	0	-1	
1674	2023-10-06	73-11-20231006-865399529-1004805	1004805	1,40895	8	0.0	-3,58	0	0	8.654e+08	...	7521	-1	0	-1	
1675	2023-10-06	73-11-20231006-865399529-1004805	1004805	1,58507	9	0.0	-2,09	0	0	8.654e+08	...	7316	135	0	-1	
1903	2023-10-05	73-11-20231005-865374540-3118898	3118898	1,40895	8	0.0	-3,58	0	0	8.65375e+08	...	7521	-1	0	-1	
2145	2023-10-03	72-01-20231003-865321486-7724609	7724609	1,58507	9	0.0	-2,12	0	0	8.65321e+08	...	7316	135	0	-1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
6506171	2023-12-28	73-11-20231228-867202822-1004929	1004929	2,09229	8	0.0	-5,07	0	0	867203000.0	...	7346.0	-1.0	0.0	-1.0	
6506257	2023-12-28	73-11-20231228-867205272-3460848	3460848	2,09229	8	0.0	-5,07	0	0	867205000.0	...	7346.0	-1.0	0.0	-1.0	
6506287	2023-12-28	73-11-20231228-867205594-1000713	1000713	4,88201	8	0.0	-11,83	0	0	867206000.0	...	7346.0	-1.0	0.0	-1.0	
6506319	2023-12-28	73-11-20231228-867198079-3399047	3399047	4,75521	27	0.0	-6,41	0	0	867198000.0	...	7316.0	135.0	0.0	-1.0	
	2023-12-	73-11-20231228-														

Additionally, the dataset may contain other dimensions that were removed due to various reasons, such as redundancy or irrelevance to the analysis objectives.

The dataset offers valuable insights into sales performance and various transaction dimensions. However, it lacks specific information regarding customer churn, a critical aspect for understanding customer behavior and retention strategies. To bridge this gap, we took the initiative to create a new dataset derived from the original transactional one in which RFM scores are calculated for each unique customer

In this dataset, customers are scored based on their recency, frequency, and monetary (RFM) values extracted from their transaction history [2]. The RFM model is a widely used method that segments the customers based on their purchasing behavior [2].

-Recency: Reflects how recently a customer made a purchase, assigning higher scores to recent purchasers.

-Frequency: Indicates how often a customer makes purchases, assigning higher scores to frequent purchasers.

-Monetary: Represents the monetary value of a customer's purchases, assigning higher scores to those with higher spending.

By analyzing RFM scores, we can identify customer segments more prone to churn and customize retention strategies accordingly. This enhanced dataset offers valuable insights into customer behavior, facilitating more informed decision-making to enhance customer retention and overall business performance [2].

This RFM dataset will act as a valuable sample for training our machine learning algorithm to predict customer churn. Through advanced analytics techniques, we aim to discern patterns and trends within the data that may indicate potential churn behavior. This predictive model will empower us to proactively address customer attrition by implementing targeted retention strategies [6].

Table 2: RFM Table

	Recency	Frequency	Monetary	Recency_Score	Frequency_Score	Monetary_Score	RFM_Score
count	111197.000000	111197.000000	111197.000000	111197.000000	111197.000000	111197.000000	111197.000000
mean	9.898873	42.770012	6.931689	2.380154	2.452027	0.138556	4.970737
std	12.450935	46.781669	147.693650	1.779269	1.729945	0.345484	1.627764
min	0.000000	1.000000	-441.000000	0.000000	0.000000	0.000000	0.000000
25%	2.000000	14.000000	0.000000	1.000000	1.000000	0.000000	4.000000
50%	5.000000	28.000000	0.000000	2.000000	2.000000	0.000000	5.000000
75%	13.000000	55.000000	0.000000	4.000000	4.000000	0.000000	6.000000
max	59.000000	1583.000000	37609.000000	5.000000	5.000000	1.000000	11.000000

Additionally, we thought about joining our new RFM data set with costumer data table on customer ID in order to provide our predictive model with more feature variables than can say a lot about what is hidden behind our data, this includes:

CUSTOMER_ID: A unique identifier for each customer.

COMPANY_COD: The code representing the company associated with the customer.

TERRITORY_ID: Identifier for the territory associated with the customer.

TERRITORY: The name or description of the territory.

DISTRIBUTION_MODEL: Describes the distribution model used for supplying products or services to the customer (Direct/ Indirect).

CUSTOMER_SEGMENT_ID: Identifier for the segment to which the customer belongs (AG/Café/AGHybrid/DL/SNACK/Dry Channel/Boulangerie/Hotel/Night).

LOCATION_ID: Identifier for the location of the customer.

DISCOUNT_COD: Code representing the discount applied to the customer.

SEGMENTATION: Indicates the segmentation category of the customer (Bronze/Silver/Gold).

COM_DELETE: A flag or indicator for whether the customer record has been deleted.

GEN_DELETE: Another deletion indicator, possibly for general reasons.

PARTNER_CLASS: Classification or category of the customer as a partner (RGB Non RGB).

SALES_ROUTE: Route or method used for sales to the customer.

LONGITUDE: Geographic longitude coordinate of the customer's location.

LATITUDE: Geographic latitude coordinate of the customer's location.

Table 3:Merged table

	COMPANY_COD	TERRITORY_ID	TERRITORY_COD	TERRITORY	DITRIBUTION_MODEL	CUSTOMER_SEGMENT_ID	LOCATION_ID	LOCATION_COD	CUSTOMER_SEG	CANAL	...	Recency	Frequency	Monetary	Recency_Score	Frequency_Score	Monetary_Score	RFM_
0	M100	106	M00005 ...	Sud ...	DIRECT ...	2	13	M120	AG ...	AG ...	...	1	517	165.0	0	5	4	
1	M100	106	M00005 ...	Sud ...	INDIRECT ...	76	13	M120	AG HYBRIDE ...	AG ...	...	1	343	57.0	0	5	3	
2	M100	106	M00005 ...	Sud ...	DIRECT ...	2	13	M120	AG ...	AG ...	...	2	196	59.0	1	5	3	
3	M100	106	M00005 ...	Sud ...	DIRECT ...	2	13	M120	AG ...	AG ...	...	2	102	19.0	1	5	1	
4	M100	106	M00005 ...	Sud ...	DIRECT ...	2	13	M120	AG ...	AG ...	...	1	274	19.0	0	5	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
140	M100	108	M00007 ...	Centre ...	DIRECT ...	2	15	M140	AG ...	AG ...	...	2	115	62.0	1	5	3	
141	M100	103	M00002 ...	Atlantic ...	DIRECT ...	7	9	M100	DL ...	DL ...	...	2	33	89.0	1	4	4	
142	M100	108	M00007 ...	Centre ...	DIRECT ...	2	10	M110	AG ...	AG ...	...	2	377	93.0	1	5	4	
143	M100	108	M00007 ...	Centre ...	DIRECT ...	7	10	M110	DL ...	DL ...	...	2	317	31.0	1	5	2	
144	M100	108	M00007 ...	Centre ...	DIRECT ...	2	10	M110	AG ...	AG ...	...	2	608	191.0	1	5	4	

145 rows × 38 columns

2. Environment and libraries :

For the project, we employed a diverse range of tools and technologies to facilitate data analysis, model development, and predictive analytics. The following is an overview of the environment and tools utilized:

1. Jupyter Notebook: We utilized Jupyter Notebook as our primary platform for data exploration, analysis, and model development. Its interactive interface and support for various programming languages, including Python, made it an ideal choice for our project.

2. Python: Python served as the primary programming language for implementing data preprocessing, feature engineering, and model building tasks. Its extensive libraries and frameworks provided robust support for data science and machine learning tasks [6].

3. Pandas: Pandas, a powerful data manipulation library in Python, was instrumental in handling and processing large datasets efficiently. It facilitated tasks such as data cleaning, transformation, and aggregation.

4. Matplotlib and Seaborn: Matplotlib and Seaborn were utilized for data visualization purposes. These libraries enabled us to create insightful charts, plots, and graphs to visualize trends, distributions, and relationships within the data.

5. Scikit-learn (sklearn): Scikit-learn, a popular machine learning library in Python, was utilized for implementing various machine learning algorithms, including classification, regression, and clustering. Its comprehensive set of tools and algorithms made it indispensable for model development and evaluation.

6. Orange3: Orange3, an open-source data visualization and analysis tool, was utilized for its intuitive interface and visual programming capabilities. It provided a user-friendly environment for data exploration, feature selection, and model building.

7. SQL: SQL (Structured Query Language) was employed for data querying and manipulation tasks, particularly for extracting data from relational databases and performing data preprocessing tasks.

8. Azure Machine Learning Studio: Azure Machine Learning Studio was utilized for cloud-based model deployment and experimentation. Its scalable infrastructure and integrated

development environment facilitated collaborative model development and deployment in a cloud environment.

9. Power BI : a comprehensive business analytics solution designed by Microsoft, offering users the capability to visualize and analyze data from diverse sources. This powerful tool enables the creation of interactive reports and dashboards for effective data-driven decision-making.

By leveraging these tools and technologies, we were able to create a robust and scalable environment for data analysis, model development, and predictive analytics, enabling us to derive actionable insights and drive informed decision-making processes.

3. State of the Art: RFM Churn Analysis on SAP transactional data :

Before delving into the specifics, let's grasp the concept of RFM. Then, we'll address the crucial question: How can this data be leveraged to expedite sales funnel and business growth in a B2B wholesale and distribution setting?

RFM, or Recency, Frequency, and Monetary analysis, serves as a potent instrument for retail companies to gain insights into their customer base. By dissecting customer behavior across three fundamental dimensions, RFM enables businesses to customize their marketing strategies, enhance customer experiences, and ultimately fuel growth [2].

Recency pertains to how recently a customer conducted a transaction. This dimension underscores the significance of staying top-of-mind with customers and prompts businesses to engage promptly after a sale. Understanding recency allows retailers to spot inactive customers who may require reactivation campaigns or personalized incentives to reignite their engagement.

Frequency gauges how frequently a customer makes purchases within a specific timeframe. By scrutinizing frequency, companies can differentiate between sporadic buyers and loyal patrons. Identifying high-frequency customers presents opportunities for loyalty programs, subscription models, or targeted cross-selling and upselling endeavors to optimize customer lifetime value.

Monetary value quantifies the monetary amount a customer spends on purchases. This dimension assists retailers in pinpointing their highest-spending customers, guiding decisions on resource allocation, pricing strategies, and providing VIP treatment to valuable segments.

Additionally, analyzing monetary value allows retailers to discern patterns in spending behavior and adapt their product offerings or pricing accordingly.

By demystifying RFM principles and harnessing cost-effective analytics solutions, companies can harness the power of data-driven decision-making without the need for significant investments in consultants or data analysts. Embracing RFM empowers companies to gain deeper insights into their customer base, identify growth opportunities, and compete more effectively in today's dynamic marketplace [2].

Why is RFM analysis essential?

In essence, RFM emerges as a paramount strategy for wholesalers and distributors. But what exactly does this entail?

Every business desires a systematic and foreseeable approach to foster growth. Effectively, it's akin to possessing the ability to anticipate each customer's needs at every stage of their interaction with your business. This is precisely what RFM facilitates [2].

However, before delving deeper, let's briefly explore modern theories of sales and marketing.

In any business, the goal is to attract new customers who match your Ideal Customer Profile (ICP) and guide them through the sales process to improve conversion rates.

Achieving these objectives effectively will expand our customer base, which is beneficial. However, in wholesale and distribution, customer acquisition goes beyond securing new leads [2]. It involves effectively onboarding customers, enhancing their average order value and frequency, and fostering long-term relationships to maximize Customer Lifetime Value (CLTV). Essentially, the focus shifts towards existing customers to optimize profitability[2].

