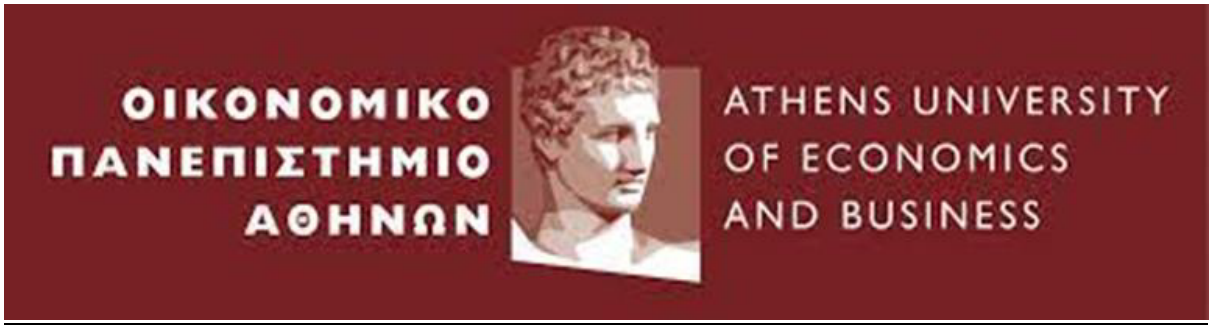

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS

DEPARTMENT OF MANAGEMENT, SCIENCE & TECHNOLOGY

MSc BUSINESS ANALYTICS

Business Analytics Practicum I

Assignment

MSc Students:

ALEXANDRIS ATHANASIOS – p2822202

STROUMPAKIS LAMPROS – p2822223

Professor: ZARAS ANDREAS

ATHENS, 2024

Table of Contents

Table Of Figures

Case Study 1

Executive Summary

Through Market Basket Analysis, we aim to uncover insights into books frequently bought together and recommend effective cross-selling strategies to drive sales growth. We employed Market Basket Analysis to explore customer behavior and identify books frequently bought together. Leveraging the Market Basket Analysis feature of SAS Viya, we analyzed past sales data to uncover insights into customer preferences and purchasing trends.

Utilizing the Market Basket Analysis feature of SAS Viya, we analyzed the sales (in units) of each book of the given dataset. The book 'Data Science and Business Analytics' emerged as the top seller, with 1596 units sold.

We utilized association rules to identify pairs of books with strong associations, recommending effective cross-selling strategies. The association rules revealed the following book pairs with the highest lift values:

Managerial Analytics: Implementing Analytics and Web Analytics 2.0

Implementing Analytics: Data Science and Big Data Analytics, Managerial Analytics

Customer Analytics for Dummies: Decision Analytics and Enterprise Analytics

Enterprise Analytics: Customer Analytics for Dummies and Managerial Analytics

We identified the three books most frequently bought together by customers: 'Data Analytics Made Accessible', 'Data Science and Business Analytics', and 'Business Analytics for Managers'. These books were associated with a count of 794 occurrences, indicating a strong association among them.

Question 2

In the bar chart below(Figure 1) are presented the sales (in units) of each book of the given dataset. We observe that the books with most units sold is the 'Data Science and Business Analytics' in the first place with 1596 units sold, and the 'Data Analytics Made Accessible' in the second with 1210 units sold.

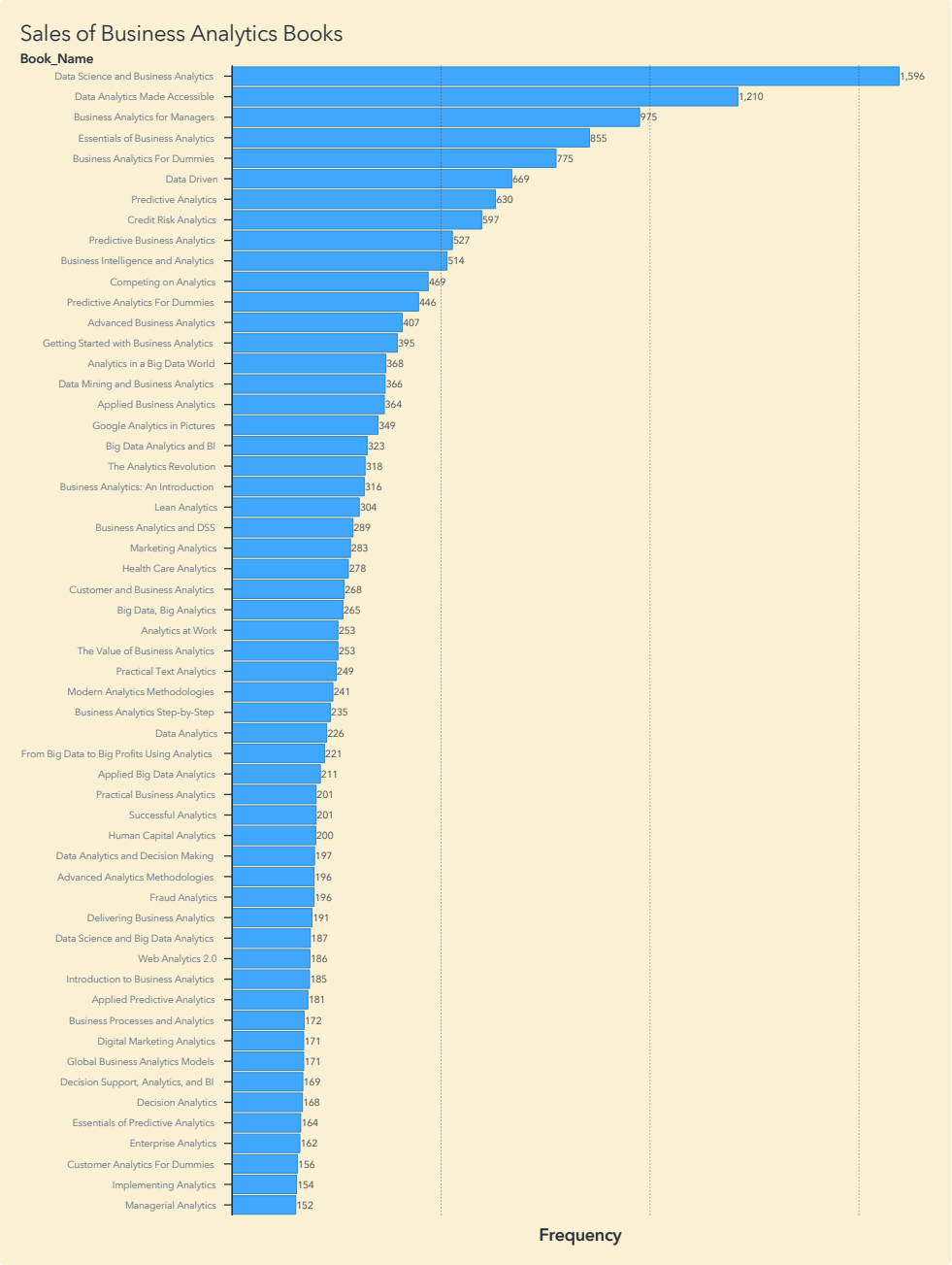

Figure 1 Sales of Business Analytics Books in Units

Question 3

We used the rules table and for each one of the requested books of the question and we performed the following actions. First we filtered the LHS column for value 1 and column RHS for value 2, as we are interested of which two books should the store advertise to customers who bought/ are searching to buy only one of the requested four books. Then we filtered column ITEM1 with each name of the book, and then we sorted in descending order the column LIFT , to identify the association rule with the largest lift for each book on the left-hand side (LHS).

The association rules with the largest lift for each book are listed below:

1. LHS: Managerial Analytics

   RHS: 1) Implementing Analytics , 2)Web Analytics 2.0

   Lift: 11.472147522

2. LHS: Implementing Analytics

   RHS: 1)Data Science and Big Data Analytics, 2)Managerial Analytics

   Lift: 11.330321852

3. LHS: Customer Analytics for Dummies

   RHS: 1)Decision Analytics , 2)Enterprise Analytics

   Lift: 11.192030991

4. LHS: Enterprise Analytics

   RHS: 1)Customer Analytics for Dummies 2)Managerial Analytics

   Lift: 11.073504274

The book pair with the highest lift for each book on the LHS is as follows:

Managerial Analytics: Implementing Analytics and Web Analytics 2.0

Implementing Analytics: Data Science and Big Data Analytics, Managerial Analytics

Customer Analytics for Dummies: Decision Analytics and Enterprise Analytics

Enterprise Analytics: Customer Analytics for Dummies and Managerial Analytics

The lift indicates the strength of association between the books. Higher lift values suggest stronger associations. Therefore, the book pairs identified above are recommended for cross-selling to customers who bought or are interested in buying the respective books on the LHS.

