**Business Analytics Practicum I**

**Assignment**

**Names**

Eleni Ralli (f2822312)

Viktor Kalatzis (f2822318)

**Athens, 2024**

**Case Study 1**

1. **Executive Summary**

The goal of the study was to determine how to increase sales for "Buy-books-on-line.com," an online retailer of science and IT books. The "Business Analytics" category, one of their store's most popular sections, is the focus of the analysis. This section contains 56 different books, at least one of which was purchased by 1,896 patrons in the previous year. The store's sales team wants to use the data it has to recommend books to customers that they might enjoy, so they carried out a Market Basket Analysis using a dataset of 19,805 historical sales from the "Business Analytics." These records are used to determine the best-selling and worst-selling books. In order to determine which additional two books to suggest to customers who are interested in any of the four "Business Analytics" titles, we also looked for purchasing trends among a selection of four titles. Additionally, for the duration of this project, we were able to strategically promote the other two books to customers who had purchased any one of the set because we had discovered a trio of books that were frequently purchased together. SAS Visual Data Mining and Machine Learning software was utilized to help them make sense of the data and make informed decisions about how to advertise more books.

1. **The sales (in units) of each book**

As we can see in Figure 1 which shows the sales (in units) of each book (and in table 1). The number of books sold is on the x-axis, and the list of the 56 titles is organized from top to bottom on the y-axis based on their sales performance. Individual bars on the graph correspond to the respective titles, and the sales figures for each are marked alongside. The top-selling title is "Data Science and Business Analytics," reaching 1596 units, whereas "Managerial Analytics" appears at the bottom of the list with sales 152 units.

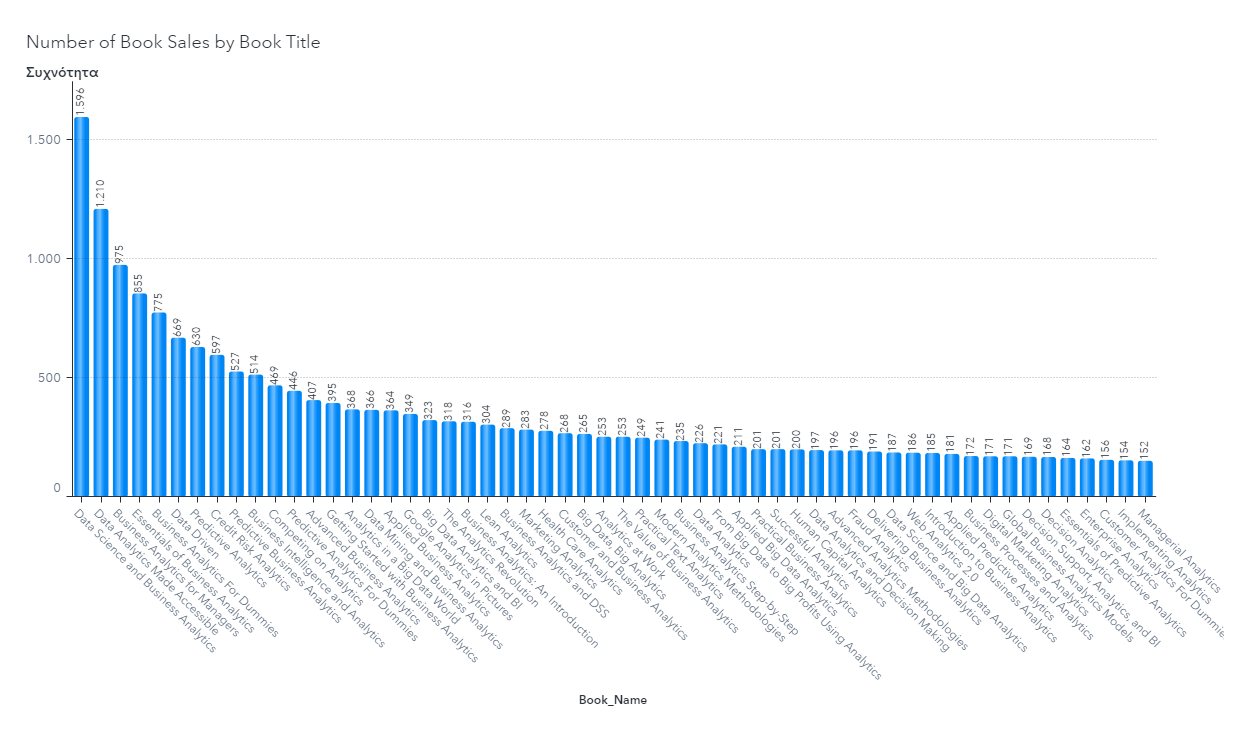


Figure 1: the sales (in units) for each book.

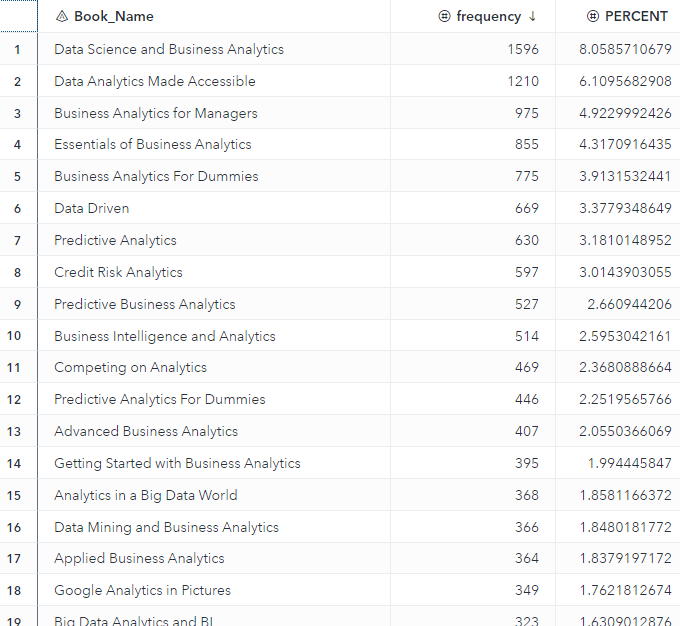


Table 1: 18 first Sales in Volume and Percentage Share of Each Book in the 'Business Analytics' Category.

**3) Book Pairing Recommendations**

To see which publications would be most appropriate to promote alongside individual titles such as “Managerial Analytics,” “Implementing Analytics,” “Customer Analytics for Dummies,” and “Enterprise Analytics,” we employed the lift metric from our Market Basket Analysis. We looked closely at the sales data to find patterns where one of these books was often bought before others (see Table 2 and in appendix table 3). We then listed these patterns by how strong the connection was, from strongest to weakest (by arranging these rules by the descending order of their lift values), and found the two best books to recommend with each one (the top complementary two titles for cross-promotion with each book). The lift index is a metric that tells us how much more likely two items are to be purchased together as opposed to being bought randomly.

Those searching for "Managerial Analytics" are likely to find "Web Analytics 2.0" and "Implementing Analytics" to be relevant choices. With a lift value reaching 11.47, these titles are more likely to be bought by someone who is already chosen "Managerial Analytics" than by an average, uncommitted browser.

The customers searching for "Implementing Analytics," the data suggests coupling it with "Managerial Analytics" and "Data Science and Big Data Analytics" for promotion. The calculated lift of 11.33 indicates a significant propensity for these books to be purchased together over random selection.

Those searching for "Customer Analytics for Dummies" might also appreciate the insights offered by "Enterprise Analytics" and "Decision Analytics." Their association carries a lift value of 11.19, predicting a higher likelihood of these books being purchased together over random selection.

Those searching for "Enterprise Analytics," the data show coupling it with "Managerial Analytics" and "Customer Analytics for Dummies", considering their lift value of 11.07, show a greater tendency for these books to be purchased together over random selection.



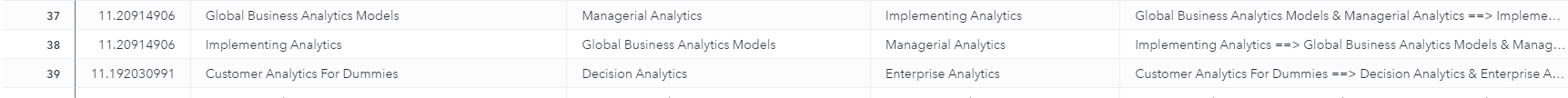




Table 2: Market Basket Analysis table instance of some lift rules (see Appendix for more columns and books).

**4)** **Analyzing the Association of Three Frequently Purchased Books**

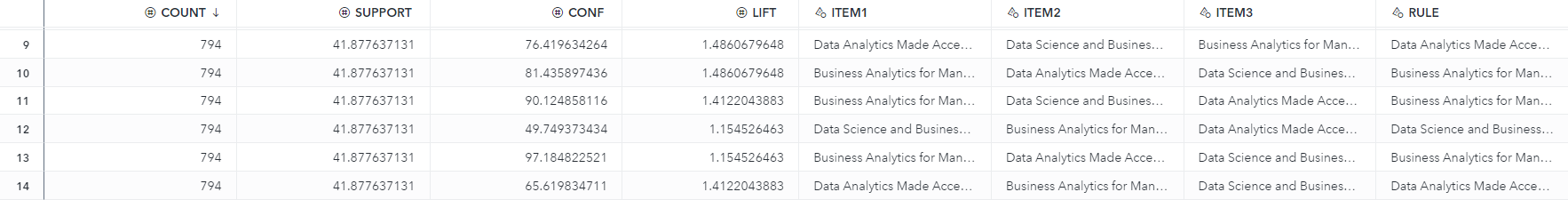
The books "Business Analytics for Managers," "Data Analytics Made Accessible," and "Data Science and Business Analytics" are frequently purchased together when we restrict our analysis to rules involving three items. These books are well-liked by customers, as evidenced by the 794 transactions that included them in the same purchase (see Table 4).

How frequently a set of items is purchased together in all transactions is indicated by the support metric. The support metric is 41.87% for the group that includes "Data Science and Business Analytics," "Data Analytics Made Accessible," and "Business Analytics for Managers." This indicates that these three books were purchased together in 41.87% of all purchases. We divide the total number of transactions (1896) by the number of transactions with these three books (794), to determine this percentage.

Support = (Number of transactions with the specific book set) / (All transactions)

= 794 / 1896

A high support metric suggests that the item set is common, making such rules useful because they apply to many purchases. For instance, a support of 50% means the rule is relevant to half of all transactions, while a support of 5% shows it’s relevant only to a few. Thus, the support metric is vital in analyzing data and finding buying patterns.



In zoom to show the Rule :

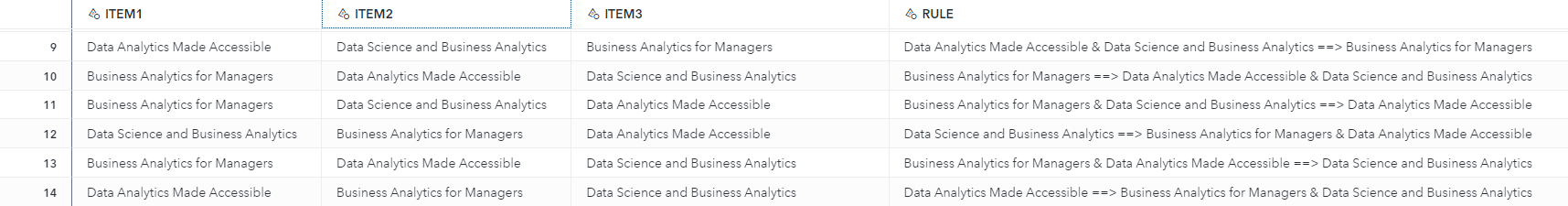


Table 4: Association Rules for the 3 books most bought together

**Case Study 2**

1. **Executive Summary**

Sports-OnLine.com is a web store-retailer that sells sports clothes and shoes. They want to use the data they have collected to understand the market better. They will do an RFM (Recency, Frequency, Monetary segmentation) analysis to find out who their best customers are by grouping them. They found five types of customers. "Lost Customers" have the longest time since last purchase, infrequent transactions, and the lowest spending. "New Customers" have recently made purchases, though they do not purchase often and their overall spending is modest. "Churners" have been inactive for a while, yet historically they made purchases more often than others and have spent considerable amounts. "Risk Customers" are on the verge of becoming inactive, with substantial periods since their last purchase, average purchase frequency, but they have spent more than most. Finally, "High Value Customers" are the most engaged, with the most recent purchases, most frequent transactions, and the highest overall spending. To help these groups to buy more and have a better experience, the store plans to ask for customer opinions, make their service better, make discounts, give special deals, and start a rewards program.