To establish predictable and sustainable growth, it's imperative to secure several initial orders from customers, establishing your brand as their preferred supplier. Subsequently, maintaining long-term relationships, upselling, and reactivating customers if their engagement diminishes are vital. But how does RFM analysis fit into this equation?

To achieve profitable growth, it's crucial to precisely understand where each customer stands in your sales funnel and flywheel, tailor your approach accordingly, and communicate effectively at each stage of the process, efficiently and at scale. Whether it's mass marketing or personalized account management, RFM analysis enables targeted outreach, ensuring the right message reaches the right customer at the right time [2].

4. Methods Used:

Initially, you analyze your customer data to establish sensible ranges for recency, frequency, and monetary value. Then, you divide these ranges into three or five equal segments, either based on value or size. While it's crucial for the segments to be evenly distributed, it often

works best when each segment contains an equal number of customers. Subsequently, each customer is placed into one of these segments and assigned a corresponding score.

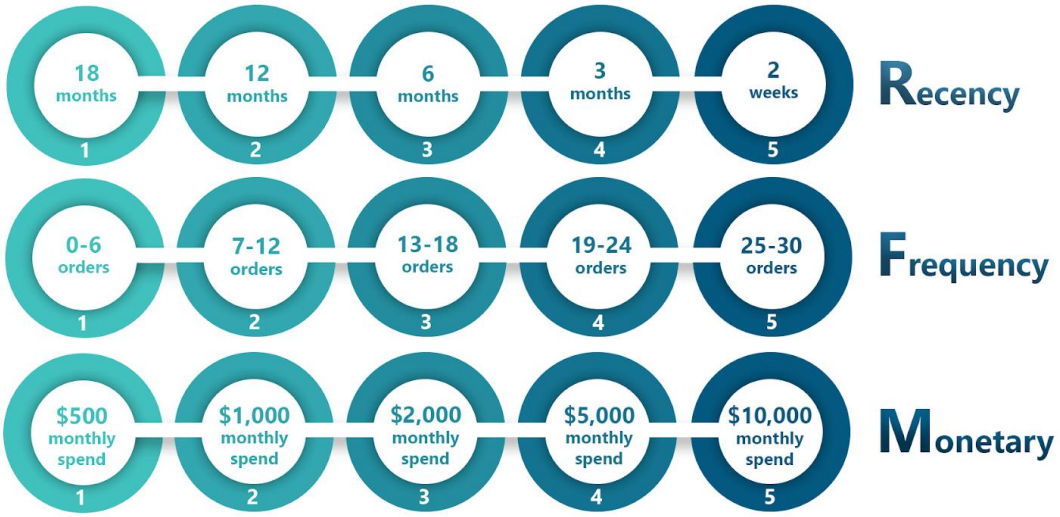


Figure 4: Scores ranges

Here, we observe a spectrum of values representing recency, frequency, and monetary value. Utilizing the provided chart, we can assign scores for recency to our customers. For instance, a customer who made a purchase 11 months ago would receive a score of "1", whereas a customer who made a purchase just two weeks ago would be assigned a score of "5". Naturally, these categories can be customized to better align with the typical purchasing behavior of customers, accounting for factors such as average order frequency and expenditure.

5. Definition of Target Variable:

In this study, the target variable is "churn", a variable previously inexistant which characterizes the churn behavior of business customers based on their transactional history within a specified timeframe. To determine whether a user churned or not, it was necessary to establish the length of the historical data used for training the algorithm. Due to data quality considerations, the transaction history was restricted to the year 2023, Consequently, the observation window for each user was defined as the last mounth of 2023, due to the BtoB context.

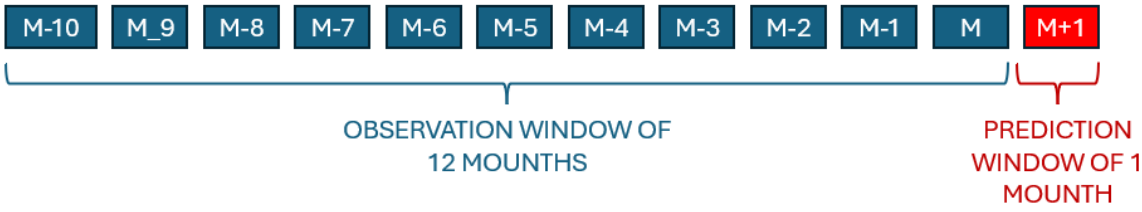


Figure 5:Prediction window in timeline

We aimed to enrich our churn prediction model by broadening the range of the target variable beyond the traditional 'Churn' and 'Not churn' labels. To bolster early detection capabilities, We introduced additional classifications, specifically 'Low Risk' and 'High Risk'. These supplementary categories offer a more nuanced comprehension of customer behavior, enabling the model to identify potential churn indicators before they fully materialize. By incorporating 'Low Risk' and 'High Risk' designations, the model gains the capacity to proactively address customer disengagement or dissatisfaction, fostering a proactive retention strategy.

6. Assigning Churn Status :

Based on extensive discussions and multiple interviews with the marketing and sales teams, we developed a set of conditions to accurately assign churn status to customers. These conditions consider various factors such as the recency and frequency of customer interactions, as well as whether the customer is new or existing. The following criteria outline how we determine the risk level and churn status for each customer:

-If the recency score is less than or equal to 3 and the customer is a new costumer (next_1_months = 'Yes'), the churn status is assigned as 'No', indicating no churn.

-If the recency score is less than or equal to 3 and the customer is not a new costumer (next_1_months = 'No'), it further checks the frequency score:

-If the frequency score is less than or equal to 3, indicating low frequency, the churn status is assigned as 'High Risk'.

-Otherwise, if the frequency score is greater than 3, indicating higher frequency, the churn status is assigned as 'Low Risk'.

-If the recency score is greater than 3 and the customer is new (next_1_months = 'Yes'), it further checks the frequency score:

-If the frequency score is greater than 3, indicating higher frequency, the churn status is assigned as 'No', indicating no churn.

-Otherwise, if the frequency score is less than or equal to 3, the churn status is assigned as 'Low Risk'.

-If the recency score is greater than 3 and the customer is new (next_1_months = 'Yes'), the churn status is assigned as 'Yes', indicating churn.

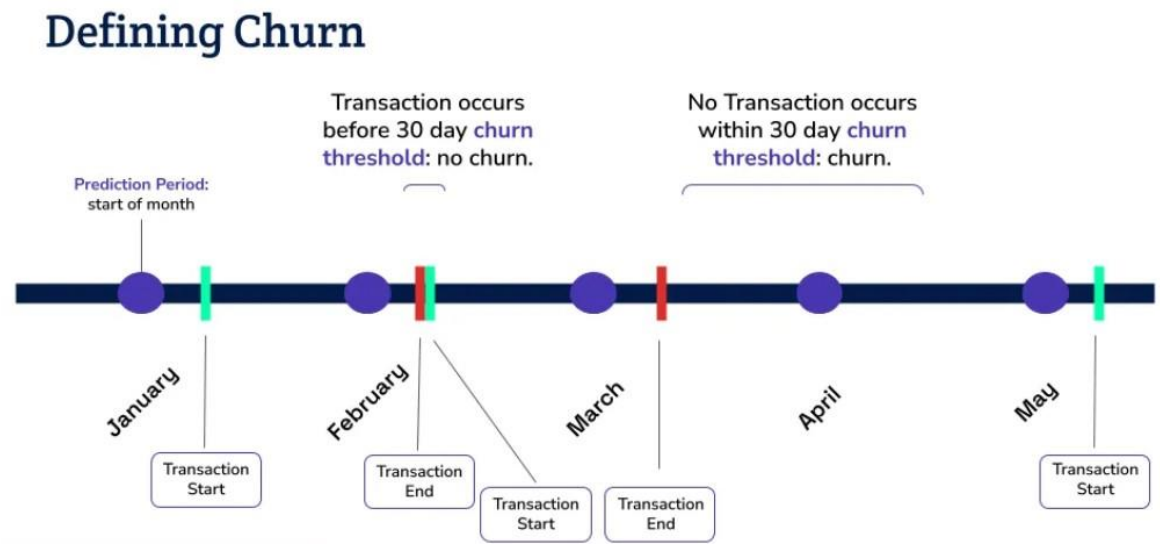


Figure 6:churn labelization procedure

The calibration period's duration is crucial, as an excessively long interval may overlook imminent churn events, while a too short interval may yield false positives. Determining the optimal period involves empirical experimentation and collaboration with the marketing team to align with business objectives.

7- RFM results :

Table 4:RFM table

	Recency	CUSTOMER_ID	Frequency	Monetary	Recency_Score	Frequency_Score	Monetary_Score	RFM_Score	Segment_Label	Seg_Num	R_Next_3Months	R_Next_1Month	Churn
0	1	100009.0	517	165.0	5	0	4	9	Prime Customer	3	No	No	Yes
1	1	100010.0	343	57.0	5	0	3	8	Prime Customer	3	No	No	Yes
3	2	100013.0	196	59.0	4	0	3	7	Standard Customer	2	No	No	Yes
4	2	100016.0	102	19.0	4	0	1	5	Standard Customer	2	No	No	Yes
5	1	100017.0	274	19.0	5	0	1	6	Standard Customer	2	No	No	Yes

The RFM dataframe table, encompassing data from 15407 unique customers, provides a comprehensive snapshot of customer engagement and transaction behavior it offers valuable insights into customer interactions and spending habits. These insights facilitate targeted marketing strategies, personalized promotions, and effective customer retention initiatives, ultimately driving business growth and enhancing customer satisfaction.

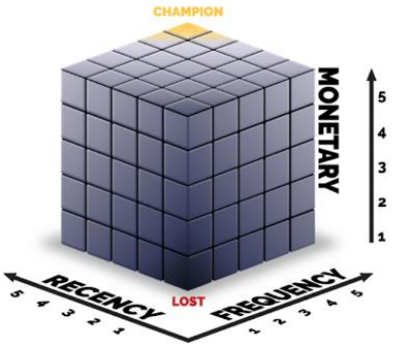


Figure 7 : 3d RFM representation

- **Chapter 3: Methodology, Proposed Solutions, and Results**

1. Methodologies :

1.1 Churn Prediction Procedure

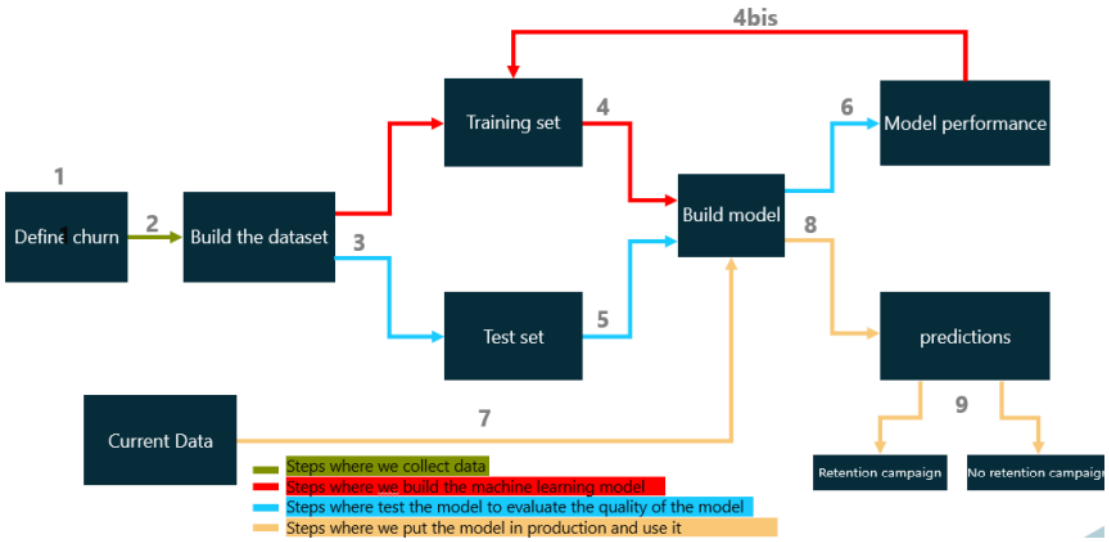


Figure 8: Prediction procedure

The creation of a churn model involves a series of steps outlined in Figure bellow. This methodology has been implemented, and the tasks executed at each stage will be elaborated upon in subsequent sections.