To identify what is the biggest lift of the rules with three items where each one of the above mentioned 4 books is on the left side of the rule, we filtered the ITEM 1 column with the names of those four books. The biggest lift was observed for the following association rule:

LHS: Managerial Analytics

RHS: 1) Implementing Analytics , 2)Web Analytics 2.0

Lift: 11.472147522

Interpretation: A customer who has buys the book 'Managerial Analytics' , is 11.472147522 times more possible to buy also the two books 1) Implementing Analytics 2)Web Analytics 2.0 , than any random customer of the dataset who does not buy 'Managerial Analytics' .

Question 4

To identify which are the 3 books most bought together by customers we performed the following actions. First, we configured the analysis to generate association rules with a maximum of three items per rule. Then, we filtered the dataset to include only rules where all three items were present, by filtering column ITEM3 not to include missing values, ensuring we focused on combinations of exactly three products. Next, we sorted the rules based on the count of occurrences in descending order to pinpoint the most popular combinations.

Among the top-ranked rules, we found that the same three books appeared consistently across different combinations: 'Data Analytics Made Accessible', 'Data Science and Business Analytics', and 'Business Analytics for Managers'. These three books were associated with a count of 794 occurrences, indicating that they were frequently purchased together by customers.

This high count suggests a strong association between these books, implying that customers who bought or showed interest in one book or two books of them were highly likely to purchase the rest one or two as well.

The support metric for this set of three books is 41.88%, which is calculated by dividing the number of transactions containing all three books by the total number of transactions in the dataset. This metric provides insight into how popular this combination is among customers, with a higher percentage indicating a stronger association between the books.

Case Study 2

Executive Summary

Our initiative aimed to harness six years of operational data to enrich customer engagement and retention strategies. The purpose of our analysis was to gain deeper insights into market dynamics and customer behavior, facilitating the segmentation of our customer base. By understanding distinct customer segments, we aimed to recommend tailored actions to maximize profitability and foster long-term customer relationships.

Utilizing Recency Frequency Monetary (RFM) analysis, we processed the data using SAS Viya, ensuring its quality and reliability. Subsequently, we identified five distinct customer segments: High-Spending Churners, First Timers, Churners, Weak Customers, and Best Customers.

These segments were named based on their unique characteristics, providing valuable insights into customer behavior. Further profiling revealed demographic and preference insights, guiding our proposed marketing actions tailored to each segment's specific needs and preferences.

Our proposed actions aim to optimize engagement, retention, and revenue generation, aligning with our overarching goal of maximizing customer satisfaction and lifetime value.

Question 2

Preprocessing

In preparing for the clustering analysis, several preprocessing steps were taken to ensure the quality of the data. Here's a breakdown of the actions taken:

Variable Transformations (logs):

Histogram Analysis: Initially, histograms were used (Figure 2) to assess the skewness of the RFM variable. It was observed that variables R and F exhibited moderate right skewness (0.7016 and 0.5717, respectively), while variable M showed a slight right skewness (0.3661). Given that R and F exceeded the cutoff point of 0.5 for moderate skewness, they were transformed using logarithm functions. Although M's skewness was below the cutoff, it was still transformed to maintain consistency and ensure the desired number of clusters.

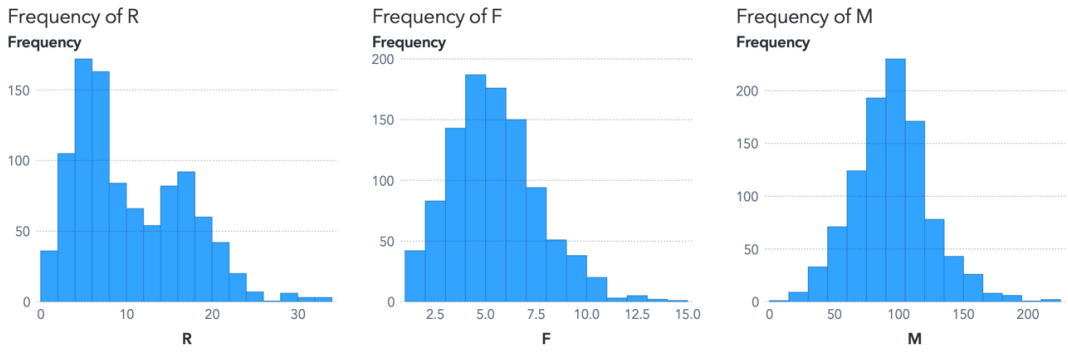

Figure 2 Histograms for RFM Variables

Correlation Analysis:

Correlation Matrix: A correlation matrix (Figure 4) was constructed to examine the relationships between the RFM variables. None of the pairs exhibited strong correlations, with weak to moderate correlations observed. Consequently, no further actions were necessary to address variable correlation.

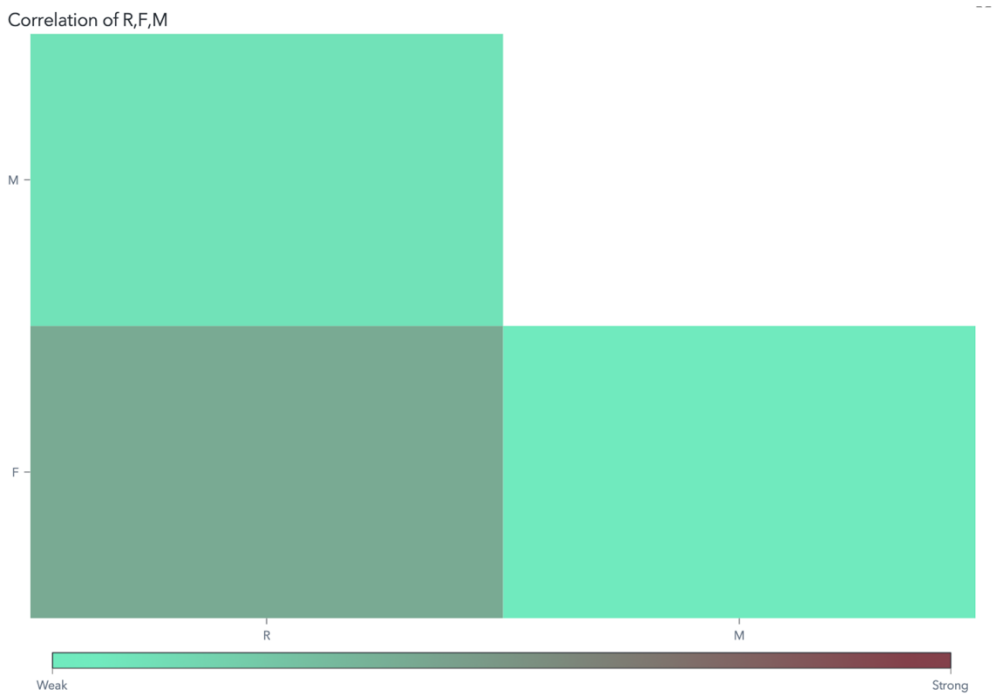

Figure 3 Correlation Matrix for RFM Variables

Outlier Detection and Treatment:

Box Plot Analysis: Box plots were utilized (Figure 4) to identify outliers within each variable. Outliers were identified based on their location outside the upper and lower whiskers of the box plots. For variable R, outliers were observed above the upper whisker (set at 30) and were subsequently filtered out. Similarly, outliers for variable F were identified above the upper whisker (set at 10) and filtered accordingly. For variable M, outliers were observed both above the upper whisker (set at 162.25) and below the lower whisker (set at 27), necessitating filtering for both extremes.