1. **Data preparation for RFM Analysis: Transformations and Outlier Management**

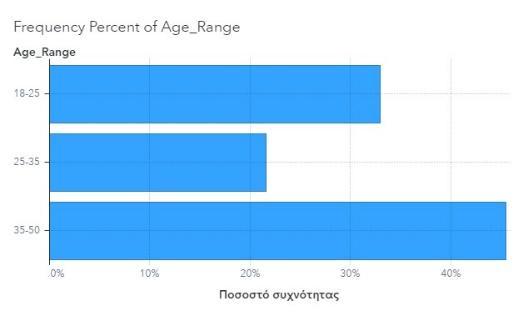
First, we make some graphical representations to understand better the data. The three bar graphs show data on Age Range, Gender, and Payment Method (see-Figure 2). The Age Range graph indicates the highest frequency percentage falls in the 18-25 range. In the Gender graph, females are represented slightly more than males. The Payment Method graph shows debit cards as the most common method, followed by credit cards, with PayPal and cash being less frequent. These visuals likely reflect certain consumer demographics and payment preferences within a dataset. The histograms-distributions (see- Figure 3, Appendix Figure 4) show customer purchasing patterns through the RFM ((Recency, Frequency, Monetary) values.

The histogram of Recency presents two customer categories, one of them that tends to purchase within a recent 2-7 months window, and another one that generally waits 15 months or more to buy.

The histogram of Frequency depicts that most customers have 3-7 transactions with the company.

The distribution of monetary presents an important portion of customers spend in the range of 200-800 euros.

These insights are important to understand how customers interact with the business and will guide future marketing and sales tactics. The plots show a positive skew in the R, F, and M variables, which shows we need to convert our data with log transformation to normalize the data distribution.



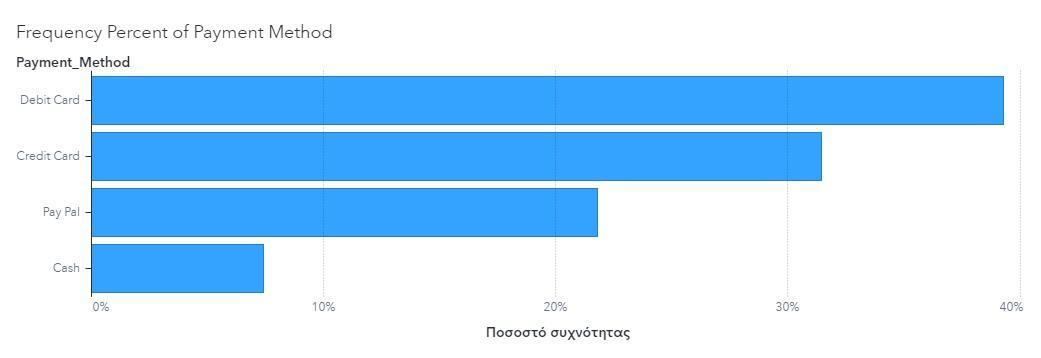
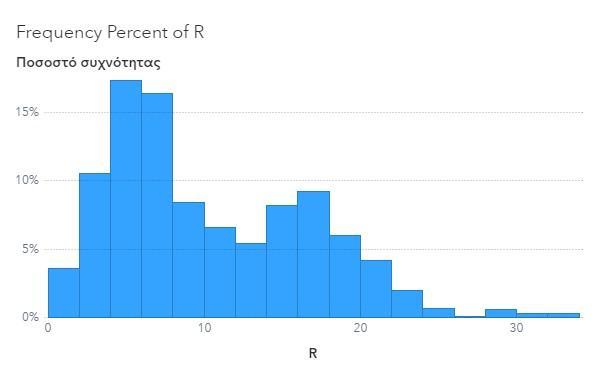
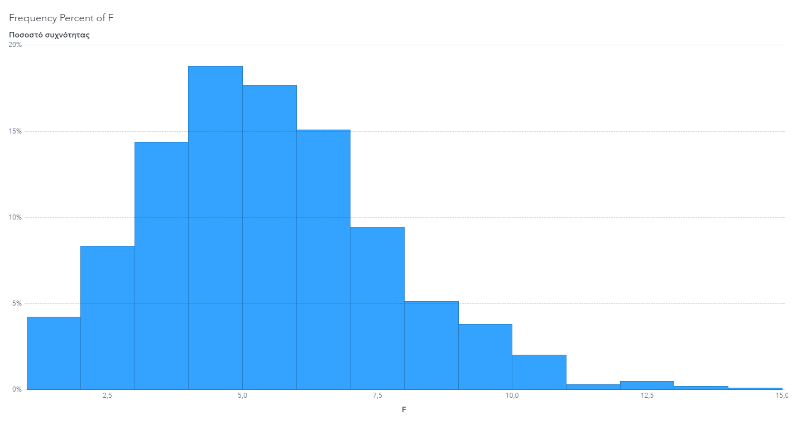


Figure 2: Barplots for Frequency Percent of Age, Gender, Payment Method 



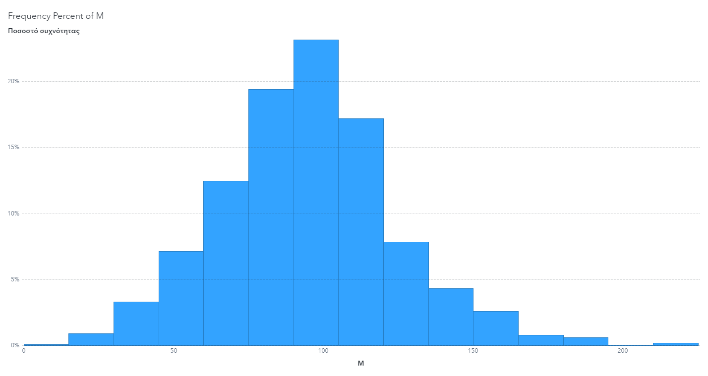


Figure 3: RFM histograms

In order to see if we have outliers, we make boxplots for the RFM variables (e.g., Figure 5). As we can see there are data points that significantly diverged from the rest of the distribution (dots above the whiskers in the plot). To maintain the integrity of the dataset, entries that were considered extreme were excluded based on specific thresholds for each variable: for the R data with values that exceeded 30, for the F data with values that was over 10, and for the M data with values that surpassed 1118.

The median figures for these variables stood at 8 months for R, 5 transactions for F, and a monetary value of 435€ for M.

Data cleansing proceeded by omitting records that deviated markedly, more than three standard deviations from the mean, and transforming skewed variables logarithmically. The process began with 995 individual customer records, which, after removing 19 outliers, we remain with 976 complete entries for subsequent analysis. Also, we check for missing values and we don’t have to do to imputation (e.g., Figure 4 appendix).

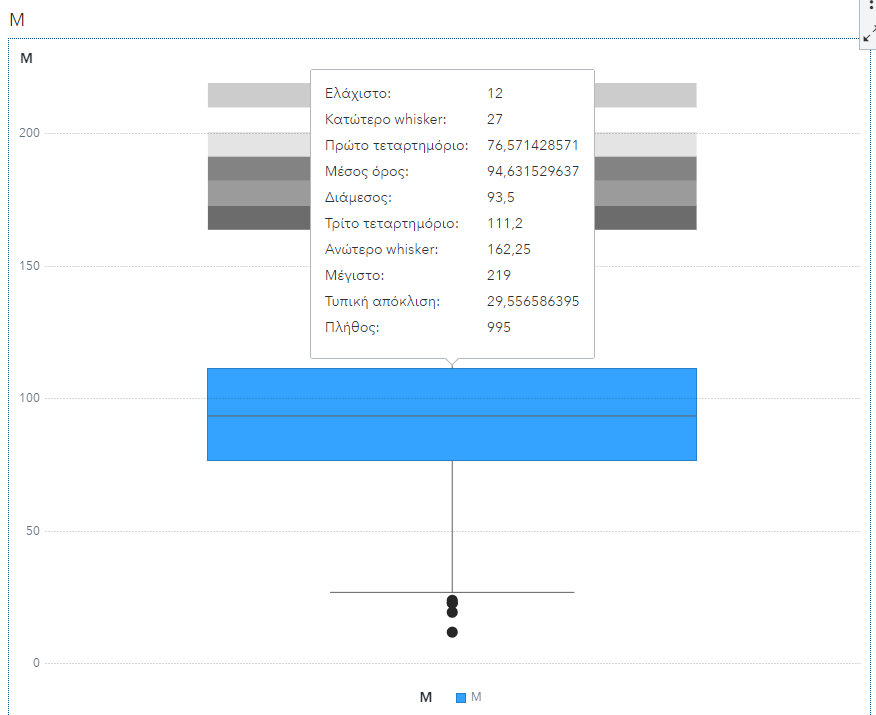
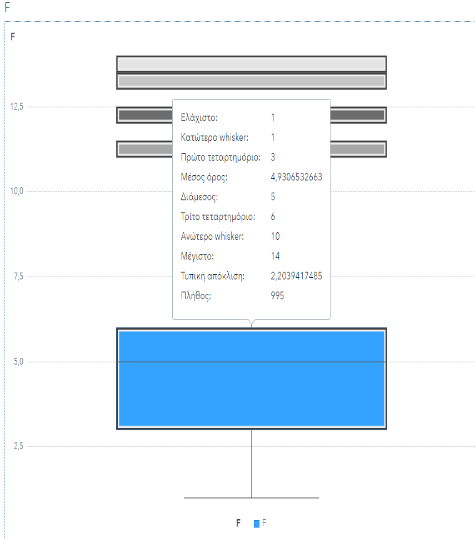
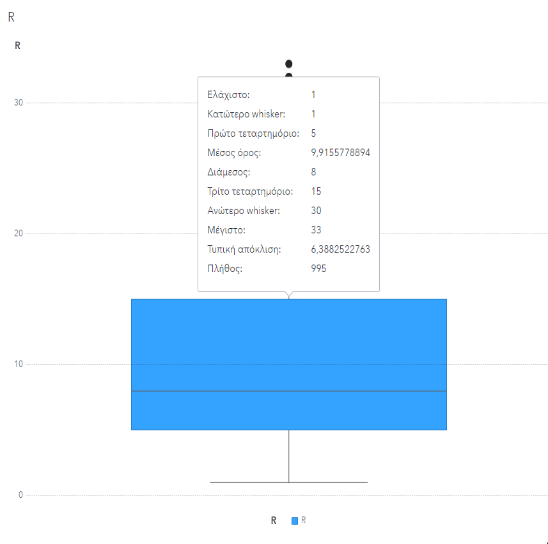


Figure 5: Boxplots for RFM

1. **Naming and Understanding Customer Groups**

The RFM analysis was conduct be using the k-means clustering method with Euclidean distance to categorize customers into four specific groups, each displaying unique buying behaviors. The analysis segmented the customers into five groups.

A detailed segmentation helps to understand customer behavior per cluster, the average profile of a customer per cluster (average Recency, Frequency, Monetary and the frequency percentage of each cluster), providing valuable insights for the Business Analytics team to devise targeted marketing strategies and promotional offers (see - Table 5). insights into customer engagement levels, from those who may require re-engagement strategies to those who exhibit high-value purchasing behaviors.