1.2 Customers segmentation:

Following the churn prediction analysis, we extended our exploration by segmenting the customer base using the same dataset. Utilizing the elbow method within K-means clustering, we determined the optimal number of clusters, resulting in the classification of customers into four categories: Basic Customer, Standard Customer, Prime Customer, and Elite Customer. This segmentation approach offers valuable insights into customer behavior and preferences, enabling tailored strategies and personalized interactions to enhance customer management and engagement.

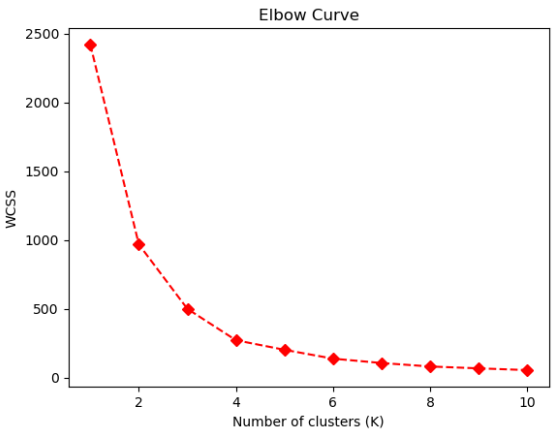


Figure 9: Elbow curve

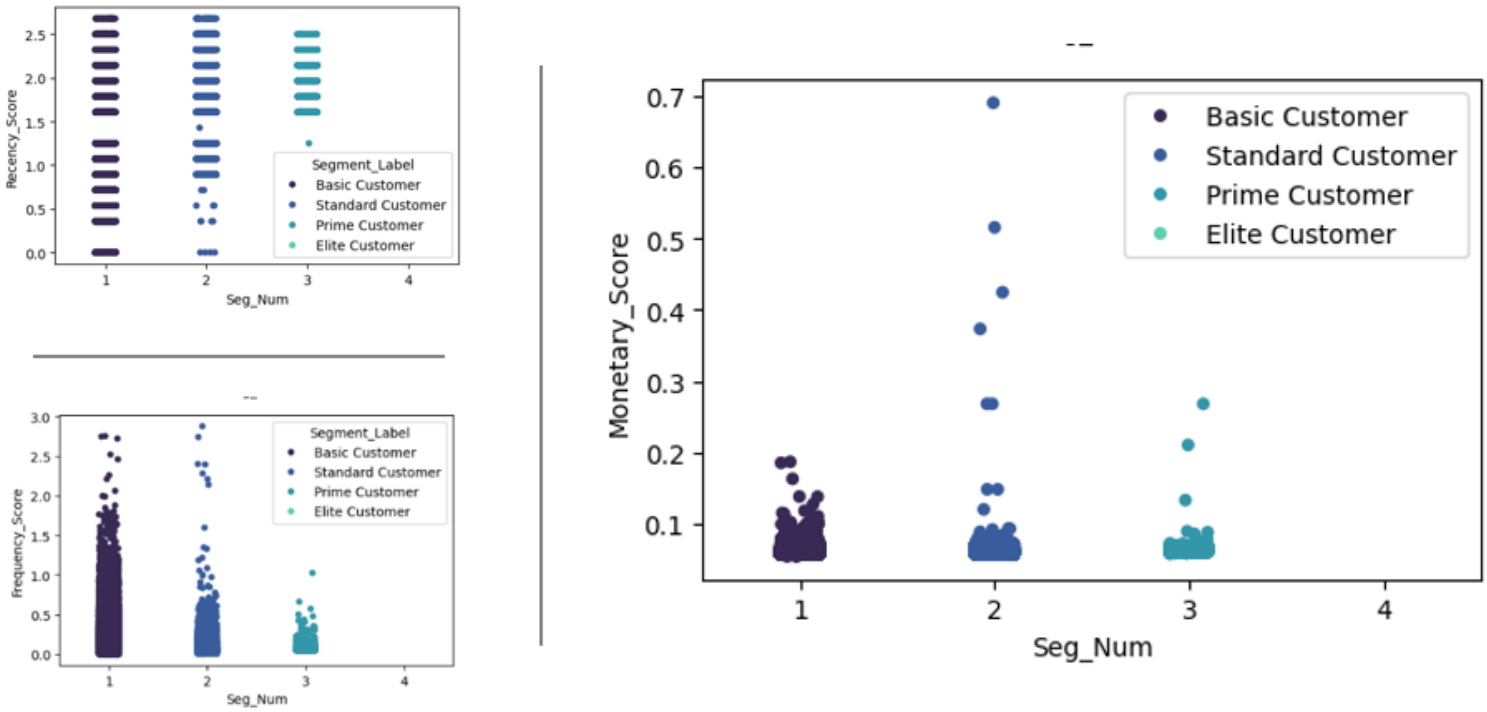


Figure 10 :Distribution of RMF scores by Segment

1.3 Data Quality Assessment and Preprocessing Methods

Data preparation, often termed pre-processing, involves transforming raw data into a usable format by cleansing and formatting it to enhance its quality and suitability for analysis. Additionally, data transformation techniques such as scaling, attribute decomposition, and aggregation further refine the data and prepare it for machine learning algorithms [6].

The analysis utilized a sales table as the primary data source, focusing on transactional data to create an RFM dataset with unique customer IDs and their recency, frequency, and monetary scores, alongside customer segment details.

We delved into the nuts and bolts of data cleaning and preprocessing on our dataset. Imagine it as tidying up before a big event, making sure everything is in its place and looking its best.

First up was handling any missing or incomplete data. Just like tidying a messy room, we sorted through our dataset to fill in the gaps where data was missing or decide if it was best to leave those parts out. This ensured that our analysis wouldn't be thrown off by any holes in the data.

Next, we tackled the fact that a lot of our data was categorical, meaning it was in the form of categories rather than numbers. To make it easier for our analysis, we transformed these categories into numbers—a bit like translating between languages so everyone can understand each other.

Then came scaling the data, which is like making sure all the ingredients in a recipe are in the right proportions. We adjusted the numerical values so that they all played nicely together, without any one ingredient overpowering the dish—or in our case, skewing the analysis.

By taking care of these behind-the-scenes tasks, we set the stage for our analysis to shine. We ensured that our dataset was clean, organized, and ready for action, giving us the best possible chance of uncovering meaningful insights and making informed decisions.

Following the integration of datasets and employing additional cleansing and preparation using pipelines, we proceeded with feature selection to refine dataset quality and integrity. During this phase, we generated visual representations to illustrate feature distributions, with a focus on identifying irrelevant ones.

These visualizations unveiled intriguing patterns. Some features exhibited significant bias, with data predominantly concentrated within limited ranges. Such skewed distributions could introduce biases into subsequent analyses or machine learning models.

For instance, multiple graphs portrayed features heavily skewed towards specific values, indicating a dominant presence of specific categories or attributes. This bias might inflate the importance or predictive power of the feature in downstream analysis, potentially compromising model performance.

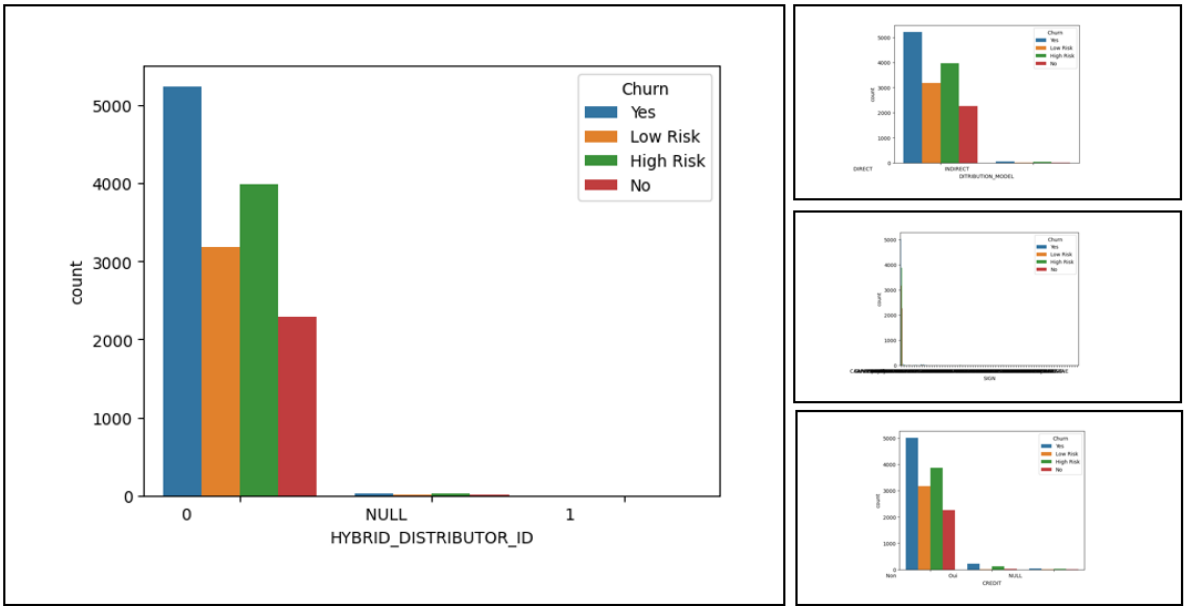


Figure 11:Examples of skewed features

Conversely, other features displayed more balanced distributions, with data spread relatively evenly across different values or categories. These features are likely to contribute more effectively to the analysis without introducing undue biases or distortions.