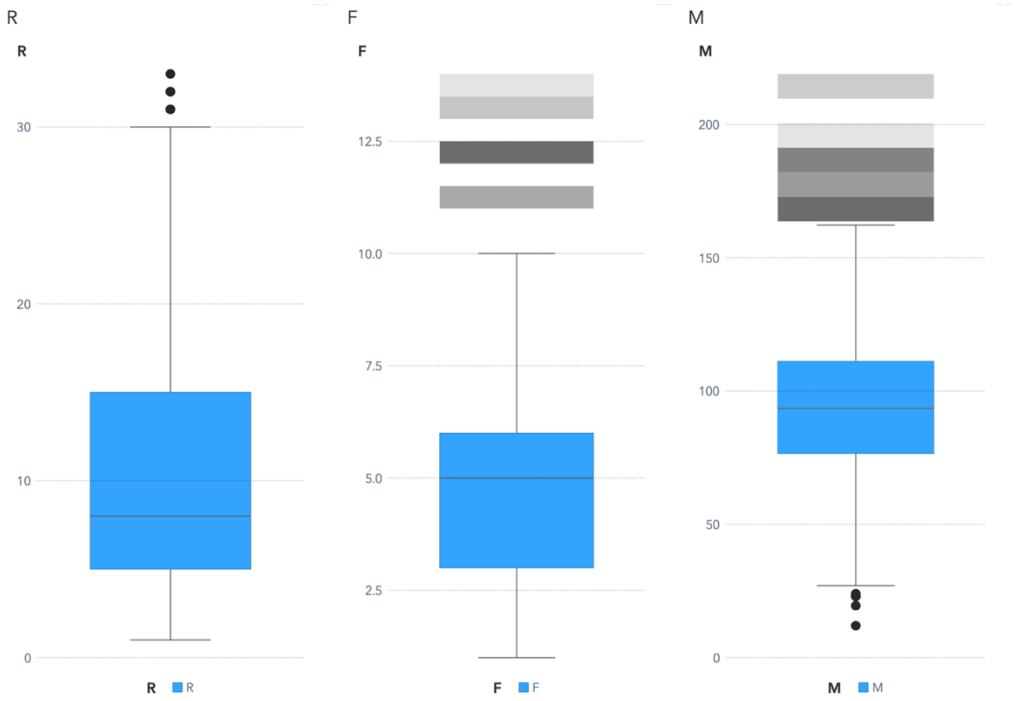

Figure 4 Box Plots for RFM Variables

Handling Missing Values:

Missing Value Check: It was confirmed that there were no missing values present in the dataset. Therefore, no imputation step was required.

In summary, prior to running the clustering analysis, variable transformations were applied to address skewness, outliers were identified and filtered out, and missing values were confirmed to be absent. These preprocessing steps ensured that the data were appropriately prepared for the subsequent clustering analysis, contributing to the reliability and accuracy of the results.

Question 3

After conducting a clustering analysis on the RFM variables, we identified five distinct clusters. Using a tabular format table (Figure 5) , we examined each cluster's average values for the Recency (R), Frequency (F), and Monetary (M) variables. Additionally, we compared these averages to the overall averages for each variable to gain insights into how each cluster differs from the dataset as a whole.

Cluster ID	Segment Names	Frequency	Frequency Percent	R	F	M
1	High Spending Churners	296	30.93%	14.043918919	3.8141891892	115.05286626
2	First Timers	129	13.48%	4.7906976744	3.976744186	66.641454411
3	Churners	153	15.99%	14.68627451	5.2352941176	75.342107584
4	Weak Customers	45	4.70%	15.933333333	1.9777777778	43.312962963
5	Best Customers	334	34.90%	4.8083832335	6.5508982036	99.813032269
Total		957	100.00%	9.7648902821	4.9320794148	93.486284852

Figure 5 Segments Comparison of RFM Variables Averages with Population Totals

Based on the analysis of the five clusters created from the RFM variables, here are the segments identified along with the justification for their names:

High-Spending Churners:

Recency: The average recency for this cluster is significantly higher (14 months) than the total average (9.76 months), indicating that these customers have not interacted with the organization for a long time.

Monetary: Despite their lack of recent transactions, they exhibit the highest average monetary value (115), well above the total average (93.489), suggesting they were once high-spending customers.

Justification: Named as "High-Spending Churners" because they spent significant amounts in the past but have not made recent transactions, indicating churn.

First Timers:

Recency: Customers in this cluster have a much lower average recency (4.79 months) compared to the total average, indicating recent interaction with the organization.

Frequency and Monetary: Both frequency (3.976) and monetary value (66.6) are below the total averages, suggesting these customers are new or infrequent buyers.

Justification: Named as "First Timers" because their recent interaction and lower frequency and monetary values suggest they are new or relatively inexperienced customers.

Churners:

Recency: The average recency for this cluster (14.68 months) is significantly higher than the total average, indicating a long time since their last transaction.

Frequency: Despite having a higher than average frequency (5.235), the high recency suggests they were once frequent buyers but have since churned.

Justification: Named as "Churners" because of their long absence in making transactions despite previously being frequent buyers.

Weak Customers:

Recency, Frequency, Monetary: This cluster has the lowest averages for all three metrics compared to the other clusters and the total averages, indicating poor engagement and spending.

Justification: Named as "Weak Customers" because their metrics suggest they are the least engaged and valuable customers in the dataset.

Best Customers:

Recency, Frequency, Monetary: This cluster has the best averages for all three metrics compared to the other clusters and the total averages, indicating high engagement and spending.

Justification: Named as "Best Customers" because they exhibit the highest levels of engagement and spending, making them the most valuable segment for the organization.

In the Parallel Coordinate Plot below (Figure 6), distinct customer segments emerge based on their behavior. The "Weak Customers" exhibit low values across Recency, Frequency, and Monetary variables, suggesting minimal engagement. Conversely, the "Churners" display high Recency values, moderate Frequency, and lower Monetary values, indicating a propensity to discontinue services. Meanwhile, the "Best Customers" demonstrate low Recency, high Frequency, and high Monetary values, representing valuable, actively engaged clients. Notably, the "High Spending Churners" exhibit high Recency, low Frequency, but high Monetary values, indicating potential churn despite significant spending. Lastly, the "First Timers" showcase high Recency, low Frequency, and low Monetary values, implying a new, yet to be fully engaged customer segment.

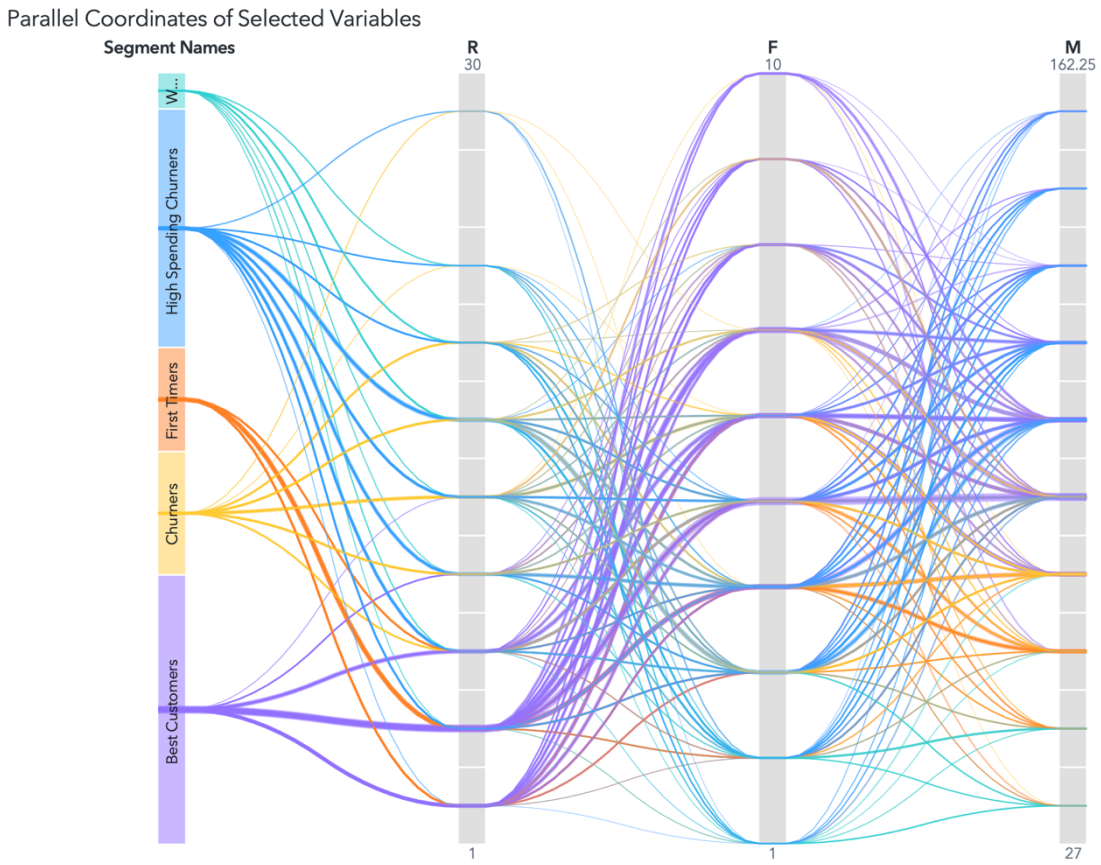

Figure 6 Parallel Coordinates Plot of 5 segments with respect to RFM Variables

Question 4

Expanding on our segmentation analysis, we delve deeper into customer profiling by incorporating additional variables such as age, most frequent payment method, and age range. Through the creation of an interactive dashboard (Figure 7 , Figure 8), we aim to gain comprehensive insights into each segment's demographics and preferences.