In more details, the 'Recency' column measures the average time (in months) since the last transaction made by a typical customer in each cluster. The 'Frequency' is the average number of transactions each customer group has made with the company over the past seven years. The 'Monetary' quantifies the total expenditure of the typical customer in each segment throughout their transactions.

To classify these groups, the RFM values of their ideal customers were evaluated against the overall average values of the dataset. In the table 5, a color scheme of green and red was applied to highlight whether each value in the R, F, M columns was above or below the dataset's mean, guiding the labeling of each cluster based on their performance in these categories. Specifically, for recency, a value below the average is preferable, whereas for frequency and monetary values, higher figures are more favorable.

With 45 customers, the first cluster, titled "Lost Customers," represents 4.7% of the total customer base. Customers in this category have, on average, not made a purchase in 15.93 months, have made purchases 1.98 times on average, and have spent an average of 43.31€. A small amount of the total sales are attributed to this cluster.

Let's move on to the "New Customers" segment, which accounts for 129 customers or 13.48% of the total customer base. This group's recent and introductory activity with the company is indicated by their average purchase recency of 4.79 months, average purchase frequency of nearly 3.98 times, and average spending value of 66.64€.

The third cluster, referred to as "Churners," comprises 153 members and represents 15.99% of the customer base. Consumers in this category may be at risk of doing away with their business with the company because they have, on average, spent 75.34€, made purchases 5.24 times, and it has been 14.68 months since their last purchase.

With 296 customers, the fourth group, referred to as "Risk Customers," makes up the largest segment and accounts for 30.93% of the customer base. These customers spend an average of 115.05€, haven't made a purchase in almost 14.04 months, and buy about 3.81 times per year, suggesting a substantial but potentially declining level of engagement with the business.

Finally, the "High Value Customers" cluster include 34.9% of the total customers, equal to 334 customers. These customers made on average a purchase recency of about 4.8 months, a higher purchase frequency averaging around 6.55 times, and have an average expenditure of 99.81€. This segment appears to be the most engaged and valuable group, contributing an important portion to the company’s revenue.



Table 5: RFM Analysis

The first chart, such as Figure 6, shows how much money each segment contributes, with "High Value Customers" and "Risk Customers" making up the largest shares of the total revenue at 37.3% and 38.1%, respectively. This demonstrates how these groups have a major financial impact on the company's profits; of particular concern are "Risk Customers" because of their decreasing engagement, which may call for proactive retention strategies.

The frequency of the customer base across the segments is the main topic of the second pie chart (see, for example, Figure 7). With 34.9% of the frequency share, "High Value Customers" once again make up the majority and demonstrate their ongoing interaction with the business. On the other hand, "Lost Customers" account for the smallest portion (4.7%), which indicates their lower engagement level and the possibility of focused reactivation efforts.

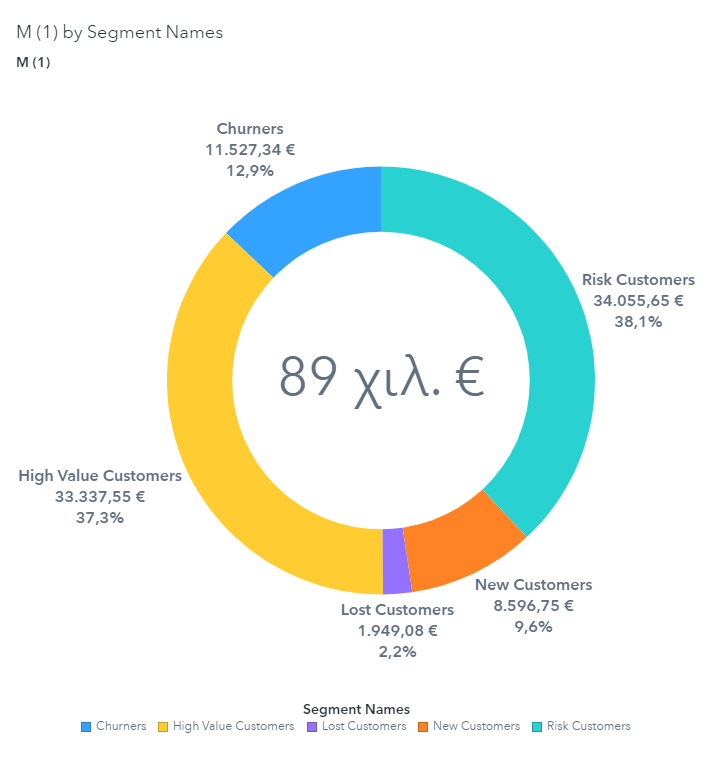
All things considered, these graphs highlight the different degrees of involvement and monetary contribution of every customer category, providing the Business Analytics team with a solid basis upon which to effectively customize their marketing and CRM tactics.

Figure 6: Pie chart for the monetary contribution of each segment

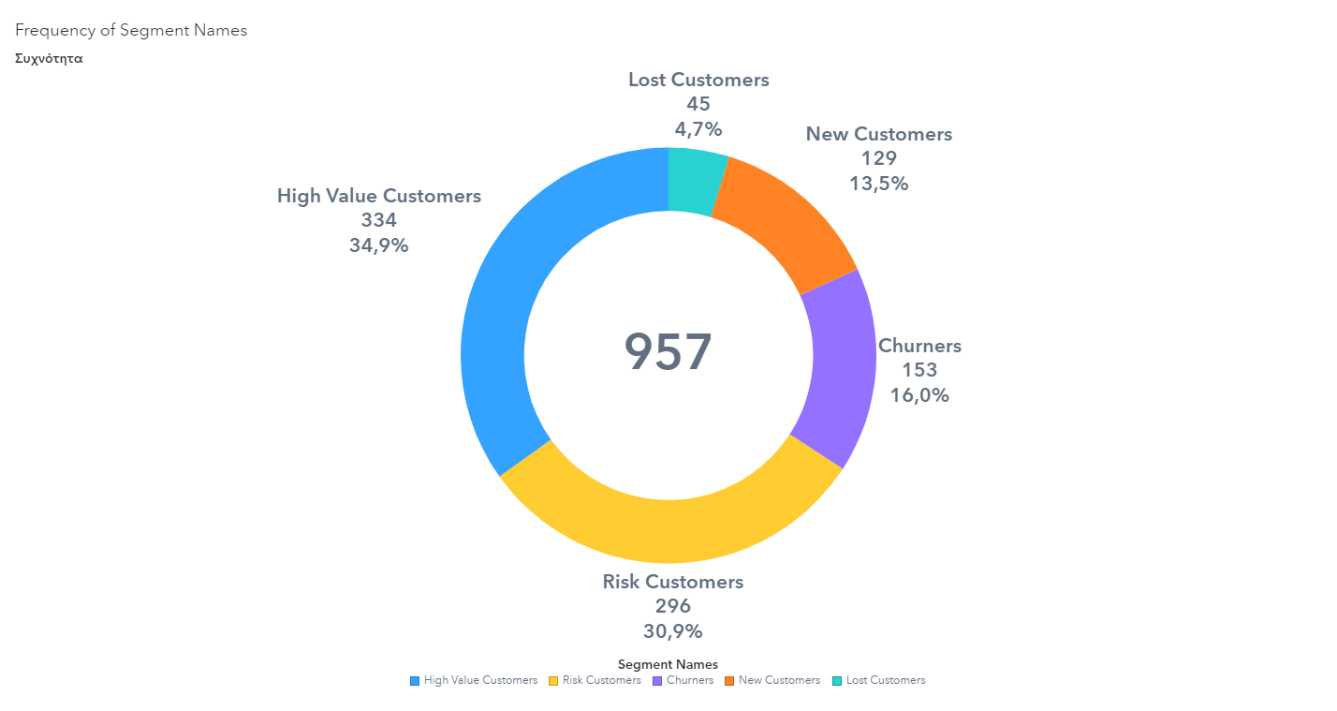


Figure 7: Pie chart for the frequency of the customer base across the segments

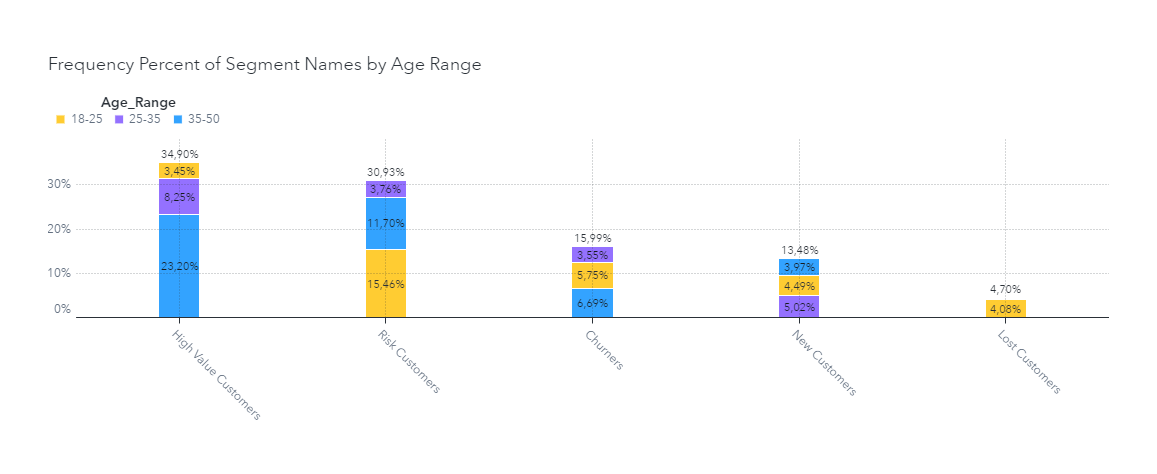
1. **Dashboard Profiling with Demographics and Payment Preferences**

From the first diagram, "Frequency Percent of Segment Names by Age Range," we can observe the distribution of age ranges within different customer segments (see-Figure 8, Appendix Table 5). High Value Customers have a significant representation in the 25-35 age bracket. Risk Customers and New Customers show a more even distribution across the age ranges, with a slight predominance in the 18-25 range for New Customers.

The interactive dashboard figure 9 show a segmented analysis focused on "Churners," highlighting the role of age range and various predictive attributes in determining the likelihood of customers discontinuing service. A significant note from the analysis is that Segment Names have a 15.99% chance of being Churners, which is the third most common outcome. The bar chart shows that the 35-50 age range is most common among Churners, and the text details probabilities linking specific transformed variables to the chance of being a Churner, with age being a significant factor. For instance, within the 18-25 age bracket, if certain conditions involving transformed variables are met, there's over a 60% chance that those customers will churn.

The diagram "Frequency Percent of Payment Method by Segment Name," details the preferred payment methods across customer segments (see-Figure 10, Appendix Table 5). Debit cards are the most popular across all segments, but notably more so among Churners and High Value Customers. Credit cards follow, with High Value Customers showing a higher preference for this method than other segments. PayPal and cash have lower frequencies but are still relevant payment methods, with Lost Customers showing a slightly higher tendency to use PayPal.

The interactive dashboard figure 11 show an analysis of payment method preferences amongst Churners, juxtaposed with predictive data. The bar chart contrasts Churners with non-Churners, emphasizing that debit card is the most common payment method among Churners. In terms of predictive power, specific ranges of transformed variables combined with the use of debit cards and PayPal significantly increase the likelihood of a customer being a Churner, with chances up to 70.59%. This nuanced view suggests that while payment method alone provides some insight, its predictive value is greatly enhanced when considered in conjunction with other customer attributes.