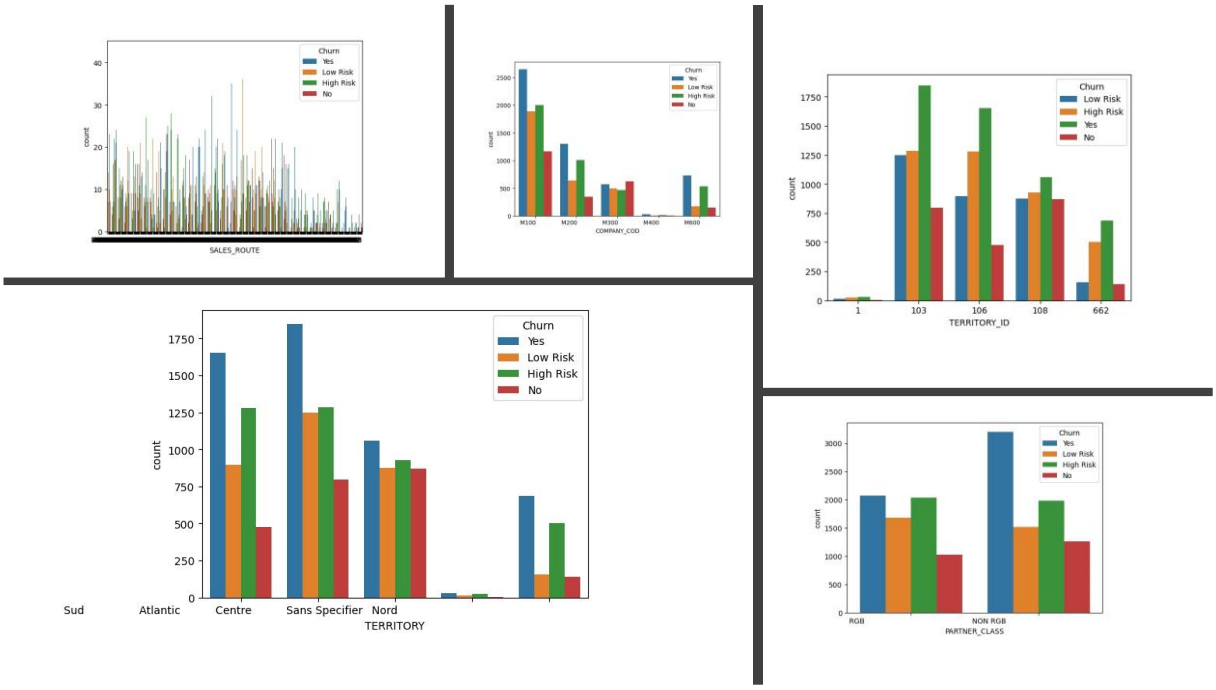


Figure 12:Example of good feature distributions

In summary, examining feature distributions visually emphasized the importance of robust data preprocessing and feature selection techniques. By identifying and mitigating biases in the dataset, we can ensure that subsequent analyses and models are built upon a sturdy foundation, yielding more dependable insights and predictions.

3. Predictive Modeling:

3.1 Exploratory Data Analysis (EDA)

It encompasses diverse statistical techniques such as text analytics, and predictive modeling to forecast future outcomes based on historical and existing data. Exploratory Data Analysis (EDA) serves as an approach to summarize key characteristics of datasets and extract meaningful insights [6].

Through statistical graphics and visualization methods, EDA maximizes insights into datasets, tests hypotheses, and facilitates model development. EDA represents not just a set of techniques but an attitude towards data analysis, emphasizing structured and comprehensive data exploration before engaging in statistical or machine learning modeling. Central to EDA are data preparation and visualization, involving data cleaning, formatting, and exploration to derive meaningful insights and prepare data for further analysis [11].

3.2 EDA (Exploratory Data Analysis) On RFM dataset

The pie chart illustrates the distribution of customer churn risk levels within our dataset. The majority of customers provided in the dataset, comprising 35.6%, are classified as "Yes" for churning, indicating a quite stable customer base indicating confirmed churn cases that require targeted intervention strategies for customer retention. However, a significant portion, representing 26.9%, falls under the category of "high risk," signifying a potential churn threat that warrants proactive retention efforts. A smaller proportion of customers, at 21.6%, are deemed "low risk," suggesting minimal likelihood of churn. Lastly, a segment, comprising 15.9% of the total, falls into the "no" category.

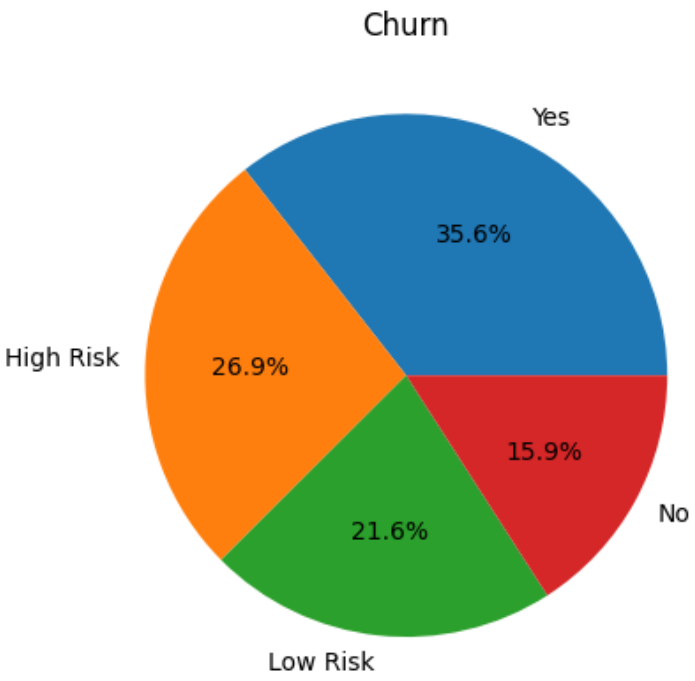


Figure 13:Distribution of Churn by target class

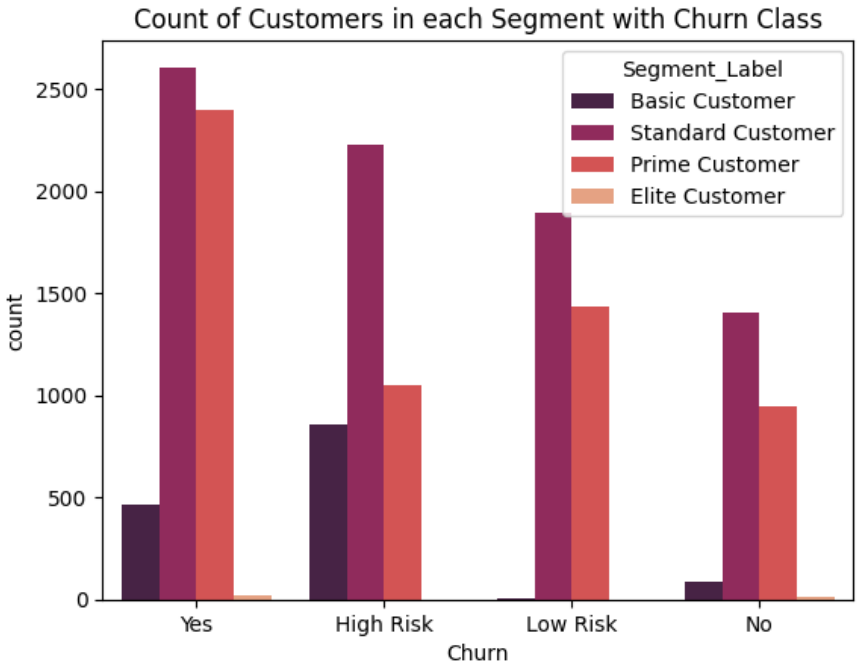


Figure 14:Count of churning Customers by Churn Class

The histogram of "Count of Churning Customers by Churn Class" offers a clear visual representation of where the majority of churn is occurring within different customer segments. Some churn classes have significantly higher counts of churning customers than others notably standard and prime customers, indicating that certain types of categories are more prevalent in our dataset.

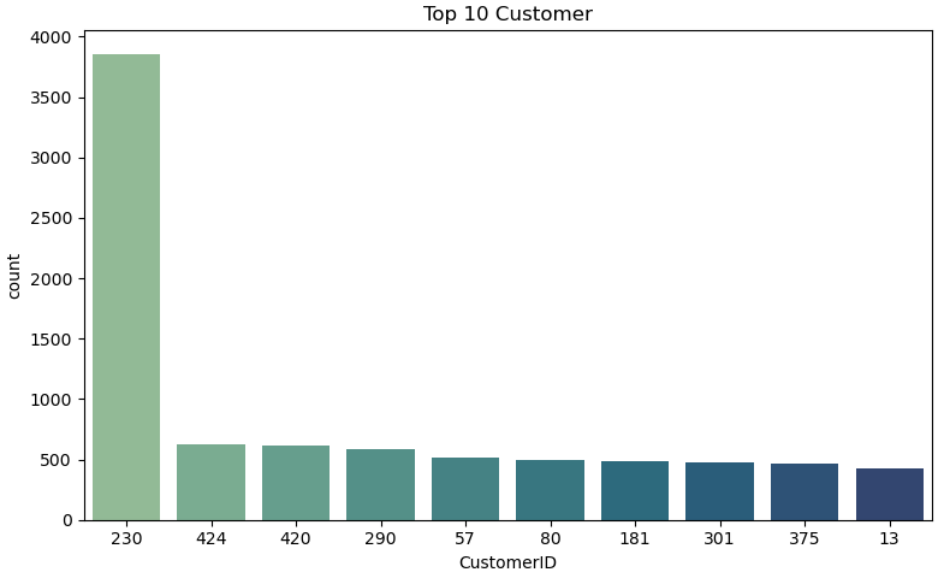


Figure 15:Top 10 customers by Monetary value

The bar chart showcases the top 10 customers ranked by the total revenue each one generates. There's a noticeable gap in revenue contributions among these customers, with one major distribution chain standing out significantly. This key customer contributes a substantial portion of the overall revenue. Understanding what drives the success of these top customers can offer valuable insights to help boost the revenue contributions from other customers as well .

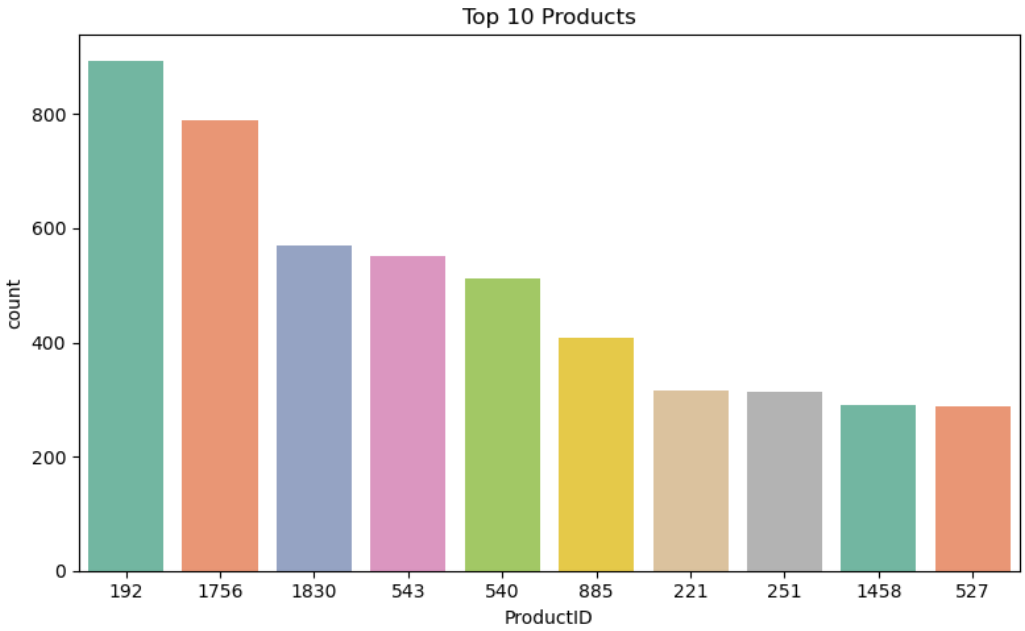


Figure 16:Top 10 products by Monetary value

The bar chart displays the top 10 products based on their monetary value, representing the revenue generated by each product.

Among these products, five stand out significantly, indicating their substantial contribution to the total revenue.