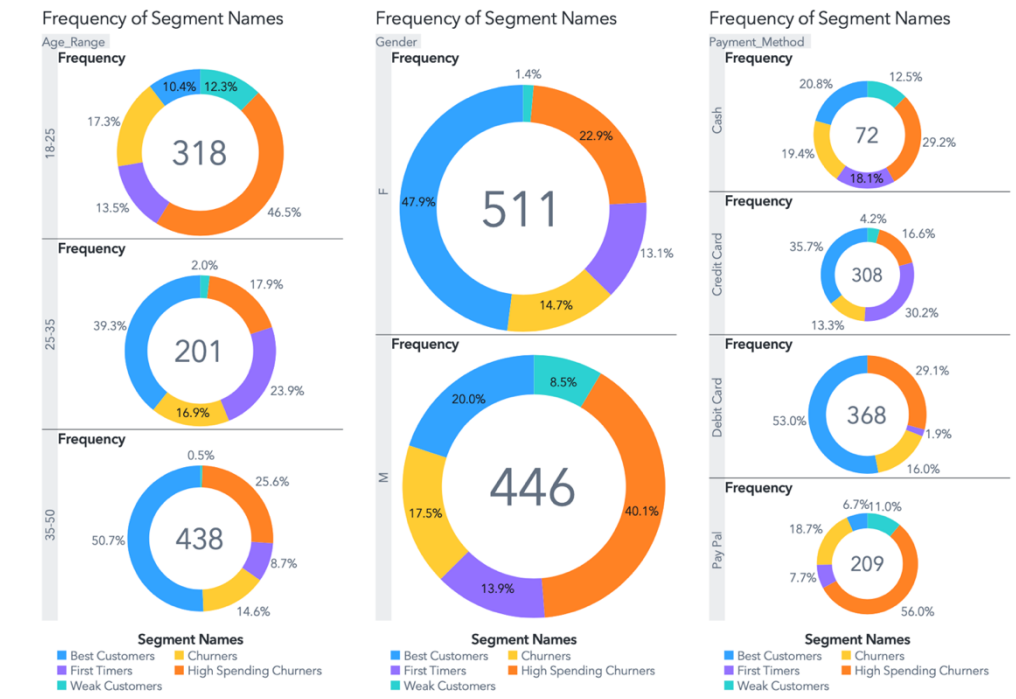

Figure 7 Frequency of Segments over Age, Gender, Payment Method

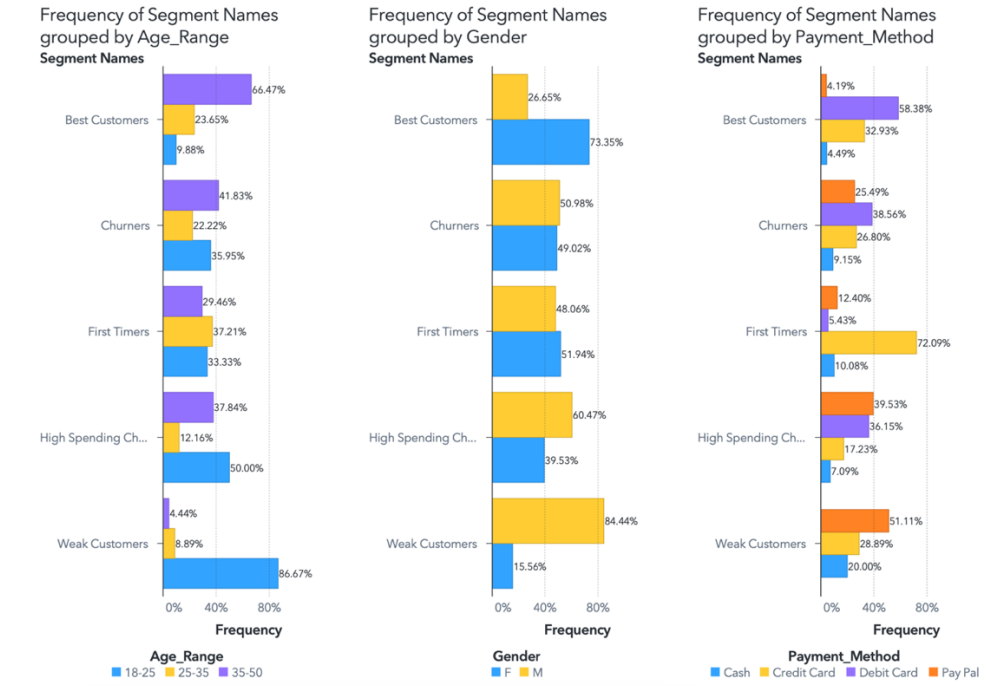

Figure 8 Frequency of Segments Grouped by Age, Gender, Payment Method

Best Customers: Predominantly female customers aged between 35-50, with a strong preference for debit and credit cards. This segment represents the most loyal and valuable customer base.

High Spending Churners: Mixed gender distribution with a notable proportion in the 18-25 age range. Show high spending tendencies and a propensity to churn. Prefer both PayPal and debit cards as the most frequent payment methods.

Churners: Fairly balanced gender distribution across different age demographics, representing customers at risk of discontinuing their relationship with the company. Display varied preferences for payment methods.

First Timers: Balanced gender distribution with a younger demographic profile, particularly concentrated in the 18-25 and 25-35 age ranges. Show openness to digital transactions and new experiences, with an impressive preference for credit cards.

Weak Customers: Predominantly male and primarily composed of customers aged 18-25, exhibiting low engagement and spending behavior. Stand out as the only segment with a significantly increased percentage in using cash, suggesting a preference for offline transactions. This segment also shows a higher-than-average preference for PayPal, indicating a diverse payment behavior.

Question 5

In order to maximize customer engagement and retention, it's crucial for businesses to implement targeted marketing actions tailored to the unique characteristics of each customer segment. Through a thorough analysis of customer behaviors and preferences, we've developed a comprehensive set of marketing proposals aimed at optimizing engagement, retention, and revenue generation. From personalized loyalty programs to targeted reactivation campaigns, these actions are designed to address the specific needs and preferences of each segment, ultimately fostering stronger customer relationships and driving sustainable business growth.

Best Customers:

VIP Events and Exclusive Access

Organize VIP events or offer exclusive access to new products for Best Customers. By providing special privileges, businesses can strengthen their bond with this segment, fostering loyalty and advocacy.

Personalized Thank You Gifts

Send personalized thank you gifts or handwritten notes to express appreciation for their continued patronage. Demonstrating gratitude reinforces positive feelings and encourages repeat purchases.

Early Access Promotions

Offer early access to sales, promotions, or new product launches exclusively for Best Customers. This creates a sense of exclusivity and makes them feel valued, encouraging them to remain loyal to the brand.

High Spending Churners:

Reactivation Discounts

Provide special reactivation discounts or incentives to encourage High Spending Churners to return. Offering discounts on past favorite items or complimentary upgrades can reignite their interest and bring them back to the brand.

Personalized Outreach

Reach out to High Spending Churners with personalized emails or phone calls to understand their reasons for churning. Showing genuine interest and offering assistance can help rebuild trust and re-establish the relationship.

Limited-Time Offers

Create urgency by offering limited-time offers or flash sales exclusively for High Spending Churners. Limited-time promotions capitalize on their previous high spending tendencies and encourage them to act quickly to take advantage of the offer.

Churners:

Win-Back Incentives

Provide win-back incentives such as special discounts, free shipping, or extended warranties to entice Churners to return. Offering value upfront can help overcome any hesitations and convince them to give the brand another chance.

Personalized Apology Campaign

Launch a personalized apology campaign acknowledging past issues or shortcomings and offering solutions to address them. Demonstrating accountability and a commitment to improvement can help rebuild trust and loyalty.