Figure 8: Frequency Percent of Segment Names by Age Range

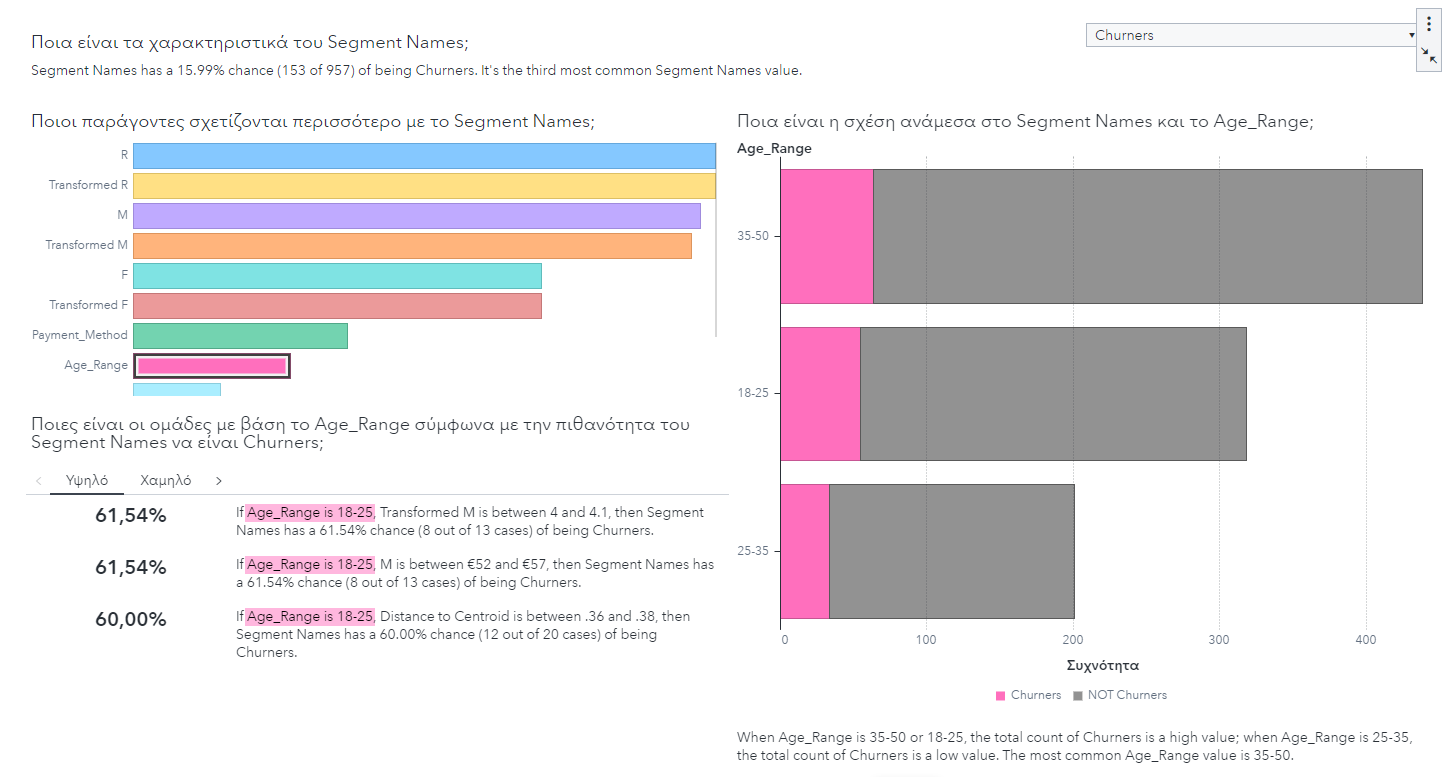
****

Figure 9: Churner Profile by Age Range and Predictive Attributes

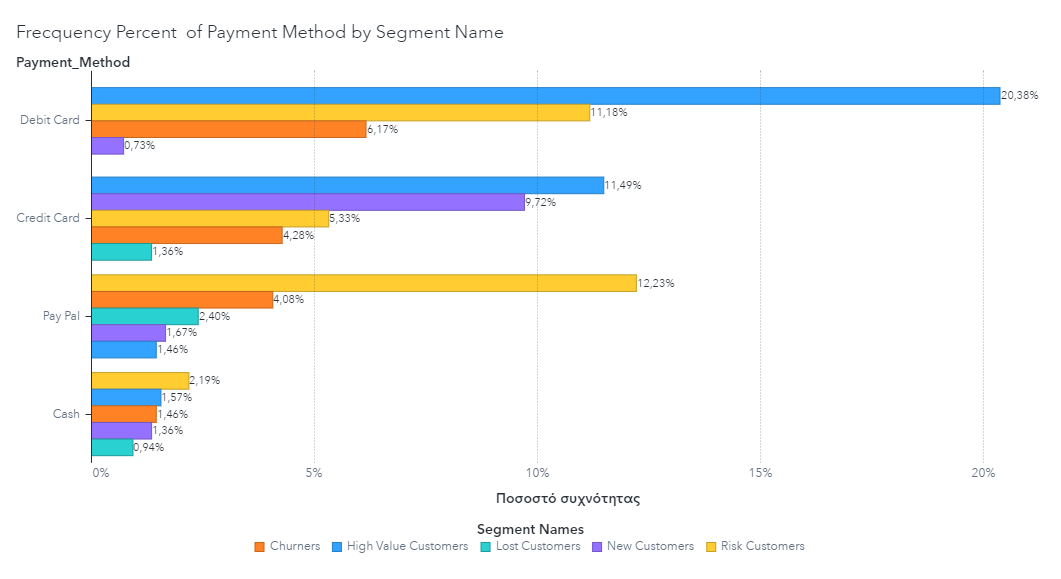
****Figure 10: Frequency Percent of Payment Method by Segment Names

Figure 11: Churner Payment Method Preferences and Predictive Analysis

1. **Marketing Strategies for RFM Analysis**

To be able to optimize marketing efforts, we will design specific actions for each customer segment by aligning them a with the specific characteristics and behaviors of each group, thereby increasing engagement, customer satisfaction, and, ultimately, revenue for the company. a

For Lost Customers (4.7%): A customer reactivation program should be applied for this group. As these customers a have a not a made purchases a for an extended period (approximately 15.93 months), re-engagement efforts could be sending them personalized emails highlighting major improvements a or new products since their last interaction, in the same time with exclusive offers to make a new purchase. A dh

For New Customers (13.5%): This segment has presented recent interaction with the company, depicting potential for development into loyal customers. A good strategy would be to contact them for feedback to understand their experience and to provide personalized recommendations based on their first purchases (cross-sell activities). A next action may be a special welcome promotion that can enhance their customer experience. Apply

For Churners (16%): As this group has a recency gap and is at risk of leaving, a dual strategy of feedback solicitation and special promotions would be good. The feedback will provide insights into their inactivity, and the promotions should be designed to rewin their interest, perhaps through a loyalty program that rewards consecutive purchases.

For Risk Customers (30.9%): In spite of being the largest segment and having a high average spend, their engagement is declining. It is important to identify the reasons for the potential churn and face them immediately. On their next purchase, offering special promotions or discounts can help. Furthermore, personalized engagement through loyalty programs or VIP customer events could reinforce their relationship with the company. a

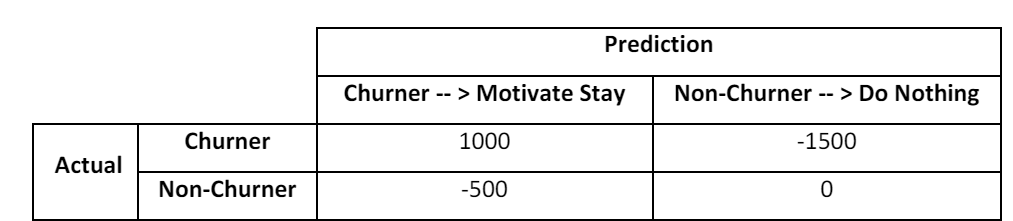
For High Value Customers (34.9%): This segment is important as they are frequent customers and spend considerably. To keep and grow this segment, exclusive benefits and rewards should be offered through a tiered loyalty program. Personalized services, early access to new products, and special appreciation events can improve their experience. As well as Cross-selling and upselling techniques may be effective by providing these customers with personalized product recommendations according to their purchase history data.

**Case Study 3**

1. **Executive Summary**

XYZ Mobile is using a strategic use of machine learning to increase client retention in an effort to counter growing market saturation and rivalry. Their rigorous strategy of using past customer data to identify which clients are most likely to cancel their subscription is demonstrated in the case study. The main goal of the project is to develop a statistical model that can analyze customer interaction data spanning six months in order to predict the likelihood of churn for the upcoming quarter. First, in order to better comprehend the data, we performed an exploratory data analysis. XYZ Mobile uses patterns such as call frequency, service usage, and customer service engagements to identify consumers who may be at danger of leaving. The data was divided into training and validation sets, using the 70%-30% split, to guarantee the accuracy of the model. This makes it possible to train and validate models rigorously, which is necessary for them to generalize effectively and prevent overfitting. The management team has created a profit matrix that acts as a compass for the model, directing the distribution of incentives. Using SAS tools for machine learning and analytics, the team established a cut-off point of 16.6% for churn probability, which is the tipping point for deciding whether a customer might leave. This threshold plays a crucial role in strategy deployment, ensuring resources are focused where they are most likely to yield a return on investment. We generated four models. These were the Decision Tree, the Maximal Tree, the Logistic Regression, and the Neural Network. The best model of the forementioned comparison was a logistic regression model, which was finally applied to new “unseen” data provided by the company to evaluate the model developed. The total number of customers in the data is equal to 1,884. From those customers the 1,365 of them are predicted by the best model as non-churners and 519 of them as churners.

1. **Interpretation of the Profit Matrix**

****Table 6: Profit Matrix

A tool used in the customer churn prediction model to quantify the financial outcomes associated with various scenarios is the profit matrix. Based on the customer's actual status (i.e., whether they churn or not) and forecast status (i.e., whether the model identifies them as likely to churn or not), the matrix presents four alternative outcomes. The predictive model's objective is to minimize False Positives (the model incorrectly predicts a non-churning customer will churn) and False Negatives (the model fails to identify a customer who will churn) while optimizing the number of True Positives (the model correctly predicts a customer will churn). This is done by weighing the benefit of reduced churn against the expense of retention strategies.

True Positive (cell on the upper-left): The model accurately forecasts a customer's churn. The business takes action by encouraging the client to stick around, which generates a profit of one thousand dollars. This figure probably represents the money kept in revenue by keeping customers from leaving less the amount spent on retention incentives. False Positive: The model predicts wrongly that a non-churning consumer would experience a churn (down-left cell). The cost to the company of 500 monetary units occurs when it gives them unnecessary incentives to stay. This amount shows the money that was lost on incentives given to clients who weren't likely to leave. True Negative (down-right cell): No action is taken because the model accurately predicts a customer won't churn. For consumers that are not at risk of churning, this results in no financial change (0 monetary units), which is the desired consequence. False Negative (upper-right cell): The model is unable to recognize a client who is likely to leave. If nothing is done, the customer will leave, costing the business a substantial 1500 in monetary units. This loss is the largest and probably represents the money that was lost from the client over time.

1. **Determining the Churn Probability Threshold**

As we want to find the minimum probability which corresponds to when a customer will be considered a churner and hence to be considered for a contact to prevent him/ her from attrition (churn), the expected profit when the company solicits the customer and the expected profit when the company ignores him is necessary to be estimated. We construct two equations for the previous occasions and we solve them. If we have that is the probability to be churner and 1- non-churner, the threshold (cut-off point) is provided by the below mathematical calculations.

To find our cut off point we should solve the equation of:

=

Consequently, the customer’s minimum probability to solicit positive is 0.166 or 16.6%.

1. **Data Partitioning and Model Evaluation Strategy**

It is common practice to split historical data into training and validation sets, usually utilizing a 70%–30% split, when splitting the data for common machine learning tasks. This section aids in the efficient construction and assessment of a model.

Model Training: The model is trained using the training set, which normally consists of a greater percentage of the data—in this example, 70%. By modifying its parameters in accordance with the supplied inputs and their related outputs, it enables the model to learn.