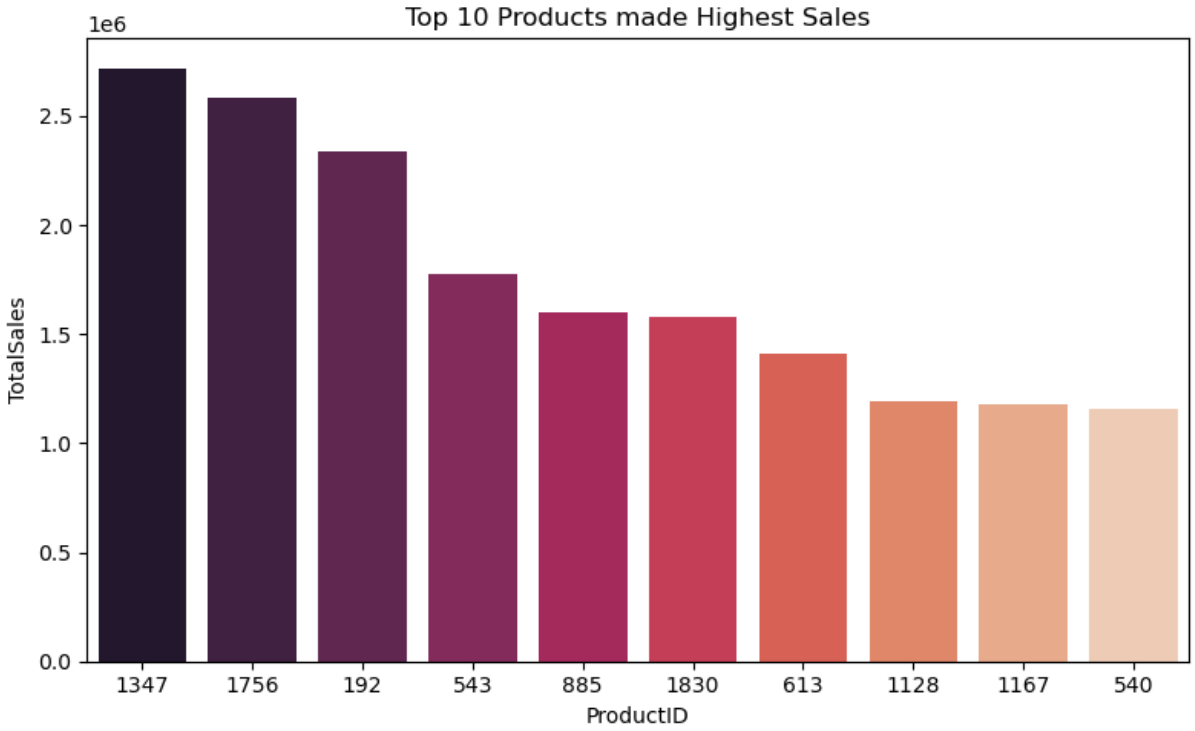


Figure 17:Top 10 products that made the highest value

3.3 EDA on the merged Dataset (RFM/CUSTOMER_DIM):

After merging the two datasets, an exploratory data analysis (EDA) was conducted on the combined dataset consisting of RFM (Recency, Frequency, Monetary) and customer dimensions. This analysis aimed to identify trends, patterns, and correlations within the comprehensive dataset. Various statistical techniques and visualization methods were employed to gain insights into the underlying data structure.

To further deepen our understanding, we also performed bivariate and multivariate analyses. These analyses allowed us to explore the interactions between multiple variables simultaneously, providing a more nuanced view of the factors influencing churn. By examining the relationships between pairs of variables (bivariate analysis) and considering the combined effects of several variables (multivariate analysis), we were able to uncover complex patterns and dependencies that might not be evident through univariate analysis alone. This comprehensive approach enabled us to identify key predictors of churn and develop more accurate and robust models.

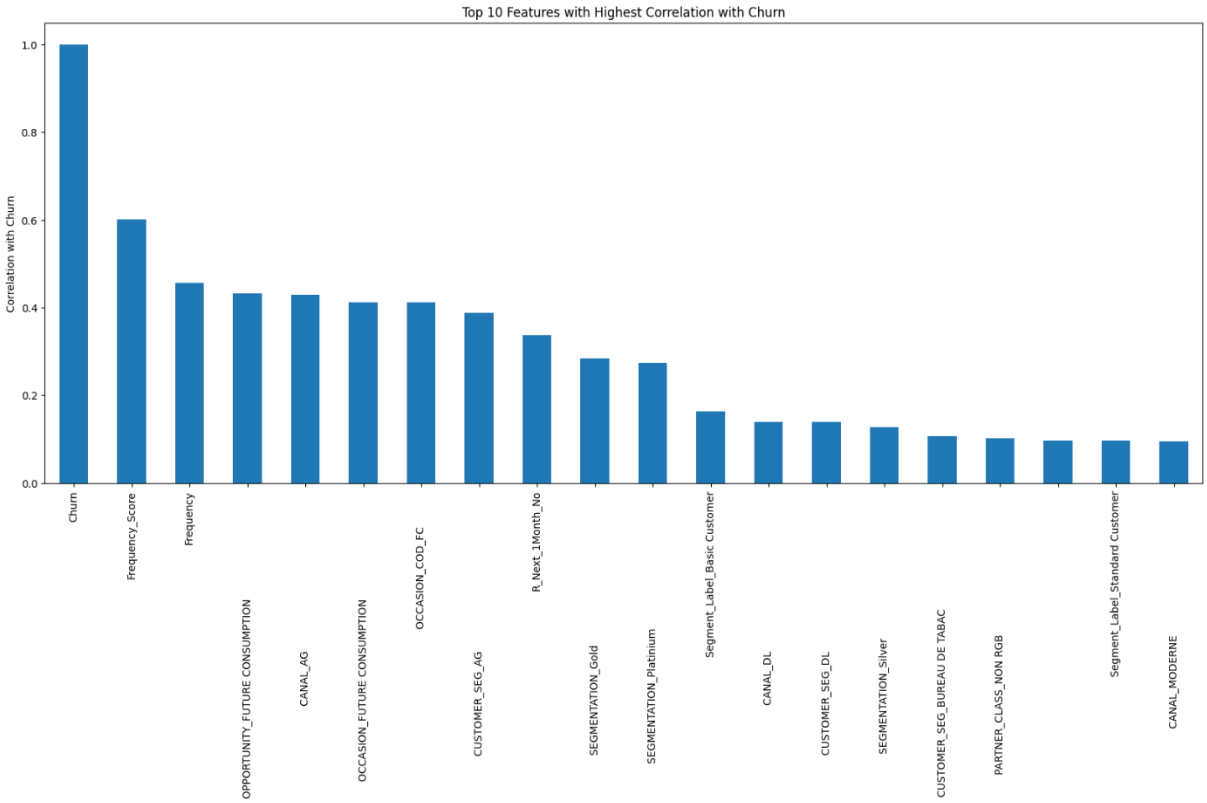


Figure 18:Distribution of correlated features with Churn

The correlations observed between churn and various customer independent variables offer valuable insights into potential factors influencing churn behavior. For instance, the Churn Frequency Score and Frequency metrics suggest that customers with lower engagement levels or less frequent interactions with the company might be at a higher risk of churn. Similarly, the ‘Future Consumption’ value in Opportunity column could indicate the likelihood of affecting churn propensity. The segmentation-related variables, such as Customer Segment and Segment Label, hint at the influence of customer segmentation strategies on churn dynamics, with certain segments possibly exhibiting higher or lower churn rates. Additionally, the channels through which customers engage with the company, represented by Channel - AG and Channel - DL, may play a role in shaping churn tendencies explaining possible Channel logistical and operational issues. Overall, understanding these correlations can empower businesses to tailor retention efforts, enhance customer experiences, and ultimately mitigate churn risks effectively.

4- Machine Learning Model :

4.1- Machine Learning overview:

Machine Learning (ML) is a subset of AI focused on computer algorithms that enhance automatically through experience and data usage. ML models are trained to recognize specific patterns, allowing systems to learn and improve without explicit programming [11]. ML pioneer Arthur Samuel defined it as systems learning from experience to improve performance on specific tasks. ML algorithms are primarily divided into four categories:

1. Supervised Learning: This method trains models using labeled data to establish a mapping function between input and output variables. It's used for classification and regression tasks.

2. Unsupervised Learning: Patterns are inferred from unlabeled input data without supervision. The goal is to find structures and patterns in the data, useful for clustering and association tasks.

3. Semi-supervised Learning: Models leverage a large amount of unlabeled data to assist in learning from a small amount of labeled data.

4. Reinforcement Learning: This area focuses on how agents should take actions in an environment to maximize cumulative rewards. It differs from supervised learning by evaluating actions based on reinforcement signals provided by the environment.

4.2- Classification models :

In pursuit of reliable churn prediction, I employed a range of statistical models tailored to handle the multiclass nature of the task. Acknowledging the necessity for predictive precision across diverse customer segments, I opted for multiclass algorithms capable of discerning intricate patterns within the data. Utilizing techniques such as Decision Trees, Random Forests, and SVM’s I aimed to capture the subtle relationships between predictor variables and the extended set of target categories, including 'Churn', 'Not churn', 'Low Risk', and 'High Risk'. These multiclass algorithms, known for their adaptability and scalability, enabled me to construct predictive models capable of effectively categorizing customers based on their churn likelihood across varying risk levels.

Used Models :

Logistic Regression (LR): Often chosen for its efficiency and ease of interpretation, LR can struggle to capture complex relationships between variables. However, its simplicity makes it a valuable baseline for comparison with other models.

$$\frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}$$

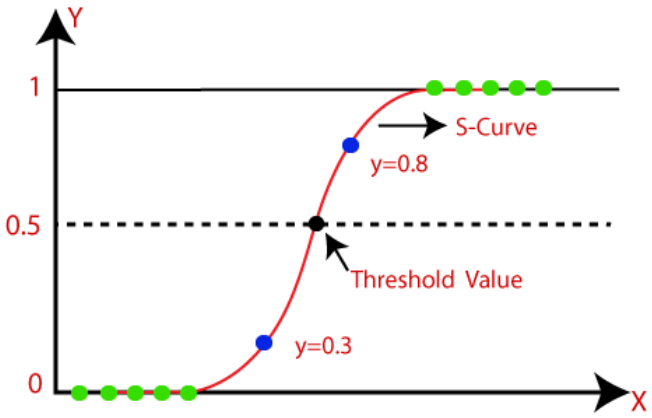


Figure 19 Logistic regression graphic representation

Decision Trees (DT): Frequently employed in churn prediction, DTs offer high interpretability, allowing us to understand the decision-making process of the model. Despite being computationally straightforward, DTs can hold their own against more complex algorithms in certain scenarios.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

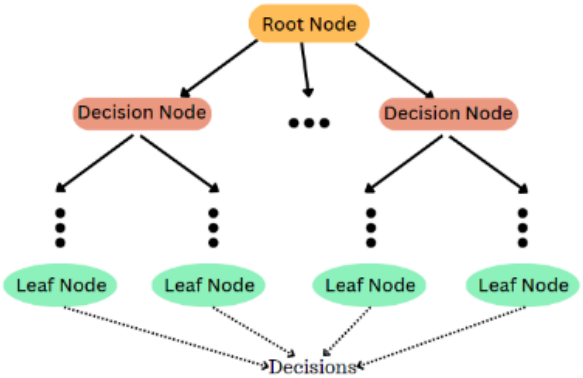


Figure 20 DT graphic representation

Random Forests (RFC): A popular ensemble method in B2C churn prediction, RFCs have received less attention in the B2B context. This research aims to fill this gap by evaluating their effectiveness in B2B churn prediction [1].