Re-Engagement Surveys

Send targeted surveys to Churners to gather feedback on their experiences and preferences. Understanding their reasons for churning can provide valuable insights for refining products, services, and customer experiences.

First Timers:

Onboarding Welcome Series

Create a series of onboarding welcome emails or tutorials to guide First Timers through the brand's offerings and benefits. Educating them about the brand's value proposition and unique features can increase their confidence and encourage further engagement.

Referral Program

Implement a referral program where First Timers can earn rewards or discounts for referring friends or family. Encouraging word-of-mouth marketing can expand the brand's reach and attract new customers.

Exclusive First-Time Offers

Offer exclusive first-time discounts or promotions to incentivize purchases and encourage repeat business. Providing a positive initial experience increases the likelihood of future purchases and builds loyalty from the start.

Weak Customers:

Personalized Product Recommendations

Provide personalized product recommendations based on Weak Customers' past purchases or browsing history. Tailoring recommendations to their interests increases the likelihood of relevant purchases and improves overall satisfaction.

Customer Support Outreach

Proactively reach out to Weak Customers to offer assistance or address any issues they may be experiencing. Providing exceptional customer support demonstrates care and commitment, potentially turning around their perception of the brand.

Loyalty Program Enrollment

Encourage Weak Customers to enroll in the brand's loyalty program by highlighting the benefits and rewards available. Loyalty program membership provides incentives for repeat purchases and strengthens the bond between customers and the brand.

Case Study 3

Executive Summary

The primary objective of this study is to develop a predictive model to anticipate customer churn, enabling proactive measures by the marketing department to mitigate attrition. Leveraging data from January to June 2017, we first used a profit matrix to determine a cutoff point of 16.667% to identify potential churners. Analysis reveals a historical churner proportion of 14.14% to 85.86%, with no missing values detected.

Subsequently, three distinct machine learning models—Decision Tree (optimal), Logistic Regression, and Neural Network—are deployed to forecast churn probabilities for customers from July to September 2017. Following rigorous evaluation on a validation dataset, the Logistic Regression model emerges as the most effective.

The Logistic Regression model is then applied to unseen data from July to September 2017 to compute churn probabilities. With a cutoff threshold of 16.667%, 496 customers are classified as potential churners out of a total of 1884.

To mitigate churn, the marketing department could consider the following actions such as implement personalized retention offers tailored to individual customer needs, enhance communication channels to provide proactive support and address customer concerns or develop targeted marketing campaigns to re-engage at-risk customers and reinforce brand loyalty.

By proactively addressing potential churners, the marketing department can enhance customer retention and drive long-term profitability.

Question 2

The profit matrix provided suggests the financial implications associated with model predictions and actual outcomes. In this context, a positive value indicates a profit, while a negative value represents a loss. Specifically, predicting a churner and successfully motivating them to stay yields a profit of 1000 monetary units, whereas failing to prevent churn results in a loss of 1500 units. Conversely, correctly identifying a non-churner and taking no action incurs a small loss of 500 units, while mistakenly inter vening with a non-churner leads to no financial impact (0 units). Therefore, the profit matrix underscores the importance of accurate prediction and strategic decision-making to optimize financial outcomes.

Question 3

To determine the cutoff point for identifying churners, we can use the profit matrix provided, which represents the outcomes of different actions taken based on predictions and actual outcomes. The profit matrix is as follows:

		Prediction	
		Churner -- >Motivate Stay	Non-Churner -- > Do Nothing
Actual	Churner	**1000**	**-1500**
	Non-Churner	**-500**	**0**

To find the cutoff point P1, we use the equation:

$$Profit\_Churner * P1 + Profit\_NonChurner * (1 - P1)$$
$$> Loss\_Churner * P1 + Loss\_NonChurner * (1 - P1)$$

Substituting the values from the profit matrix:

$$1000 * P1 - 500 * (1 - \mathrm{P1}) > -1500 * P1 + 0 * (1 - P1)$$

P1 > 1/6

So, the minimum probability P1 for a customer to be considered a churner and hence to be considered for a contact to prevent churn should be greater than 1/6.

Question 4

The partitioning is essential to train the model on a subset of the data while retaining a separate portion for evaluation, ensuring an accurate assessment of its predictive performance on unseen data.

Question 5

As we observe by looking at the 'Missing' Column of Data option in of the Project there are no Missing Values for all Variables, that we are interested in.

Enable stratified sampling to maintain the proportions of churners and non-churners in both the training and validation datasets, ensuring unbiased model evaluation.

1) The proportion of churners and non-churners in the dataset is as follows:

   Non-Churners: 85.86%

   Churners: 14.4%

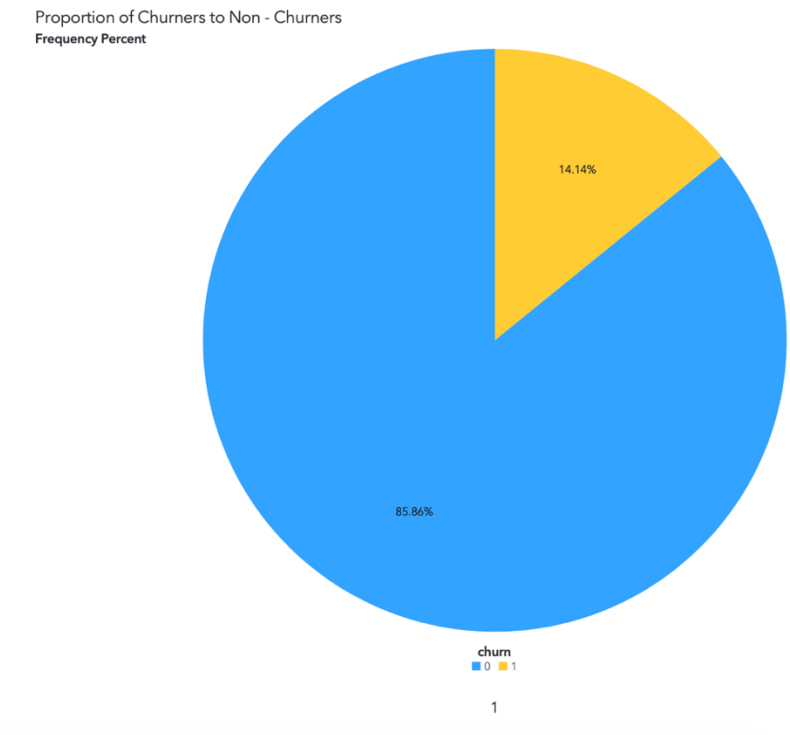

Figure 9 Pie Chart, Proportion of Churners to Non - Churners

Question 6

If the proportion of churners to non-churners in the historical dataset shifted to 3% churners and 97% non-churners, it would indicate a significant class imbalance. With churners constituting less than 10% of the dataset, it falls into the category of rare events. In such scenarios, traditional machine learning algorithms often struggle to effectively learn patterns from the minority class (churners) due to the dominance of the majority class (non-churners).

To address this imbalance, one effective strategy is to employ resampling techniques such as oversampling the minority class (churners) or undersampling the majority class (non churners) to balance the class distribution.

Question 7

The proportion of chunrers and non- chunrers for those customers that have contacted more than 6 times the customer service is as follows:

Non-Churners: 41.18%

Churners: 58.82%

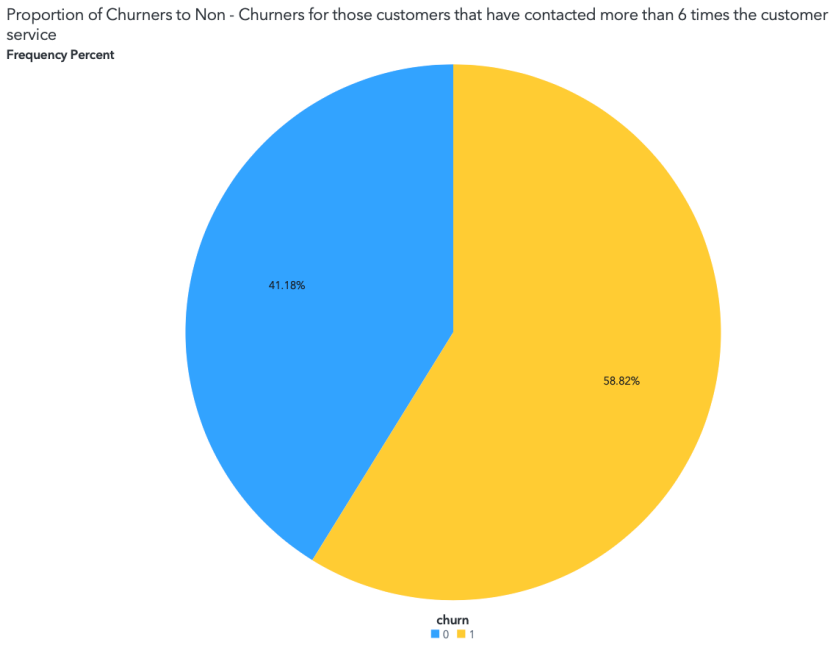

Figure 10 Pie Chart, Proportion of Churners to Non - Churners for customers that have contacted more than 6 times the Customer Service

We observe that filtering the customers that have contacted more than 6 times the customer service, the percentage of churners is higly increased and it is slighlty bigger from the percentage of non churners. This insight is meaningfull since, customers who contact so many times customer service they have faced many problems, and this could drive them to churn.