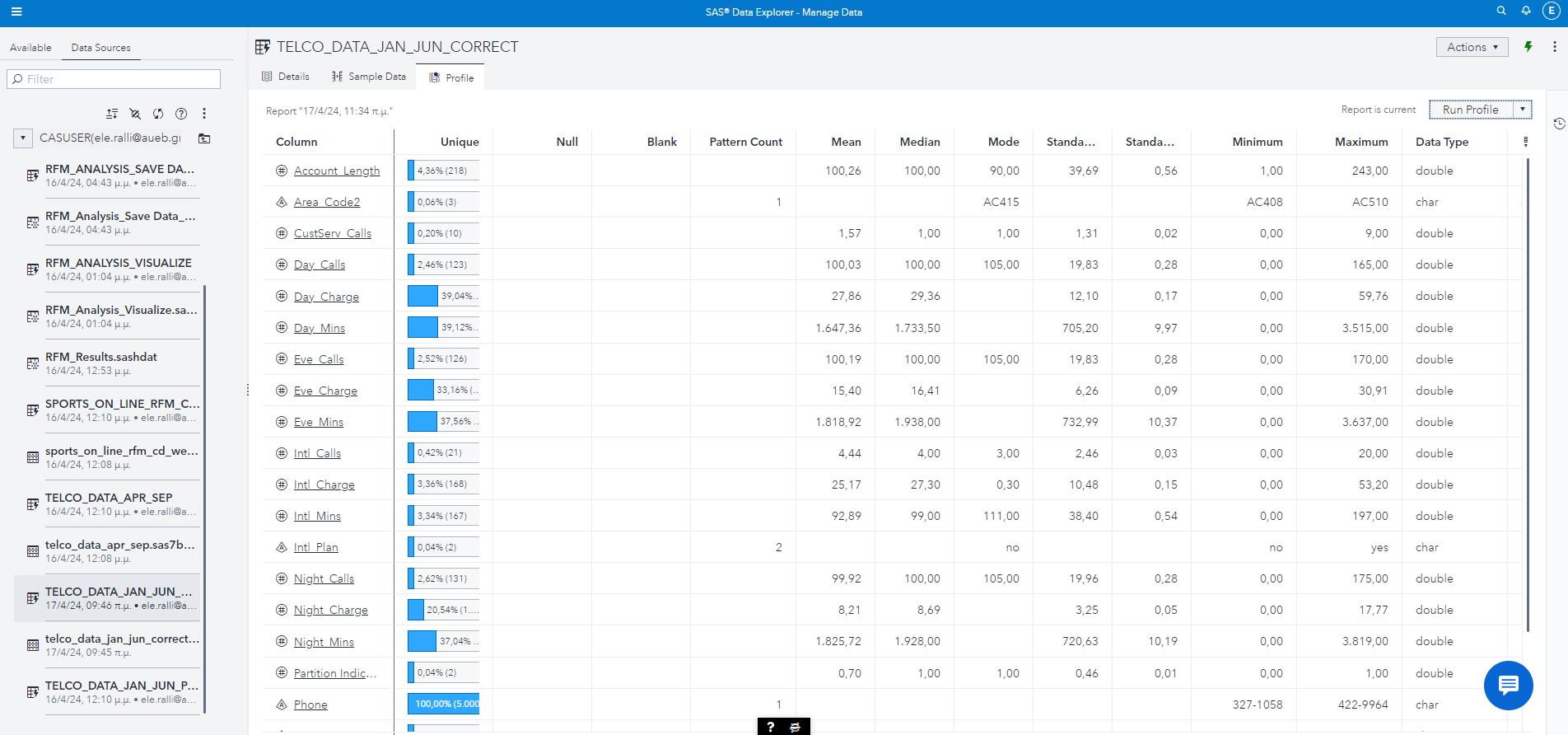
Validation of the Model: The model's performance on fresh, untested data is assessed using the validation set, which comprises the remaining 30% of the data. This aids in evaluating the model's learning process and its performance on non-training datasets. It is vital for identifying overfitting, a phenomenon in which a model works well with training data but badly with unknown data.   
An overly simplistic model may fail to identify the underlying patterns in the data, leading to underfitting, or excessive bias, in which the model repeatedly fails to identify important signals. On the other hand, an overly complex model may fit the peculiarities of the training data too well, resulting in overfitting, also known as high variance, when the model interprets noise as signal.   
The ideal model finds a happy medium, having just the right amount of complexity to allow for good generalization to new data. The training set is used to generate a number of models with different levels of complexity in order to establish this balance. Both models that are prone to overfitting and underfitting fall within this spectrum. The validation set is then used to assess these models' performance. This auxiliary data collection optimizes the complexity of the models, assisting in their fine-tuning. It offers a crucial feedback loop for figuring out whether a model is too simple or too complex, and if more training or modifications are required to get the best out of the model.

By splitting the dataset into training and validation sets so that each set is representative of the entire, a technique known as stratified sampling is used. More specifically, in both the training and validation sets, the percentage of each category of the target variable is maintained. In datasets where one class may be underrepresented, this is especially crucial. The model gains more accuracy in class prediction by guaranteeing that every class is fairly represented in both sets.

One performance metric used to assess the classification model's accuracy is the Misclassification Rate (Event). Out of all the predictions produced, it determines the percentage of inaccurate predictions (false positives and false negatives). For a binary classification job aids in comprehending how well the model predicts occurrences.

1. **Analysis of Data Completeness and Churn Distribution**

There are no missing values in the variables of the data set - where each variable has a non-null count (e.g., Figure 12). The proportion of churners and non-churners in the data set is 14,1% – 85,9% respectively (e.g., [Figure 13).](#_bookmark7)

****

****

Figure 12: Historical Data

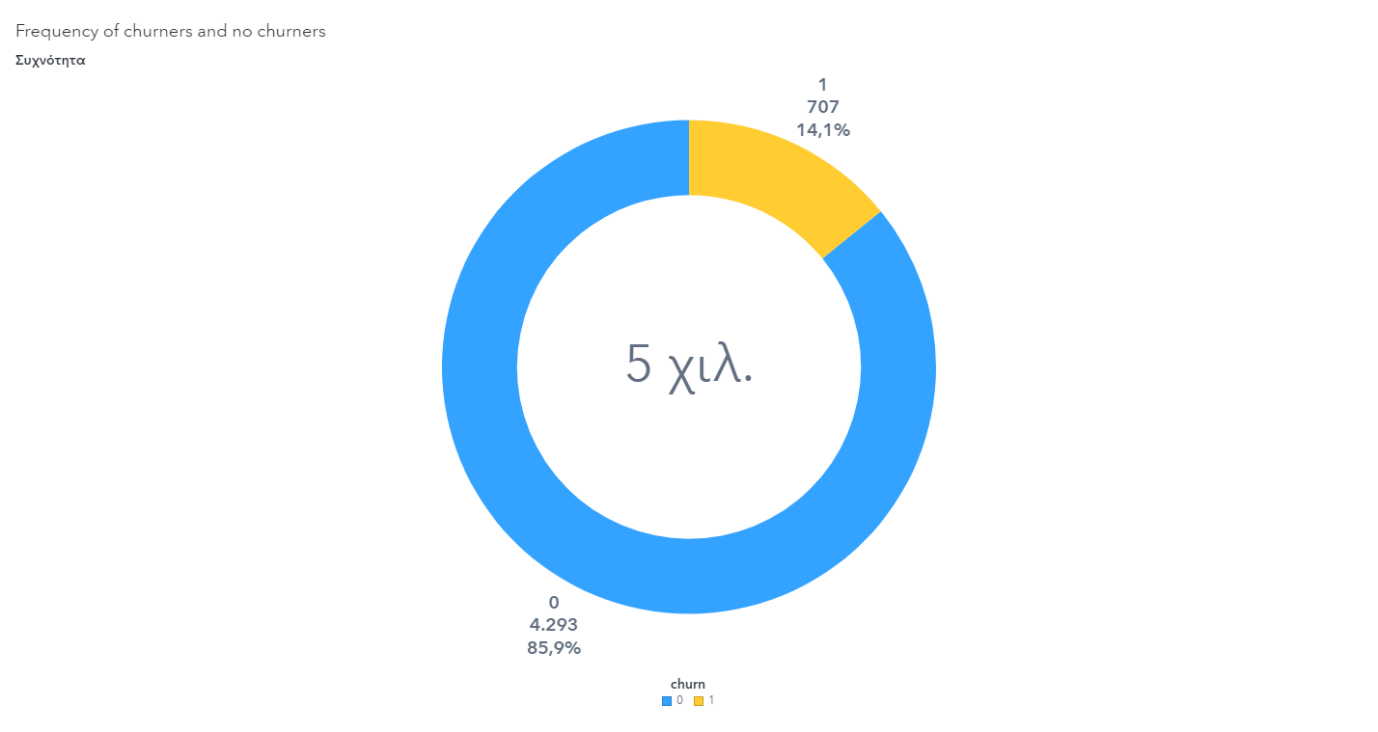
****

Figure 13: Pie Chart of the frequency of churners and non-churners.

1. **Strategies for Handling Imbalanced Data**

If the proportion of churners to non-churners in the historical dataset was significantly imbalanced, such as 3% churners to 97% non-churners, it could lead to several issues with model training and performance evaluation. In this situation we have imbalanced data, because the target of interest (churners) is characterized as a rare event in relation to the total number of samples. Using a model on this kind of uneven data without adjusting for the big difference in numbers will give us a model that is very good at predicting the majority class (non- churners) but fails to provide useful information about the less common, but often more important, class ( churners).

Resampling is a common method for handling datasets with a large imbalance between classes. For a dataset where only 3% are churners and 97% are not, we would adjust the data to better reflect a balance. To increase the churner class size, we might replicate some of their data points—a process known as oversampling. Conversely, to decrease the non-churner class size, we could randomly remove some of its data points, which is known as undersampling. While oversampling might risk overfitting, as the model sees the same churn examples repeatedly, undersampling might lose valuable information by discarding non-churner data. SAS provides tools for undersampling directly in its user interface, but for oversampling, particularly advanced techniques like the Synthetic Minority Oversampling Technique (SMOTE). Or, we can use machine learning algorithms that are less sensitive to class imbalance, such as tree-based methods or ensemble methods like Random Forest or Gradient Boosting. Or, we can adjust the weights in the algorithm to penalize misclassifications of the minority class more than those of the majority class, making the model pay more attention to the minority class. Or, we can adjust the decision threshold after model training to increase the sensitivity of the model to the minority class.

**Top of Form**

1. **Churn Analysis for High-Frequency Customer Service Contacts**

As can be seen, the majority of these customers—60%, designated as '1'—are churners within this category. This raises the possibility of a relationship between the number of customer support interactions and churn risk. It is also evident that 40% of these consumers (designated as '0') are non-churners, suggesting that not all customers who often contact customer service will experience churn. Given that this distribution differs significantly from the dataset's total turnover rate of 14.14 percent, it is possible that increased customer service interactions are a sign of dissatisfied customers and a higher risk of churning (e.g., Figure 14).