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

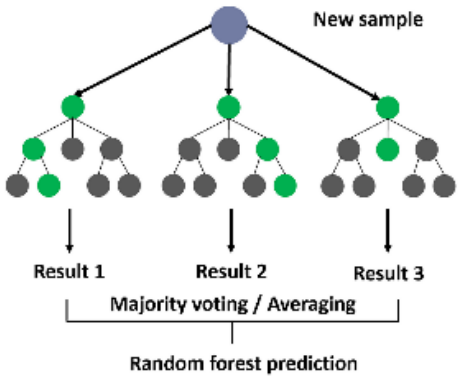


Figure 21 RFC graphic representation

K-Nearest Neighbors (KNN): This non-parametric algorithm classifies new data points based on the similarity to existing data points within a defined neighborhood. While offering ease of implementation, KNN's performance can be sensitive to the chosen distance metric and the number of neighbors considered.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

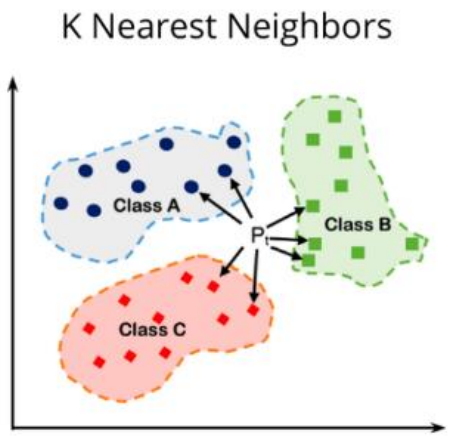


Figure 22 KNN graphic representation

Support Vector Machines (SVM): Offering a powerful alternative to LR for linear classification, SVCs can potentially achieve better performance by considering the margin between classes [11].

$$\begin{aligned}\text{maximize } f(c_1 \dots c_n) &= \sum_{i=1}^{n} c_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i c_i (\varphi(\mathbf{x}_i)\cdot\varphi(\mathbf{x}_j)) y_j c_j \\ &= \sum_{i=1}^{n} c_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i c_i k(\mathbf{x}_i,\mathbf{x}_j) y_j c_j \\ \text{subject to } \sum_{i=1}^{n} c_i y_i &= 0, \text{ and } 0 \le c_i \le \frac{1}{2n\lambda} \text{ for all } i.\end{aligned}$$

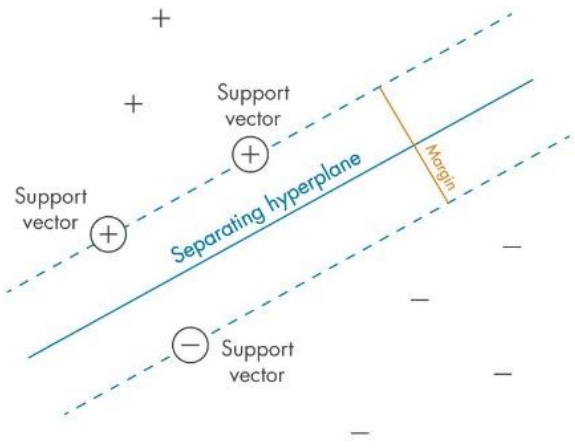


Figure 23 SVM graphic representation

XGBoost: This powerful machine learning algorithm is known for its efficiency and performance. XGBoost, or Extreme Gradient Boosting, builds an ensemble of decision trees in a sequential manner to classify new data points. It leverages gradient boosting techniques to optimize performance and reduce errors. XGBoost is highly effective in handling large datasets and complex data patterns, often outperforming other algorithms in terms of accuracy. However, it requires careful tuning of parameters and can be computationally intensive, especially with large and complex models.

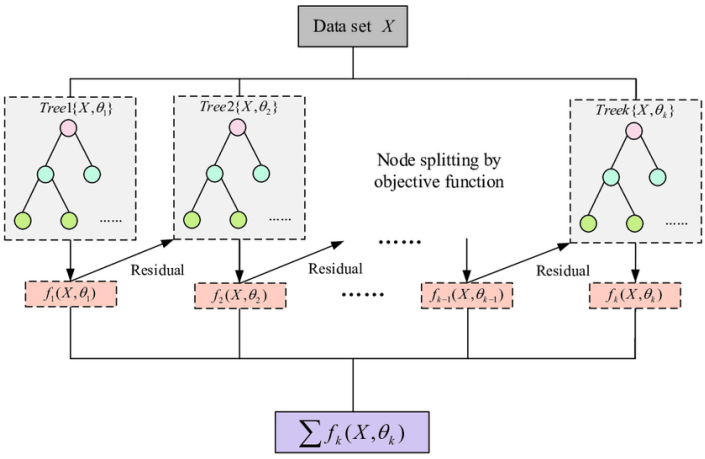


Figure 24 XGBoost

Given the absence of a single dominant algorithm in the literature, we will evaluate the proposed churn prediction methodology across this diverse set of algorithms (**LR, RFC, KNNC, MNB, SVC**) to enhance the generalizability and robustness of our findings. By comparing the performance of these algorithms, we aim to gain a comprehensive understanding of their suitability for B2B churn prediction in the context of the chosen features and evaluation metrics [1].

5- Model Evaluation:

5.1Evaluation Metrics:

The evaluation metric serves as the yardstick for measuring a classifier's performance, crucial for achieving optimal results during classification training. Therefore, selecting an appropriate evaluation metric is key to discerning and attaining the best classifier. In binary classification, evaluating the optimal solution relies on the confusion matrix,, where rows represent predicted classes and columns represent actual classes. From this matrix, the numbers of true positives (tp), true negatives (tn), false positives (fp), and false negatives (fn) can be derived. Commonly used metrics like accuracy, widely utilized in practice for both binary and multi-class classification problems, are derived from the confusion matrix, as demonstrated in previous studies.

When dealing with a multi-class target variable, the confusion matrix expands to accommodate multiple classes. Each cell in the matrix corresponds to the instances of the predicted class versus the actual class, allowing for a more granular analysis of the model's performance across different categories. For instance, in our scenario with four classes, the confusion matrix will have dimensions of 4x4, providing insights into how often each class is correctly predicted and where misclassifications occur.

From this extended confusion matrix, we can derive various evaluation metrics specific to multi-class classification, such as precision, recall, and F1-score for each class. These metrics help in understanding the classifier's performance in distinguishing between multiple classes. Additionally, metrics like macro-average and micro-average precision, recall, and F1-score can be used to provide a summary performance measure across all classes.

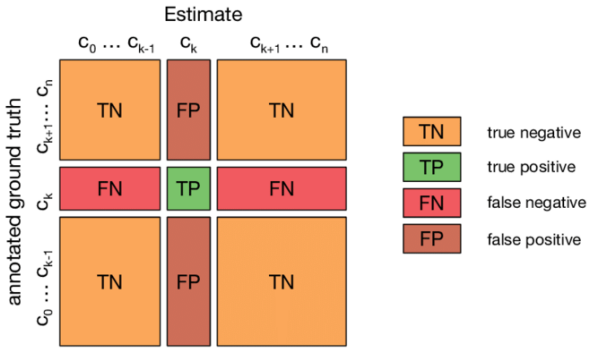


Figure 25 Multi-Class Confusion Matrix schema

Overall, whether for binary or multi-class classification, the confusion matrix remains a crucial tool for evaluating and understanding the effectiveness of the classifier in distinguishing between different classes, thus aiding in fine-tuning and improving the model.

Imbalanced data occurs when one class vastly outweighs the others, creating a skewed distribution within the dataset. This imbalance poses a challenge for machine learning algorithms, as they may exhibit a bias towards the majority class, leading to inaccuracies in model evaluation [11]. For instance, consider a scenario where Class A comprises 90% of the dataset, while Class B represents only 10%. Although identifying instances of Class B is of greater interest, a model that consistently predicts Class A could still achieve a high accuracy rate of 90%. However, this accuracy measure is misleading, as it overlooks the significance of correctly identifying instances of Class B[11] [6].

In practical terms, the costs associated with false positive and false negative predictions are not equal. Therefore, optimizing for basic accuracy alone is insufficient for the intended use case. Instead, a well-calibrated approach should prioritize maximizing the true positive rate, also known as recall. This ensures that instances of the minority class are correctly identified, even at the expense of overall accuracy [14].

Additionally, it's pivotal to outline the strategies employed to tackle these issues effectively. Following the identification of skewed columns and their potential impact on model efficacy, a preprocessing phase involved eliminating these skewed features to counteract bias and enhance the classifier's resilience using pipelines[14]. Subsequently, to address class imbalances, methods like Synthetic Minority Over-sampling Technique (SMOT) were employed [14]. By generating new instances for minority classes, SMOTE helped balance the dataset and mitigate the bias towards majority classes, thereby bolstering the classifier's capacity to generalize and accurately predict minority class instances [14].

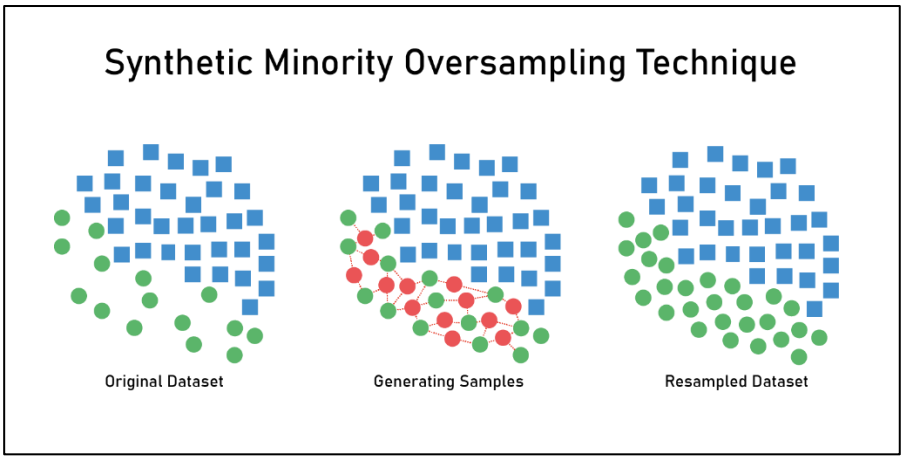


Figure 26:SMOT Sampling technique schema

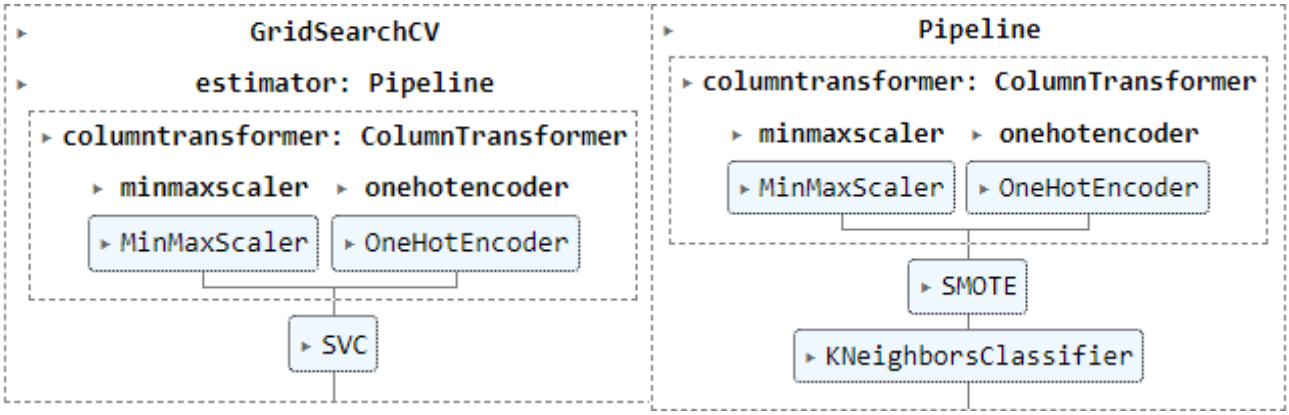


Figure 27: Pipelines examples

5.2 Validation method :

In machine learning, model validation involves evaluating a trained model using a separate validation dataset. This process entails dividing the dataset into three main parts: the train dataset, the validation dataset, and the test dataset [15].