Question 8

Churners have an average of 1,898 minutes, while non-churners have an average of 1,606 minutes. This finding suggests a potential correlation between the duration of phone calls during the day and the likelihood of churn. Specifically, churners appear to spend more time on phone calls compared to non-churners. This could imply that higher usage or engagement, as reflected in longer call durations, may be associated with a higher probability of churn.

Question 9

The decision tree analysis using the CHAID algorithm revealed that the variable utilized for the first split is "Intl_Plan," which indicates whether a customer has an International plan. This variable was selected as the initial split due to its significance in differentiating the dataset. In the context of our analysis, customers are directed to the left node if they have an International plan (Intl_Plan = Yes) and to the right node if they do not have an International plan (Intl_Plan = No).

To gauge the importance of variables, one method involves examining their logworth statistic, which represents the negative logarithm of their p-value. The decision tree algorithm strategically selects variables with higher logworth values, indicating their significant predictive capability in segmenting the dataset effectively. The selection of the "Intl_Plan" variable for the initial split underscores its considerable influence, as reflected in its logworth score.

Regarding missing values, it's noteworthy that in the first split of the decision tree (corresponding to the "Intl_Plan" variable), there is no specific handling mentioned for missing values. However, as observed in subsequent nodes related to numerical variables, missing values seem to be directed to one of the branches based on certain criteria. This suggests that the decision tree algorithm handles missing values differently for numerical variables, possibly considering them as a separate category or directing them to a specific branch based on predefined thresholds or rules.

Question 10

The second decision tree node in the workspace is named "Maximal tree" and is configured to utilize the largest (maximal) method for pruning. After running the tree node, it was observed that the tree consists of 22 terminal leaves. This tree is referred to as the maximal tree because it maximizes the number of terminal leaves without pruning.

The performance of the training and validation datasets was evaluated using the Misclassification Rate as the assessment criterion. The subtree assessment plot illustrates how the misclassification rate changes as the decision tree is pruned to different numbers of leaves. In this plot, the phenomenon presented in the blue line for the training dataset shows a decreasing trend in the misclassification rate as the number of leaves increases. This phenomenon is commonly known as overfitting, where the model learns to perform well on the training data but fails to generalize to unseen data.

To address overfitting, the solution involves pruning the decision tree based on the performance of the validation dataset. By selecting a subtree with an optimal number of leaves that minimizes the misclassification rate for the validation partition, we ensure that the model achieves better generalization performance. In this case, the selected subtree based on the pruning options has 18 leaves with a misclassification rate of 0.138 for the validation partition.

A screenshot of the largest tree is provided in the report (Figure 11), showcasing the structure and complexity of the decision tree model.

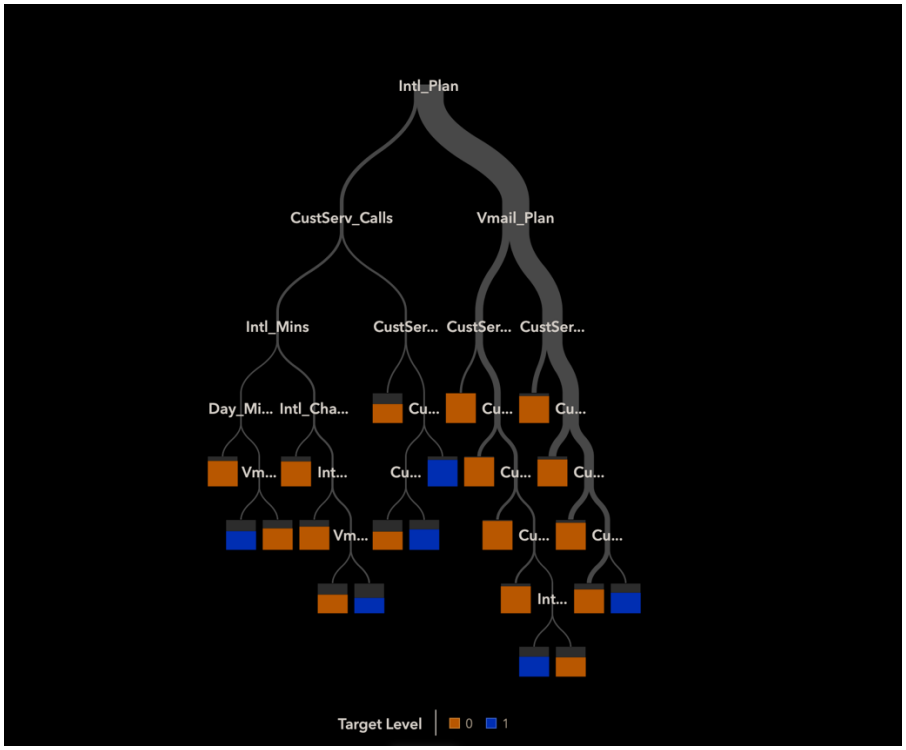

Figure 11 Maximal Tree

Question 11

After running the first decision tree node, it was determined that the optimal tree has 18 terminal leaves. Below, you can find a screenshot of the optimal tree (Figure 12):

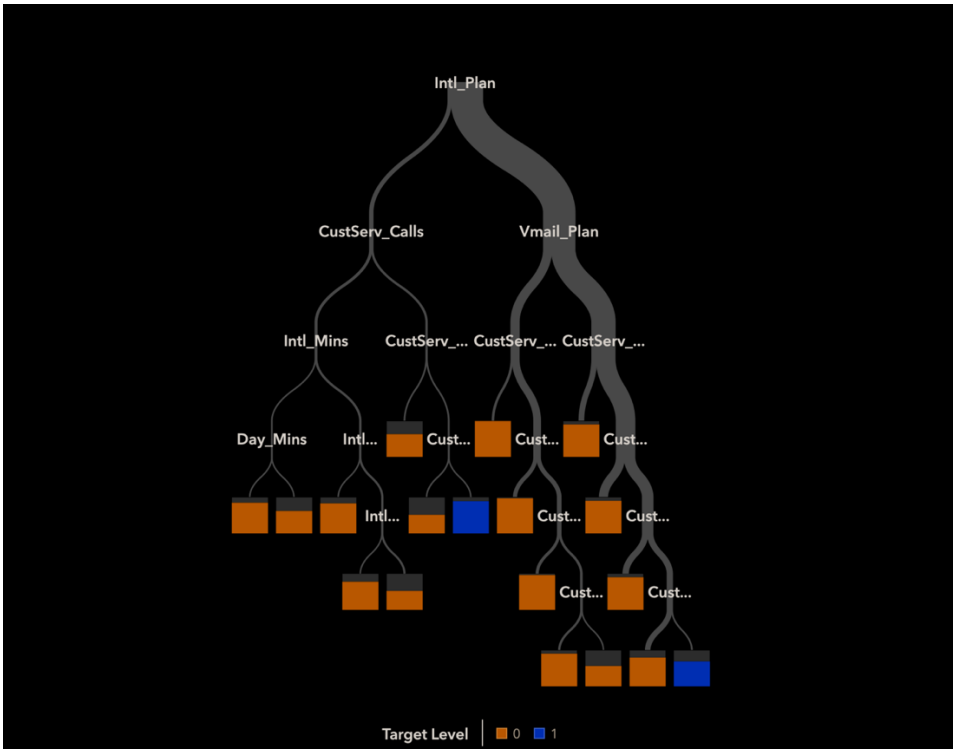

Figure 12 Optimal Tree

Additionally, a screenshot of the subtree assessment plot is provided (Figure 13):