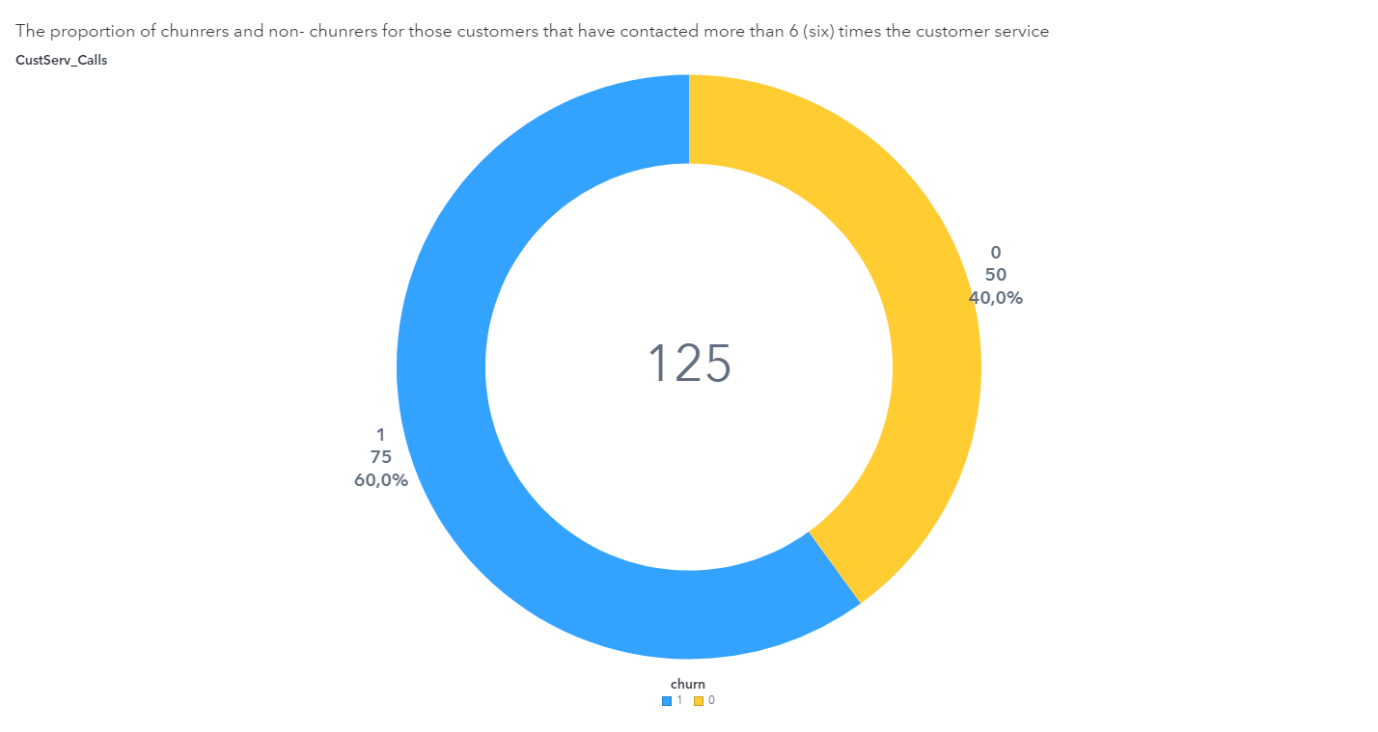


Figure 14: Pie Chart of the proportion of churners and non-churners for those customers that have contacted more than 6 times the customer service.

1. **Daytime Usage Patterns Among Churners and Non-Churners**

The average number of minutes that customers spend on the phone during the day, categorized by churn status can be presented by barchart (e.g., Figure 15). Churners (label= '1') have a higher average of daily minutes spent on the phone which is equal to 1898 minutes, whereas non-churners (label= '0') have a lower average of daily minutes, 1606 minutes.

This difference in daily phone usage between churners and non-churners could show several things with respect to the target variable. Churn, with an average of 1898 minutes, are initially more engaged or maybe more dependent on the service for their communication needs. However, this higher usage may also show a level of dissatisfaction; these customers might be experiencing service issues, make them to use more minutes to resolve these issues before deciding to switch providers. For non-churners, with a lower average of 1606 minutes, it could show that they are either more satisfied with the service, hence requiring less time on customer service calls, or they have a usage pattern that fits within the bounds of their current plans, leading to a lower likelihood of considering other companies.

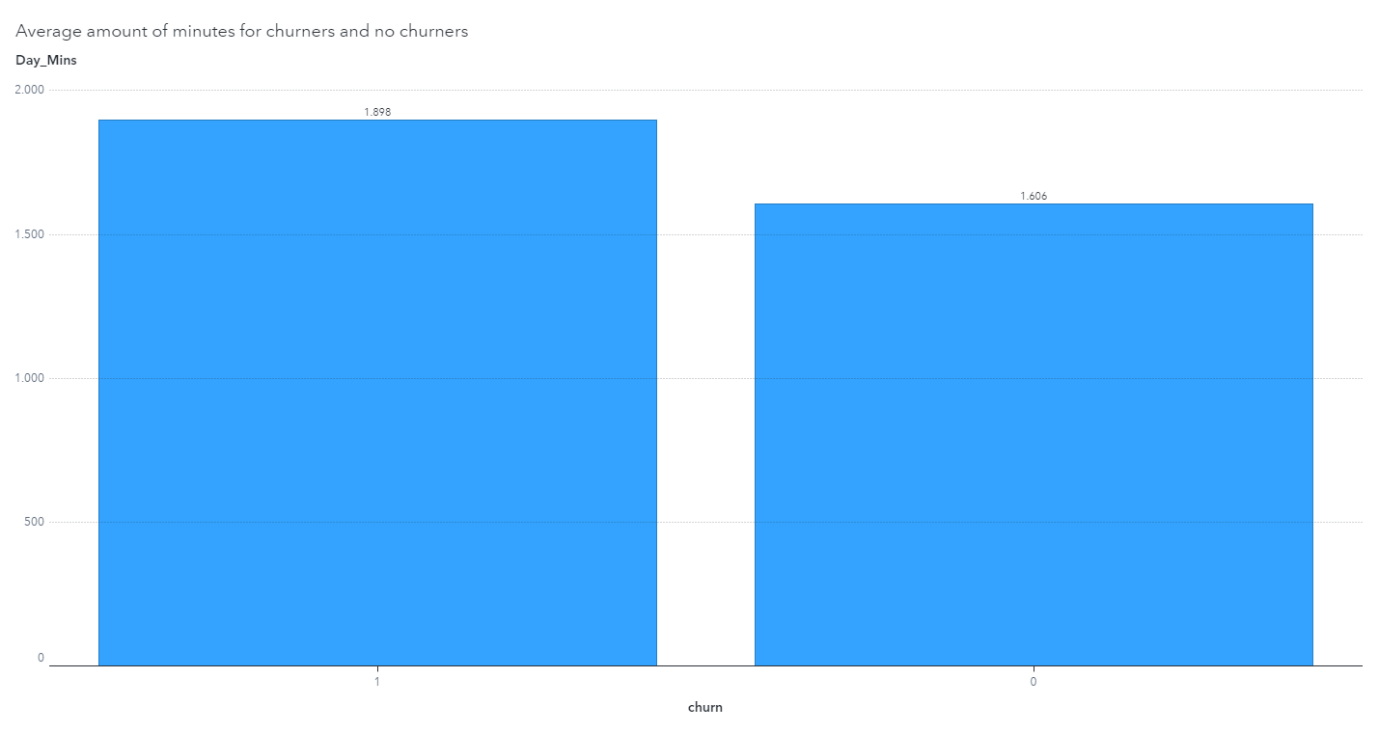
****

Figure 15: Barplot of the average amount of minutes for churners and non-churners

1. **Initial Decision Tree Analysis**

Intl\_Plan, the variable that shows if a customer has an international plan, is the first one selected for splitting in the decision tree. This choice was made because the data had a high statistical correlation with customer attrition, which was determined by a careful investigation employing a split-search technique. The effectiveness of several variables in classifying consumers into groups based on their churn rates is evaluated by this algorithm. A Pearson's Chi-Squared test is used to determine the significance of each split, including Intl\_Plan, by comparing the churn rates of the groups established by the split. We employ the logworth, a negative logarithm of the p-value, as a more trustworthy indicator of significance because p-values have a propensity to be deceptively low in big datasets. After these modifications, Intl\_Plan had the greatest logworth score, which made it the main criterion for the tree's first split. Clients without an international plan should go to the right branch, and those with one should go to the left. To ensure that no data is missed, our decision tree classifies missing data into a designated branch. It is noteworthy, nonetheless, that this specific variable in our sample contains no missing values.

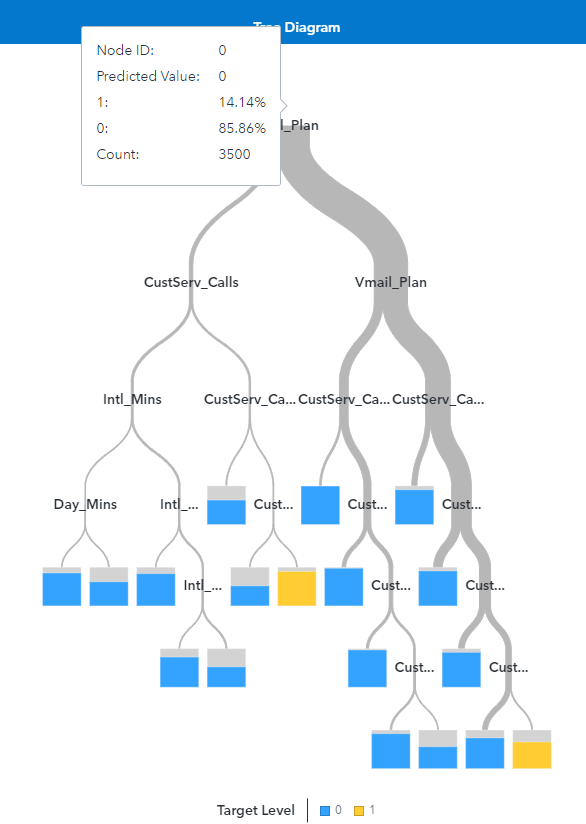


Figure 16: Optimal decision tree

1. **Complexity and Performance of the Maximal Decision Tree**

The maximal decision tree has 22 of terminal nodes (leaves). This tree, with all its splits and leaves, is known as the Maximal Tree (e.g., Figure 17). The Subtree Assessment Plot indicates how the misclassification rate varies for subtrees created by pruning the full decision tree to different numbers of leaves (e.g., Figure 18).

The training error, shown by the blue line, typically decreases as the number of leaves increases. This decrease is due to the tree increasingly fitting the randomness and noise within the training data( overfitting). Overfitting occurs when a model becomes too complex and starts to capture the random noise in the training data, rather than the underlying data pattern. While such a model may perform very well on the training data, its predictions for new, unseen data are generally poor because it has learned from the noise that isn't present in the new data set. With pruning we reduce overfitting, which involves cutting back the tree to a size where it performs optimally on validation data.

In the Misclassification Rate graph, we search for the point where the validation error is minimized, indicating a good generalization performance. This optimal subtree size typically has a higher misclassification rate on the training data compared to the maximal tree but is expected to perform better on unseen data. With pruning we reduce the complexity of the maximal tree by trimming off branches that have little to no predictive power. This process results in a smaller, more generalized tree that performs better on validation data. Also, from the pruning error plot, you can select the subtree with the number of leaves that provides the lowest misclassification rate for the validation data, not just the training data. This subtree represents a pruned version of the decision tree, balancing complexity with predictive power to avoid overfitting.