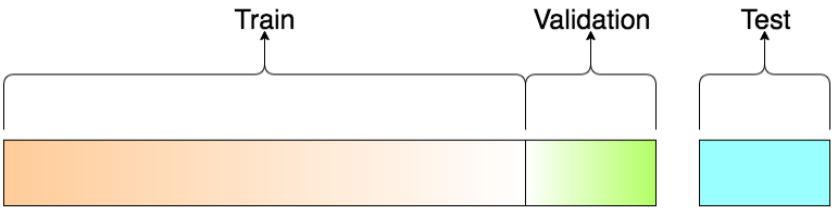


Figure 28:Data Splitting

The train dataset is used for training the model, while the validation dataset assists in fine-tuning model hyperparameters without bias[15]. Meanwhile, the test dataset serves as an impartial evaluation tool for the final model trained on the training dataset. In this study, the dataset was split into two primary segments: train and test. The test set remained separate, while a portion (e.g., 80%) of the train dataset was randomly selected for training, and the remaining portion (20%) was allocated for validation [15]. This iterative process of training and validation was repeated across different splits of the train and validation sets, a technique known as Cross Validation. Such an approach helps prevent overfitting and is increasingly favored, with K-fold Cross Validation emerging as a prominent method in cross-validation practices [15].

A grid search is also conducted using GridSearchCV to tune the hyperparameters of the different models. The grid search explores various combinations of hyperparameters, such as the number of estimators, maximum features, minimum samples per leaf, criterion, and minimum samples for split. This exhaustive search helps identify the optimal combination of hyperparameters that maximizes the model's performance.

5.3 Model Performance Results :

Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy computes the number of correctly classified items out of all classified items.

Recall:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall tells how much the model predicted correctly, out of all the positive classes. It should be as high as possible as high recall indicates the class is correctly recognized (small number of FN). It is usually used when the goal is to limit the number of false negatives.

Precision:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision tells, out of all the positive classes that were predicted correctly, how many are actually positive. High precision indicates an example labeled as positive is actually positive (small number of FP). It is usually used when the goal is to limit the number of false positives. A model with high recall but low precision means that most of the positive examples are correctly recognized (low FN) but that there is a lot of false positive.

F-score:

$$\text{F-score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

F-score is a way to represent precision and recall at the same time and is therefore widely used for measuring model performances. Indeed, if you try to only optimize recall, the algorithm will predict most examples to belong to the positive class, but that will result in many false positives and, hence, low precision. On the other hand, optimizing precision will lead the model to predict very few examples as positive results (the ones with highest probability), but recall will be very low

Actual results :

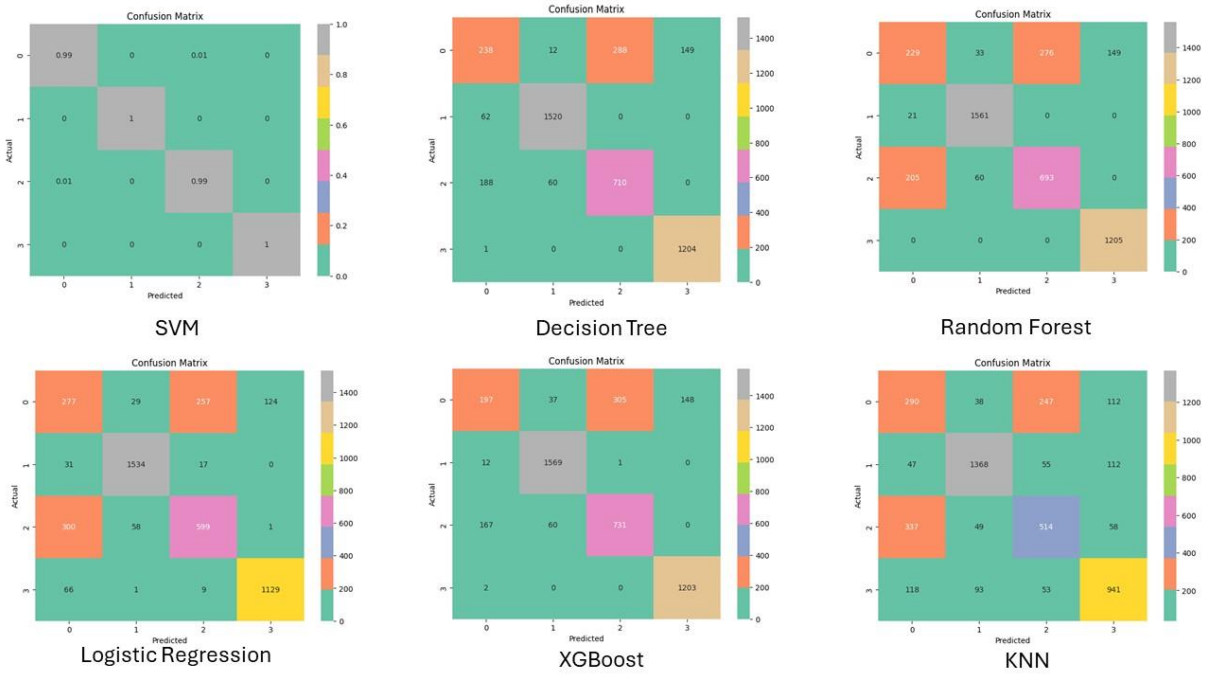
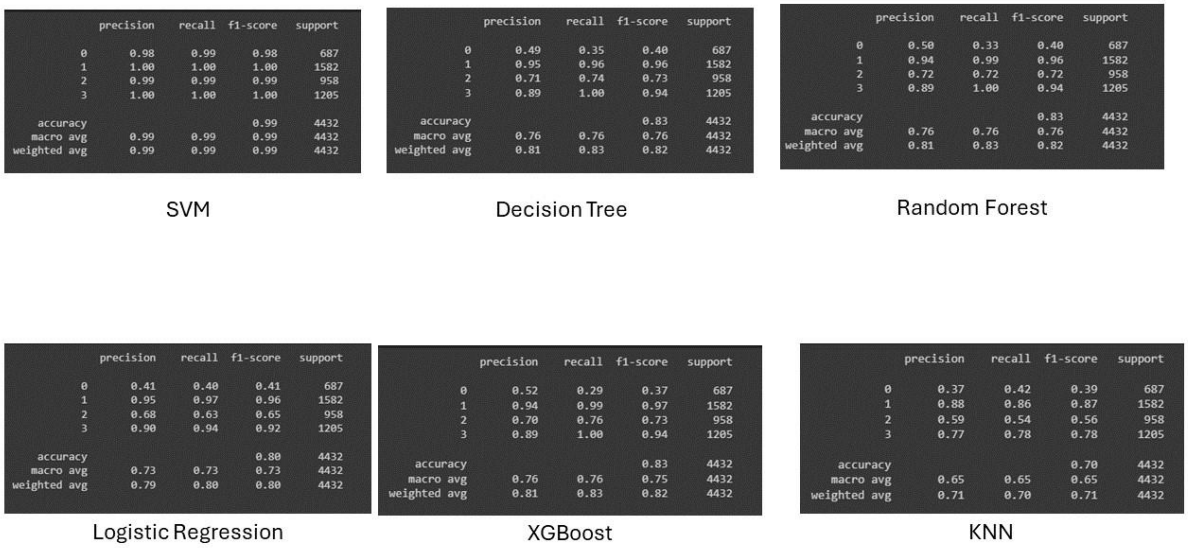


Figure 29:CONFUSION MATRIXES



	precision	recall	f1-score	support
0	0.98	0.99	0.98	687
1	1.00	1.00	1.00	1582
2	0.99	0.99	0.99	958
3	1.00	1.00	1.00	1205
accuracy			0.99	4432
macro avg	0.99	0.99	0.99	4432
weighted avg	0.99	0.99	0.99	4432

SVM

	precision	recall	f1-score	support
0	0.49	0.35	0.40	687
1	0.95	0.96	0.96	1582
2	0.71	0.74	0.73	958
3	0.89	1.00	0.94	1205
accuracy			0.83	4432
macro avg	0.76	0.76	0.76	4432
weighted avg	0.81	0.83	0.82	4432

Decision Tree

	precision	recall	f1-score	support
0	0.50	0.33	0.40	687
1	0.94	0.99	0.96	1582
2	0.72	0.72	0.72	958
3	0.89	1.00	0.94	1205
accuracy			0.83	4432
macro avg	0.76	0.76	0.76	4432
weighted avg	0.81	0.83	0.82	4432

Random Forest

	precision	recall	f1-score	support
0	0.41	0.40	0.41	687
1	0.95	0.97	0.96	1582
2	0.68	0.63	0.65	958
3	0.90	0.94	0.92	1205
accuracy			0.80	4432
macro avg	0.73	0.73	0.73	4432
weighted avg	0.79	0.80	0.80	4432

Logistic Regression

	precision	recall	f1-score	support
0	0.52	0.29	0.37	687
1	0.94	0.99	0.97	1582
2	0.70	0.76	0.73	958
3	0.89	1.00	0.94	1205
accuracy			0.83	4432
macro avg	0.76	0.76	0.75	4432
weighted avg	0.81	0.83	0.82	4432

XGBoost

	precision	recall	f1-score	support
0	0.37	0.42	0.39	687
1	0.88	0.86	0.87	1582
2	0.59	0.54	0.56	958
3	0.77	0.78	0.78	1205
accuracy			0.70	4432
macro avg	0.65	0.65	0.65	4432
weighted avg	0.71	0.70	0.71	4432

KNN

Figure 30: EVALUATION METRICS

Given the dataset's imbalance, it's pivotal to prioritize metrics that address this issue. The recall and F1 Score, which blends precision and recall, is ideal for such scenarios. Among various machine learning models assessed, the SVM classifier with SMOT emerged as highly promising. In customer segmentation, it achieved an outstanding F1 Score of 99 %, highlighting its ability to identify potential churners accurately.

The SVM with SMOT model not only excelled at addressing dataset imbalance but also showcased remarkable predictive accuracy and consistency across various validation sets. Utilizing the Synthetic Minority Over-sampling Technique (SMOT), the model effectively tackled the class imbalance issue, ensuring adequate representation of minority class instances (i.e., potential churners) during training. This strategy significantly improved the model's generalization capabilities and performance on new data.

Given its outstanding performance, the SVM with SMOT model has been chosen as the key predictive tool for detecting customer churn. Its high F1 Score reflects a strong proficiency in accurately identifying true churn instances while minimizing false positives. Consequently, this model will play a vital role in our churn prediction strategy, facilitating proactive measures to retain at-risk customers. By reliably forecasting churn, we can implement targeted interventions to keep valuable customers, thereby enhancing overall customer

satisfaction and reducing attrition rates. This data-driven approach highlights our dedication to utilizing advanced analytics for informed decision-making and sustained business growth.

6- New KPI introduction

The project played a crucial role in introducing fresh Key Performance Indicators (KPI) into the board's dashboard, focusing on churn rate and other relevant indicators. By leveraging insights from the churn prediction model and customer segmentation analyses, these new KPIs offered the board insightful metrics to assess the company's performance and customer retention efforts. This initiative not only enhanced the board's understanding of customer behavior and market trends but also highlighted the organization's dedication to data-driven decision-making. Additionally, the inclusion of churn rate and associated metrics in the dashboard underscored the company's proactive approach to tackling customer churn and optimizing customer lifecycle strategies. According to CustomerGauge, churn rates can significantly differ across various industries, reflecting the unique challenges and customer dynamics within each sector. Typically, highly competitive and subscription-based industries, like telecommunications and media, experience higher churn rates, often exceeding 20%. On the other hand, sectors with greater customer loyalty, such as financial services, generally see lower churn rates, around 10-15%.