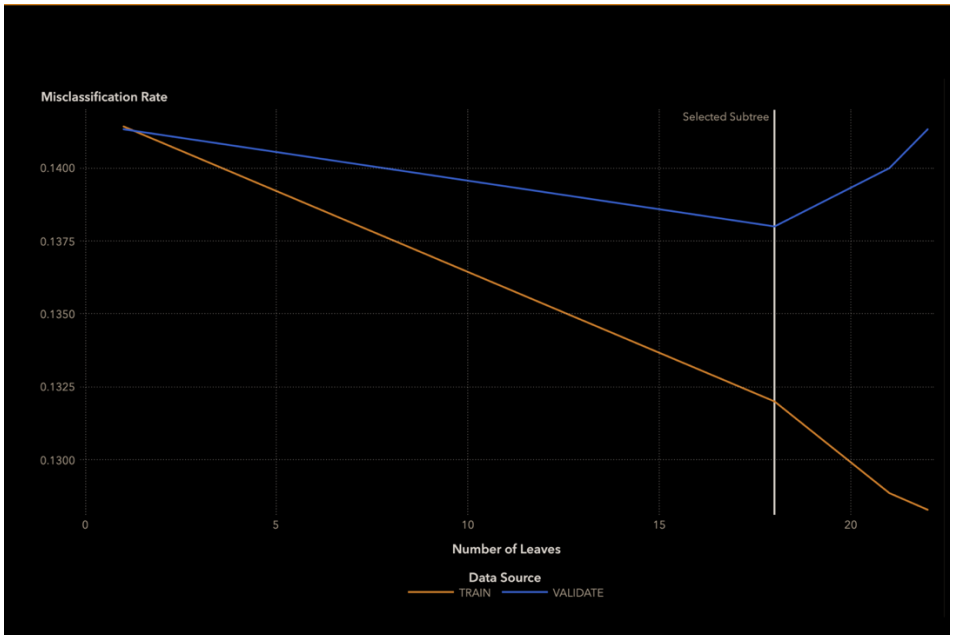

Figure 13 Optimal Tree Assessment Plot

The subtree assessment plot illustrates how the misclassification rate changes as the decision tree is pruned to different numbers of leaves. As depicted in the plot, the training error decreases as the number of leaves increases, indicating that the model performs better on the training data with a larger number of leaves. However, it's crucial to consider the performance on the validation dataset to prevent overfitting. In this case, the selected subtree based on the pruning options has 18 leaves with a misclassification rate of 0.138 for the validation partition, suggesting a balance between model complexity and generalization performance.

Question 12

The decision tree model constructed from our telecom customer data reveals several key insights into factors influencing churn. At its core, the tree represents a hierarchical structure of decision rules based on various customer attributes. Each node in the tree represents a decision point where the dataset is split based on a specific feature or attribute. The splits are determined by identifying the most predictive variables that effectively differentiate between churners and non-churners.

For instance, the initial split is based on whether a customer has an international plan. Subsequent splits further segment the data based on additional features such as customer service calls, voicemail plan, international minutes, and day minutes. These splits lead to terminal leaves where the model makes predictions about whether a customer is likely to churn or not.

Interpretation of five terminal leaves:

Terminal Leaf 1 (Node ID 11): Customers who do not have an international plan (Intl_Plan = no) AND have a voice mail plan (Vmail_Plan = yes) AND have fewer than 1 customer service call or the value of customer service calls for this customer is missing are predicted to have 0.66% probability of being churners and since it is below than the cutoff 16.667% we decide not to take any actions for this segment of customers.

Terminal Leaf 2 (Node ID 13): Customers who do not have an international plan (Intl_Plan = no) AND do not have a voice mail plan (Vmail_Plan = no) AND have one or more customer service calls (CustServ_Calls >= 1) are predicted to have 10.06% probability of being churners and since it is below than the cutoff 16.667% we decide not to take any actions for this segment of customers.

Terminal Leaf 3 (Node ID 21): Customers who do not have an international plan (Intl_Plan = no) AND have a voice mail plan (Vmail_Plan = yes) AND have one or more customer service calls (CustServ_Calls

>= 1) AND have less than two customer service calls (CustServ_Calls<2), so the value of customers calls should be exactly one, are predicted to have 2.83% probability of being churners and since it is below than the cutoff 16.667% we decide not to take any actions for this segment of customers.

Terminal Leaf 4 (Node ID 26): Customers who do not have an international plan (Intl_Plan = yes) AND have less than two customer service calls (CustServ_Calls <2) AND have 86 or more minutes in International calls (Intl_Mins >= 86), AND the amount that the customer paid for the international calls is 24.6 $ or more(Intl_Charge >= 24.6) AND have 95 or more minutes in International calls (Intl_Mins >= 95), are predicted to have 47.27% probability of being churners and since it is above than the cutoff 16.667% we decide to take any actions for this segment of customers.

Terminal Leaf 5 (Node ID 34): Customers who do not have an international plan (Intl_Plan = no) AND AND do not have a voice mail plan (Vmail_Plan = no), AND have 5 or more customer service calls (CustServ_Calls >= 5) are predicted to have 68.57% probability of being churners and since it is above than the cutoff 16.667% we decide to take any actions for this segment of customers.

Question 13

The decision tree analysis unveils critical insights into customer churn dynamics, shedding light on the most influential variables that determine whether a customer is likely to churn or not. Key variables such as international plan status, voicemail subscription, and customer service interactions emerge as pivotal factors in predicting churn.

Terminal Leaf 1: Customers without an international plan but with a voicemail plan and minimal customer service calls exhibit a mere 0.66% probability of churning.

Terminal Leaf 2: Similar to the first group, customers without an international plan and voicemail, but with some customer service calls, display a slightly higher churn probability of 10.06%.

Terminal Leaf 3: Customers without an international plan, with a voicemail plan, and a specific range of customer service calls demonstrate a marginal churn probability of 2.83%.

Terminal Leaf 4: Conversely, customers with an international plan, specific call durations, and corresponding charges exhibit a significantly elevated churn probability of 47.27%.

Terminal Leaf 5: Finally, customers without international or voicemail plans but with a high volume of customer service calls represent a substantial churn risk, with a probability of 68.57%.

Understanding these influential variables enables us to tailor retention strategies effectively, engaging customers and mitigating churn risks proactively.

Question 14

Maximal Tree

If we choose the 20% of most highly ranked customers to be churners according to the probability that the best model gives them to be churners, the 32.589% of this 20% will be churners.

If we choose the 100% of most highly ranked customers to be churners according to the probability that the best model gives them to be churners, the 14.133% of this 100% will be churners.

Optimal Tree

If we choose the 20% of most highly ranked customers to be churners according to the probability that the best model gives them to be churners, the 32.589% of this 20% will be churners.

If we choose the 100% of most highly ranked customers to be churners according to the probability that the best model gives them to be churners, the 14.133% of this 100% will be churners.

Logistic Regression

If we choose the 20% of most highly ranked customers to be churners according to the probability that the best model gives them to be churners, the 35% of this 20% will be churners.

If we choose the 100% of most highly ranked customers to be churners according to the probability that the best model gives them to be churners, the 14.133% of this 100% will be churners.

Neural Network

If we choose the 20% of most highly ranked customers to be churners according to the probability that the best model gives them to be churners, the 35% of this 20% will be churners.

If we choose the 100% of most highly ranked customers to be churners according to the probability that the best model gives them to be churners, the 23.333% of this 100% will be churners.

Question 15

The % response chart for the validation data set is constructed by sorting the dataset based on predicted response probabilities. The x-axis represents the cumulative percentage of the total population, while the y-axis represents the cumulative percentage of responders as we move through the sorted list.

Maximal Tree

If we choose the fifth bucket (20%-25%) of most highly ranked customers to be churners according to the probability that the best model gives them to be churners, the 16.489% of this bucket will be churners.

Optimal Tree

If we choose the fifth bucket (20%-25%) of most highly ranked customers to be churners according to the probability that the best model gives them to be churners, the 16.489% of this bucket will be churners.

Logistic Regression

If we choose the fifth bucket (20%-25%) of most highly ranked customers to be churners according to the probability that the best model gives them to be churners, the 21.333% of this bucket will be churners.

Neural Network

If we choose the fifth bucket (20%-25%) of most highly ranked customers to be churners according to the probability that the best model gives them to be churners, the 12% of this bucket will be churners.

Question 16

The cumulative lift chart for the validation data set illustrates how much better the model performs compared to random selection, as we target a certain percentage of the population ranked by their predicted response probabilities. At the 20% point on the x-axis, the chart shows how many times the model lifts the response rate compared to if we were to select customers randomly.