Top of Form

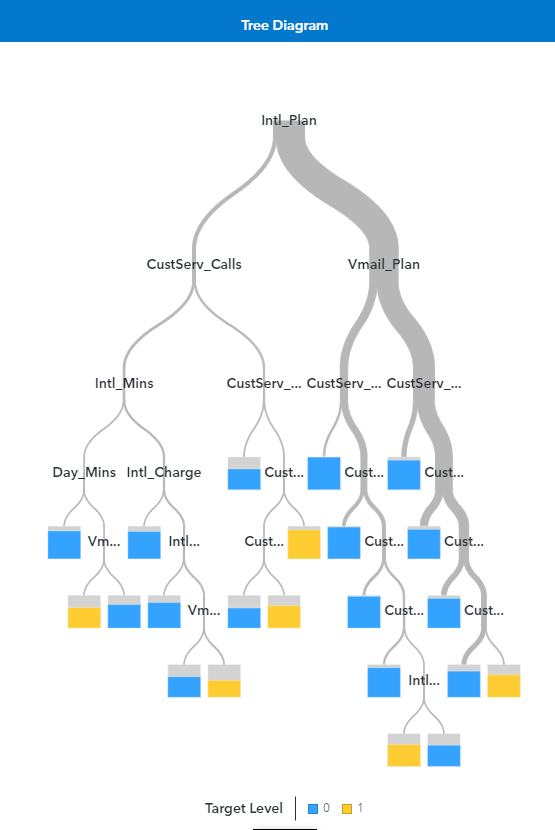
****

Figure 17: Maximal Tree Diagram

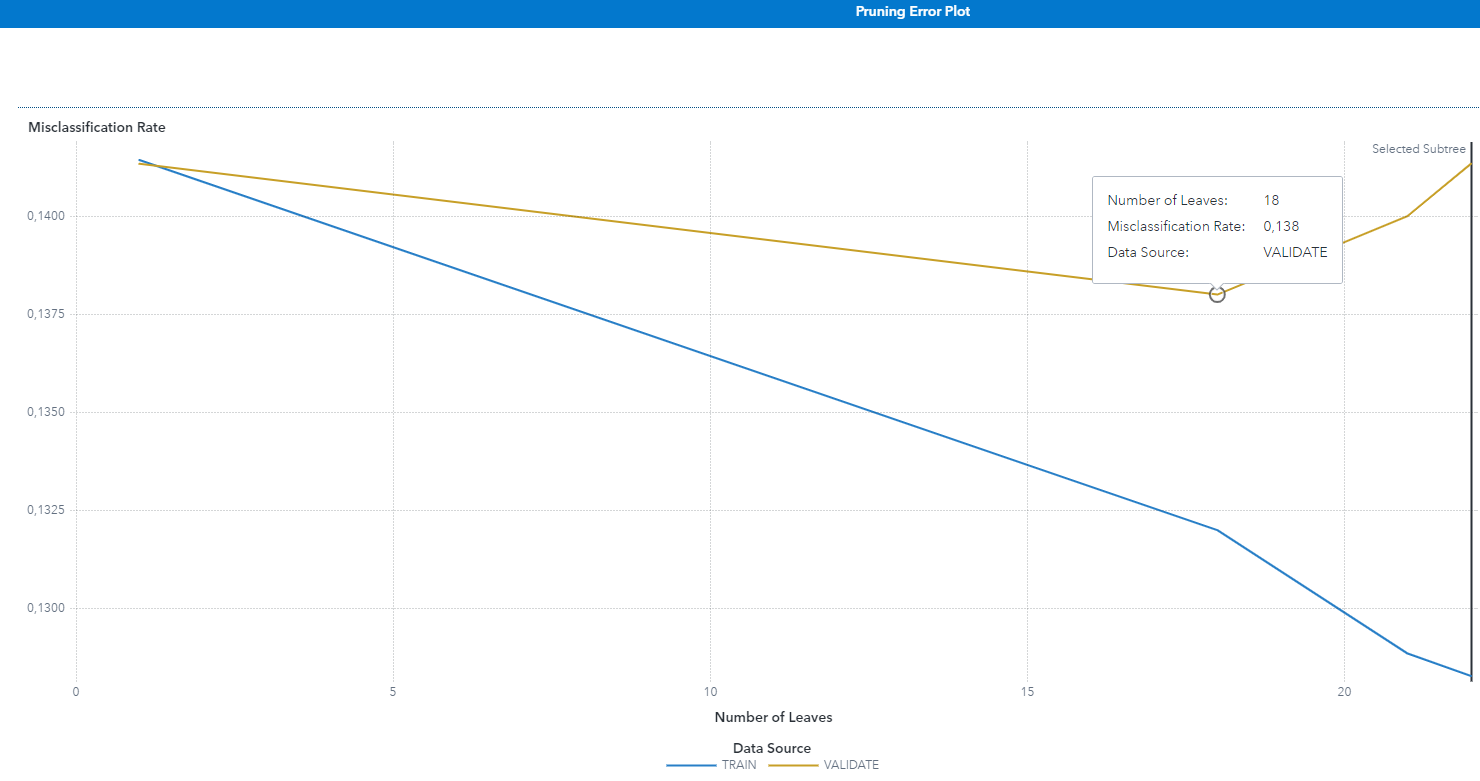
****

Figure 18: Subtree Assessment Plot- Pruning Error plot

**11) Optimal Decision Tree Structure and Performance**

The optimal tree (e.g., Figure 19), based on the subtree assessment plot where the Misclassification Rate is used to examine the performance, have 18 terminal leaves. This can be seen where the line representing the validation data set's misclassification rate (yellow line) reaches its lowest point before starting to increase, which indicates the point of optimal complexity for the tree model to generalize well to new data (e.g., Figure 20). From the plot, it can be seen that the training misclassification rate (blue line) consistently decreases as more leaves are added, which is expected because the model becomes more tailored to the training data. However, the key to selecting the optimal tree is to find a balance where the validation misclassification rate is minimized, indicating good performance on unseen data. This is the lowest point of the validation line before it plateaus or increases, suggesting that adding more leaves beyond this point only contributes to overfitting the model to the training data without improving its general predictive ability. For this we prune the tree at the point where the validation misclassification rate is minimized, for us is when the tree has 18 leaves. Pruning the tree to this size helps to prevent overfitting and ensures that the tree has the best chance of performing well on both the training data and unseen data.

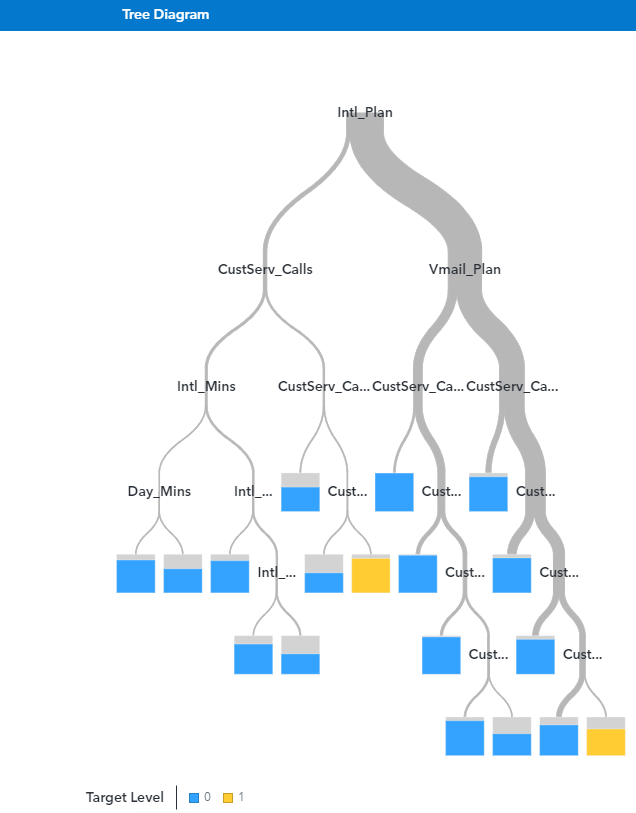
****

Figure 19: Optimal decision tree

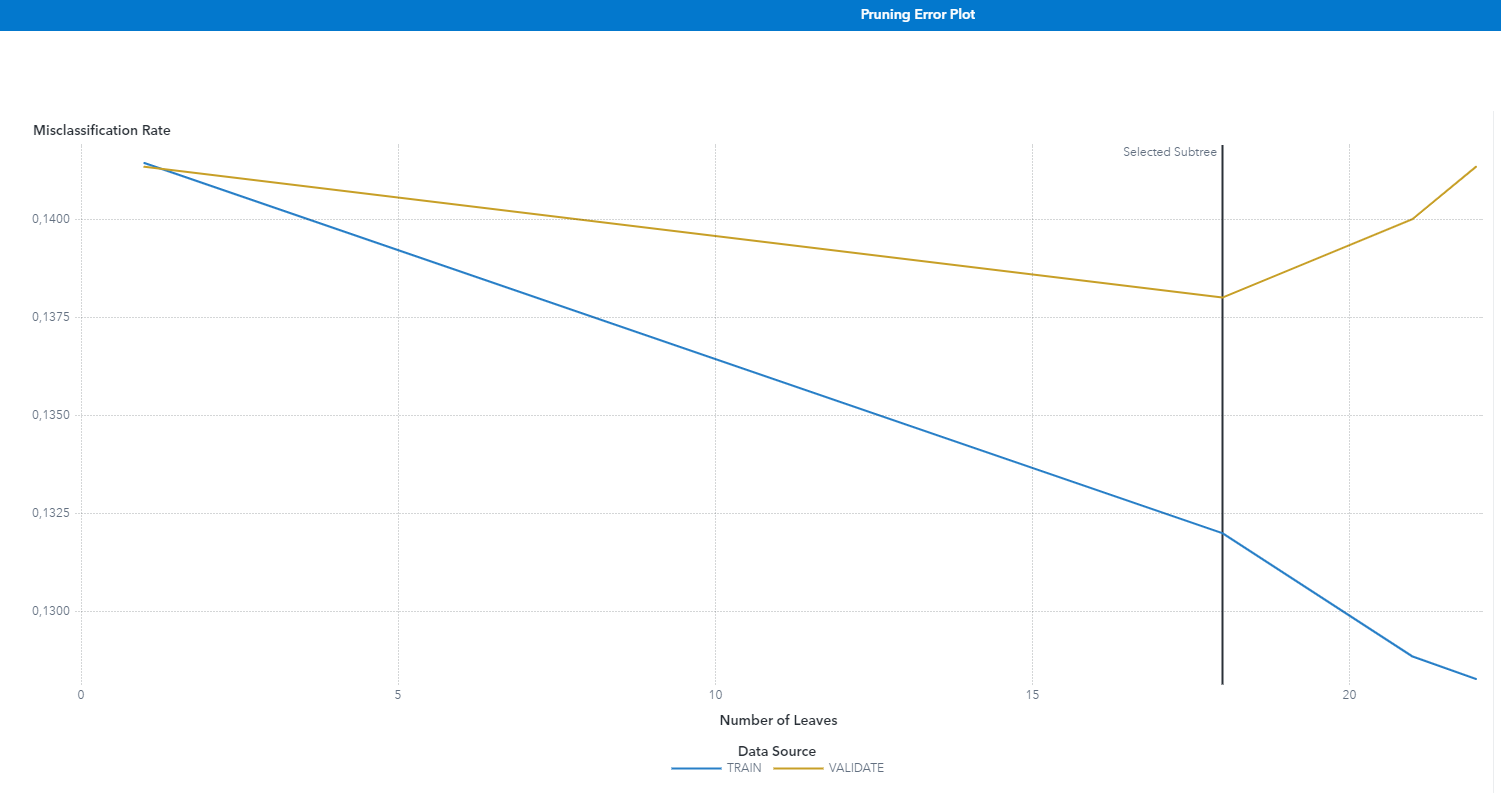
****

Figure 20: Subtree Assessment Plot- Pruning Error plot

**12) Description of the Decision Tree Model**

These leaves provide as an example of the model's subtlety, which takes into account several feature combinations, including voicemail plans, international plans, and customer service encounters, in order to estimate the probability of churn. The model illustrates how important some indicators are in predicting the churn likelihood, such as the quantity of customer support calls and plan kinds. The reduced tree offers a condensed, yet effective, representation for making decisions, attempting to strike a compromise between precision and the model's capacity to generalize to unknown inputs.There are particular configurations under which the Decision Tree Model functions. It uses the chi-square statistic to evaluate the variable separation for each possible split, utilizing the Bonferroni procedure to account for multiple comparisons with a significance level of 0.2. Each node in the binary tree structure branches into two, and the complexity of the tree is managed by restricting its growth to a maximum of 10 levels deep and requiring a minimum of 5 cases for a node to be qualified for additional division. In terms of missing data, the approach ensures a robust analysis by accounting for it within the split search itself, as opposed to rejecting it. Using the quantile approach, the model discretizes continuous variables into 50 bins. It uses a type of reduced-error pruning in that it chooses the most compact subtree that produces the lowest error upon validation for optimization. The algorithm makes this choice automatically, which expedites the process of choosing the most predictive and broadly applicable tree structure.