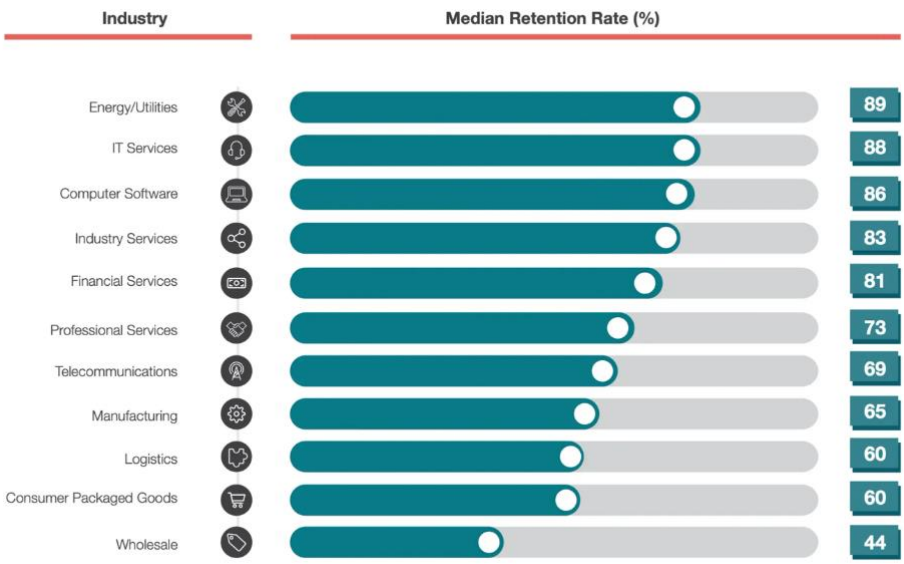


Figure 31:CHURN RATE ACROSS DIFFERENT INDUSTRIES

To calculate churn from retention just subtract the retention rate from 100.

In the food and beverage industry, churn rates also vary depending on the market segment and business model. Research indicates that the average churn rate in this industry ranges from 15% to 25%, influenced by factors like customer satisfaction, brand loyalty, and market competition. Understanding these industry-specific churn trends is essential for businesses to craft effective retention strategies and maintain their competitive position. .

7- Model Deployment and Dashboard interface :

After building the churn prediction model, it was deployed to the rest of the customer base using Azure Machine Learning. This deployment ensures that the model is integrated seamlessly into our operational workflow. To maintain up-to-date insights, the model is scheduled to run on a monthly basis. Each month, it processes the latest transactional data, providing refreshed predictive insights on potential churn. These updated predictions are then incorporated into a Power BI report, allowing us to continuously monitor and respond to changes in customer behavior. This regular execution and integration highlight our commitment to leveraging advanced analytics for proactive customer retention and strategic decision-making.

The final result of the churn prediction project was visually presented using a comprehensive Power BI dashboard. This dashboard offers a detailed and interactive view of the churn analysis, providing key insights into customer behavior and churn risk.

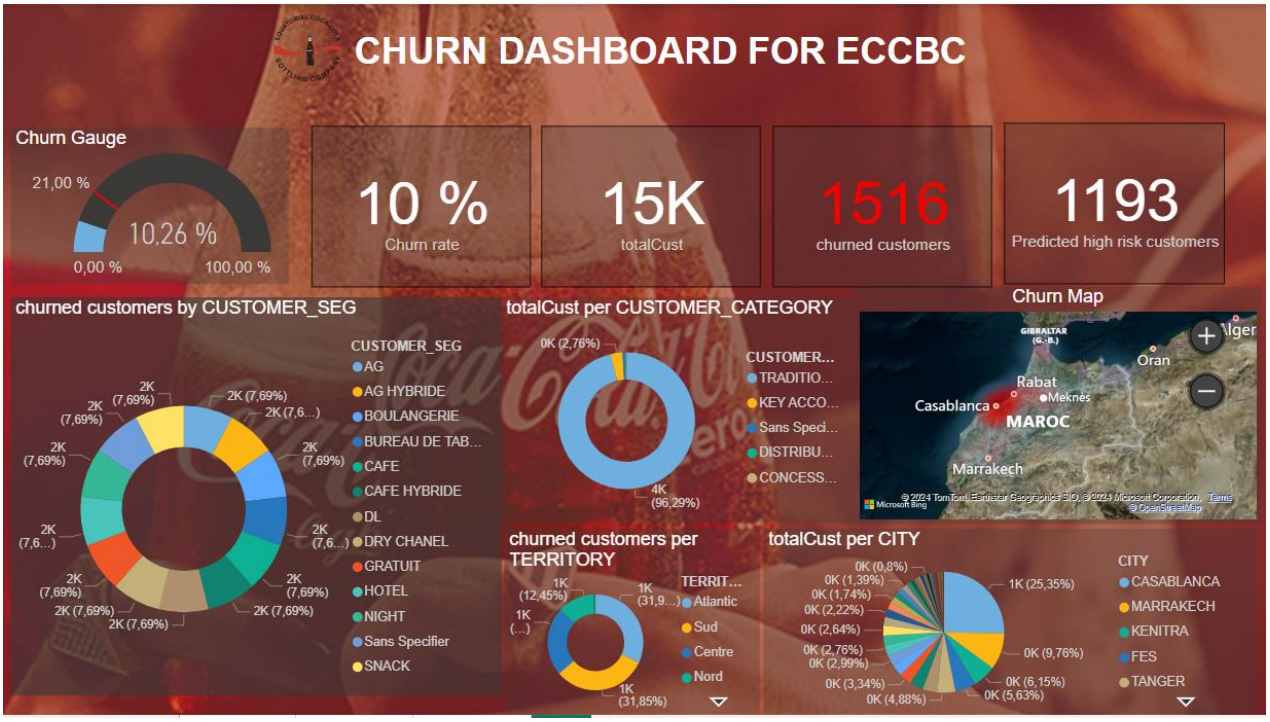


Figure 32:DASHBOARD FIRST PAGE

The first page of the dashboard features several critical visual elements, here is a non-exhaustive list:

Gauge Chart: This chart effectively illustrates the current churn rate in comparison to the permitted churn rate, offering a quick visual reference to determine if the churn rate is within acceptable limits.

Total Number of Customers: A clear and concise count of the total number of customers, giving an overview of the dataset's scale.

Total Number of Churned Customers: This metric highlights the total number of customers who have already churned, providing a clear indication of the churn problem's magnitude.

Number of High-Risk Customers: This figure focuses on customers identified as being at high risk of churning, which is essential for planning targeted retention efforts.

Heat Map of Churners: A geographical representation of where churned customers are located, helping to identify patterns and regional hotspots of customer attrition.

Churn per Segment: A breakdown of churn rates across different customer segments, offering insights into which groups are most at risk and require more focused intervention.

Figure 33:DASHBOARD SECOND PAGE

The second page of this report features a series of pie charts, each offering a unique perspective on customer churn according to various features. These pie charts visually represent the distribution of churn across different customer segments, purchase behaviors, and risk levels. For instance, one chart might display the proportion of churn among different segments, while another highlights the churn rates across various geographical regions. Additionally, pie charts segmented by customer distribution model and Opportunity offer a clear view of how these factors influence churn. This visual representation allows stakeholders to quickly grasp complex data patterns and identify key areas for targeted intervention, ultimately aiding in more effective churn management and customer retention strategies.

Figure 34:DASHBOARD THIRD PAGE

The final page of the Power BI report serves as a comprehensive summary, featuring a detailed table listing all churned customers. This table provides essential information such as customer ID, name, churn risk level and other relevant metrics. To enhance data exploration and usability, the page includes various slicers that allow users to filter the data based on specific criteria, such as customer segment, geographical location, distribution model. These interactive slicers enable stakeholders to drill down into the data and gain focused insights tailored to their needs. Additionally, the page is equipped with an export button, allowing users to seamlessly export the filtered list of churned customers to an Excel file. This functionality facilitates further analysis and action, empowering the team to implement targeted retention strategies effectively.

10. Conclusion and Perspectives

In B2B businesses, each customer holds significant importance compared to those in B2C businesses [1]. This is because the number of B2B customers is generally much lower, but the value of each transaction is considerably higher. Losing even a single customer can have a substantial impact on a B2B product provider. This highlights the critical need for effective customer churn prediction in B2B contexts. However, there is currently a lack of research on how to accomplish this effectively.

Our exploratory study aimed to address this gap by investigating methods for predicting customer churn in B2B settings, while considering the unique characteristics of these businesses. We implemented our approach in a real-world product and developed a two-stage process. This process involves first building a shared data model and then conducting the prediction itself. During the data mapping stage, we created a comprehensive data set that includes both customer and end-user information, organized by customer phases. This data set served as the input for our prediction model.

11. Bibliography

1. *Ali Tamaddoni Jahromi a 1, S. S. (2014, 06). Managing B2B customer churn, retention and profitability. Récupéré sur https://www.sciencedirect.com/science/article/abs/pii/S001985011400114X.*

2. *Bagu, N. (2021, 03). Retail Customer Churn Analysis using RFM. Récupéré sur https://www.ijert.org/research/retail-customer-churn-analysis-using-rfm-model-and-k-means-clustering-IJERTV10IS030170.pdf.*

3. *BLOOMENTHAL, A. (12, 2022). Competitive Intelligence: Definition, Types, and Uses. Retrieved from https://www.investopedia.com/terms/c/competitive-intelligence.asp.*

4. *Equatorial Coca-Cola Bottling Company (ECCBC). (n.d.). Equatorial Coca-Cola Bottling. (s.d.). https://www.eccbc.com/. Récupéré sur eccbc.*

5. *FRANKENFIELD, J. (s.d.). What Is Business Intelligence (BI)? Types, Benefits, and Examples. Récupéré sur https://www.investopedia.com/terms/b/business-intelligence-bi.asp.*

6. *Ghosh, P. (2020). REPORT ON CUSTOMER CHURN PREDICTION USING SUPERVISED MACHINE LEARNING. Récupéré sur https://ghosh-pronay18071997.medium.com/project-report-on-customer-churn-prediction-using-supervised-machine-learning-e5714462f19.*

7. *Howarth, J. (2023, 12). Average Customer Retention By Industry . Récupéré sur https://explodingtopics.com/blog/customer-retention-rates.*

8. *Meredith, D. (s.d.). Customer Churn Prediction. (n.d.). Customer Churn Prediction. . Récupéré sur https://www.investopedia.com/terms/c/customer-churn-prediction.asp.*

9. *Ramirez, J. S. (s.d.). Incorporating usage data for BtoB churn prediction. Récupéré sur https://www.orbel.be/orbel37/ORBEL_37_Booklet_final_160523.pdf#page=25.*

10. *Ramshaw, A. (s.d.). Customer Churn Prediction For B2B and B2C Industries. Récupéré sur https://www.genroe.com/blog/customer-churn-prediction/14714.*

11. *Srigopal, V. (2018). Predicting customer churn risks and optimizing retention investments using reliability and. Récupéré sur https://pure.tue.nl/ws/portalfiles/portal/107234073/Master_Thesis_Vinay_Srigopal.pdf.*

12. *Thibaut, J. (2020, 09). Optimization of a mathematical model for predicting churn and improving cross sell policies. Récupéré sur https://webthesis.biblio.polito.it/secure/19202/1/tesi.pdf.*

13. *Volz, J. (2022, 02). Solving Customer Churn on the Modern Data Stack. Récupéré sur https://medium.com/@jordan_volz/solving-customer-churn-on-the-modern-data-stack-607127cbc33a.*