Maximal Tree

If we choose the 20% of most highly ranked customers to be churners according to the probability that the best model gives them to be churners, we will capture 2.3058 times more churners than if we did the same job without a model i.e. at random.

Optimal Tree

If we choose the 20% of most highly ranked customers to be churners according to the probability that the best model gives them to be churners, we will capture 2.3058 times more churners than if we did the same job without a model i.e. at random.

Logistic Regression

If we choose the 20% of most highly ranked customers to be churners according to the probability that the best model gives them to be churners, we will capture 2.4764 times more churners than if we did the same job without a model i.e. at random.

Neural Network

If we choose the 20% of most highly ranked customers to be churners according to the probability that the best model gives them to be churners, we will capture 1.6509 times more churners than if we did the same job without a model i.e. at random.

Question 17

The cumulative % captured response graph for the validation dataset depicts the proportion of responses captured as we target a certain percentage of the population based on their predicted response probabilities. At the 40% point on the x-axis, the graph indicates the cumulative percentage of actual responses obtained when targeting the top 40% of customers according to the model's predictions.

Maximal Tree

If we choose the 40% of most highly ranked customers to be churners according to the probability that the best model gives them to be churners, we will capture the 68.021% of all the responders of the whole validation data set.

Optimal Tree

If we choose the 40% of most highly ranked customers to be churners according to the probability that the best model gives them to be churners, we will capture the 68.021% of all the responders of the whole validation data set.

Logistic Regression

If we choose the 40% of most highly ranked customers to be churners according to the probability that the best model gives them to be churners, we will capture the 76.415% of all the responders of the whole validation data set.

Neural Network

If we choose the 40% of most highly ranked customers to be churners according to the probability that the best model gives them to be churners, we will capture the 58.491% of all the responders of the whole validation data set.

Question 18

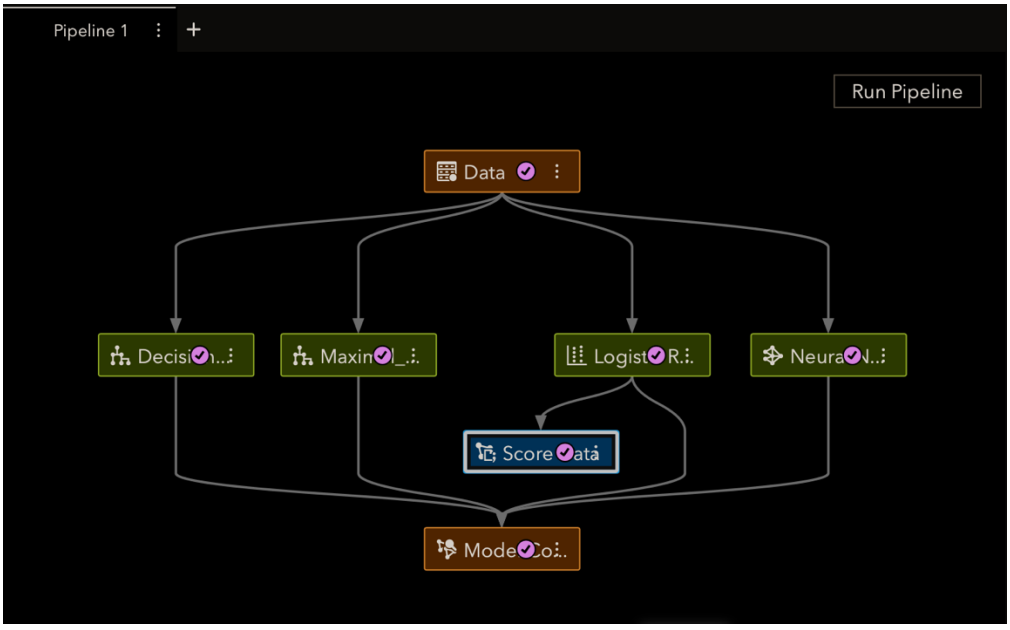

Figure 14 Model's Completed Process Flow

As observed in the following bar chart (Figure 15) generated using SAS Visual Analytics, there are 496 predicted churners and 1388 predicted non-churners in the "telco_data_apr_sep" dataset. This indicates a total of 1884 customers in the dataset.

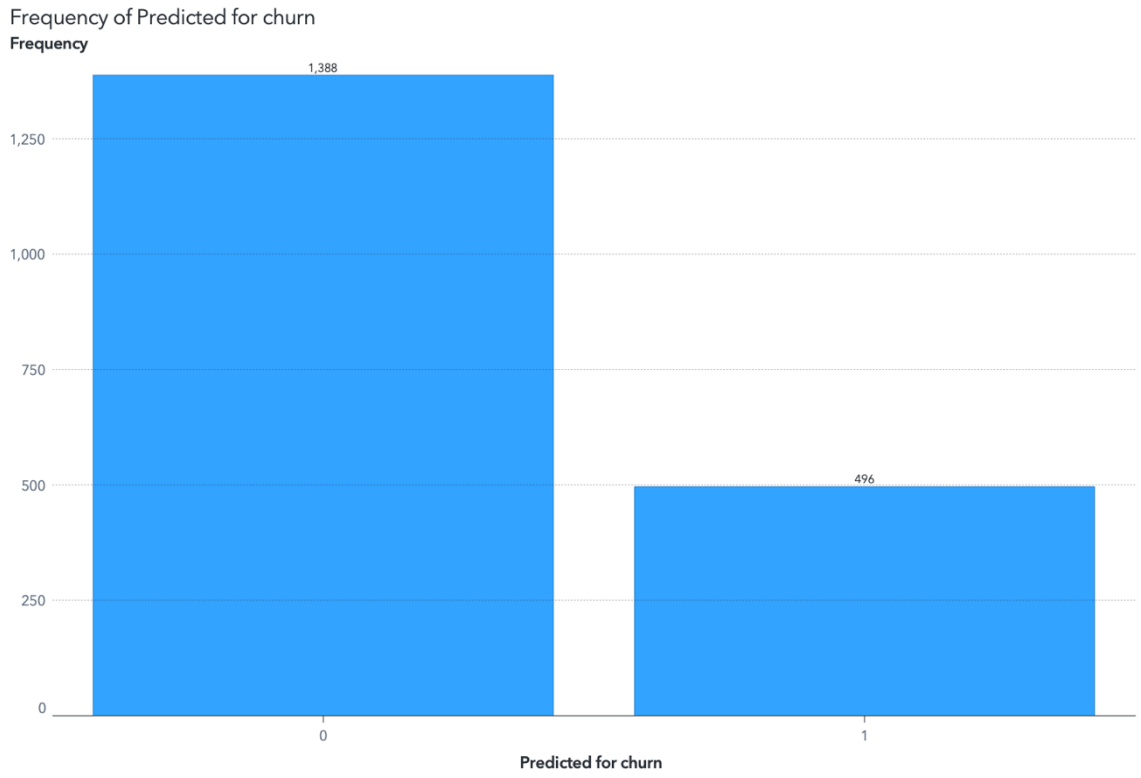

Figure 15 Bar chart - Frequency of predicted customers for churn

Question 19

Using the Key Value Object and setting the aggregation of the measure “Probability for churn = 1” to Maximum and Minimum Respectively, we identified the biggest and smallest probability of being a churner assigned to a customer, as presented below (Figure 16 & Figure 17):

Figure 16 Key Value Object - Highest Probability for Churn

Figure 17 Key Value Object - Lowest Probability for Churn

Question 20

The software assigns a prediction of 1 or 0 to customers based on the column 'Probability for Churn = 1' in the score data set. This column represents the probability that a customer will churn, as predicted by the model. Customers with a probability higher than the cutoff, which we have set at 16.667%, are classified as churners (assigned a prediction of 1), while those below the cutoff are classified as non-churners (assigned a prediction of 0). To validate this, we selected two customers (): one classified as a non-churner and one as a churner, based on their respective probabilities. The customer with the phone number 366-2273 is classified as a non-churner, with a probability of 0.16624, which falls below the cutoff. Conversely, the next customer in the sorted list, with the phone number 349-4396, is classified as a churner, with a probability of 0.16674, exceeding the cutoff.

Phone	Predicted for churn	Probability for churn =1
366-2273	0	0.1662478476
349-4396	1	0.1667413956

Figure 18 Screenshot at the change of Non Churners to Churners with respect to Probability for Churn