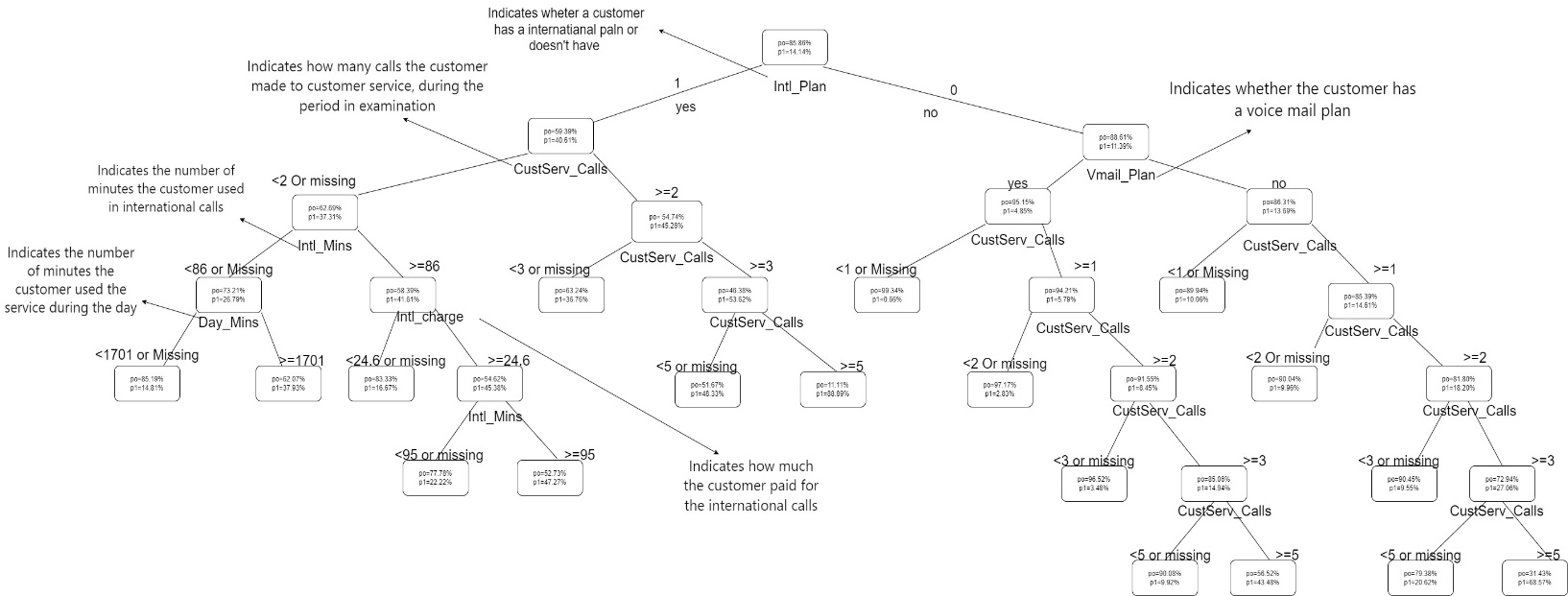


Figure 21: Optimal Decision tree

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Terminal Leaf -Node (ID:15) | Terminal Leaf -Node (ID:25) | Terminal Leaf -Node (ID:27) | Terminal Leaf -Node (ID:31) | Terminal Leaf -Node (ID:32) |
| Posterior Prob/ties | po= 85.19%  p1=14.81 % | po= 77.78%  p1=22.22 % | po=96.52 %  p1= 3.48% | po= 90.08%  p1= 9.92% | po= 56.52%  p1= 43.84% |
| Decision | 0 | 0 | 0 | 0 | 0 |
| Rules | ●Inti\_plan yes ●CustServ\_Calls<2 or missing  ● Intl\_Mins< 86 or missing  ●Day\_Mins< 1701 or missing | ●Inti\_plan yes ●CustServ\_Calls<2 or missing  ● Intl\_Mins>=86  ●Intl\_charge>=24.6  ● Intl\_Mins<95 or missing | ●Inti\_plan no  ●Vmail\_plan yes  ●CustServ\_Calls>=1  ●CustServ\_Calls>=2  ● CustServ\_Calls<3 or missing | ●Inti\_plan no  ●Vmail\_plan yes  ●CustServ\_Calls>=1  ●CustServ\_Calls>=2  ● CustServ\_Calls>=3  ● CustServ\_Calls<5 or missing | ●Inti\_plan no  ●Vmail\_plan yes  ●CustServ\_Calls>=1  ●CustServ\_Calls>=2  ● CustServ\_Calls>=3  ● CustServ\_Calls>=5 or missing |

Table 7: Table of Decision Tree Model for 5 Terminal Leaves

Three traits characterize the decision tree model. The first one is the posterior probability, which indicates the likelihood that each category will contain the observations that are enclosed in brackets. The second is the ultimate determination of the category into which the observations are assigned. The rules, which specify the threshold at which each Node splits until the Terminal Leaf is reached, are the last feature.

Terminal Leaf - Node (ID:15): Customers with an international plan who made less than 2 customer service calls or have missing data in this field (we don’t have), and who used less than 86 minutes for international calls or have missing data we don’t have), and who spent less of 1701 minutes or have missing data (we don’t have) on day calls are predicted to be non-churners. The posterior probabilities are po: 85.19% (posterior probability to be a customer non- churner) and p1:14.81% (posterior probability to be a customer churner). Our set threshold of 16.6%, which we use as a cut-off point to decide if a customer might leave. We expect that these customers won't churn, so they wouldn't be targeted for additional retention efforts.

Terminal Leaf - Node (ID:25): Customers with an international plan who made less than 2 customer service calls, used 86 or more minutes for international calls, and were charged more than or equal to $24.6, and who used less than 95 minutes for international calls or have missing data (we don’t have) are more likely to not churn. The posterior probabilities are po: 77.78% (posterior probability to be a customer non- churner) and p1:22.22% (posterior probability to be a customer churner). Our set threshold of 16.6%, which we use as a cut-off point to decide if a customer might leave. We expect that these customers won't churn, so they wouldn't be targeted for additional retention efforts.

Terminal Leaf - Node (ID:27): Customers without an international plan but with a voicemail plan, who made more or equal of 2 calls to customer service but less than 3 or have missing data in this field (we don’t have), during the period in examination are more likely to not churn. The posterior probabilities are po: 96.52% (posterior probability to be a customer non- churner) and p1:3.48% (posterior probability to be a customer churner). Our set threshold of 16.6%, which we use as a cut-off point to decide if a customer might leave. We expect that these customers won't churn, so they wouldn't be targeted for additional retention efforts.

Terminal Leaf - Node (ID:31): Customers without an international plan but with a voicemail plan, making between 3 and 4 customer service calls, or have missing data in this field (we don’t have), during the period in examination, are predicted not to churn. The posterior probabilities are po: 90.08% (posterior probability to be a customer non- churner) and p1:9.92% (posterior probability to be a customer churner). Our set threshold of 16.6%, which we use as a cut-off point to decide if a customer might leave. We expect that these customers won't churn, so they wouldn't be targeted for additional retention efforts.

Terminal Leaf - Node (ID:32): Customers without an international plan but with a voicemail plan, making 5 or more customer service calls or missing data in this field, are predicted not to churn The posterior probabilities are po: 56.52% (posterior probability to be a customer non - churner) and p1:43.84 % (posterior probability to be a customer churner). Our set threshold of 16.6%, which we use as a cut-off point to decide if a customer might leave. We expect that these customers won't churn, so they wouldn't be targeted for additional retention efforts.

**13) Management Summary of the Decision Tree Analysis**

Node (ID:15): Customers with an international plan who made less than 2 customer service calls or have missing data in this field (we don’t have), and who used less than 86 minutes for international calls or have missing data we don’t have), and who spent less of 1701 minutes or have missing data (we don’t have) on day calls, are very likely to stay with us. Their chance of leaving is only around 14.81%, which is low, so we don't need to worry much about them.

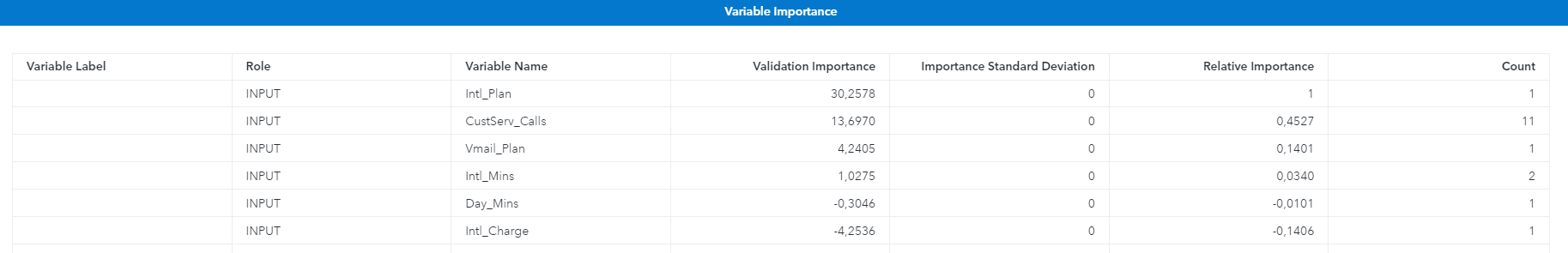
Node (ID:25): Customers with an international plan who made less than 2 customer service calls, used 86 or more minutes for international calls, and were charged more than or equal to $24.6, and who used less than 95 minutes for international calls or have missing data (we don’t have), also show a low chance of leaving, about 22.22%. They're quite content with our service, so we don't need to put extra effort into keeping them.

Node (ID:27): Customers without an international plan but with a voicemail plan, who made more or equal of 2 calls to customer service but less than 3 or have missing data in this field (we don’t have), during the period in examination are extremely unlikely to churn, with over 96.52% probability of staying. They're happy customers, and we're doing well with them.

Node (ID:31): S Customers without an international plan but with a voicemail plan, making between 3 and 4 customer service calls, or have missing data in this field (we don’t have), during the period in examination are predicted to stick around, with a 90% chance of staying loyal to our service. They seem to get what they need from us.

Node (ID:32): Customers without an international plan but with a voicemail plan, making 5 or more customer service calls or missing data in this field, are a bit more at risk of churning, with a churn probability of 43.48%. They’re more likely to stay than go.

International Plan: Having an international plan is a significant factor. Customers with such a plan are less likely to churn, especially if they also have fewer interactions with customer service and use a moderate amount of international minutes. , Customer Service Calls: The number of customer service calls is a strong indicator. Those who call less are more likely to be satisfied and stay with the service. In contrast, an increasing number of calls could signal potential issues and dissatisfaction. . Usage of Services: The amount of time customers spend on calls during the day, particularly for international calls, also influences churn. Customers with higher usage seem more content with the services and thus show lower churn probability. , Voicemail Plan: The presence of a voicemail plan correlates with a lower likelihood of churning, suggesting that additional service features may enhance customer loyalty . (e.g., Figure 22)

****Figure 22: Variables Importance

**14) Model Comparison: Scoring and Ranking Analysis**

Τhe cumulative % response graph (e.g., Figure 25) for our four models is depicted below. We can observe that if we solicit the 20% of the best customers according to the probability that the Optimal Tree model gives them to be  churners, the 32.589% of this 20% will be churners. As well as, if we solicit the 100% of the best customers according to the probability that the Optimal Tree model gives them to be churners, the 14.133% of this 100% will be churners .

We can observe that if we solicit the 20% of the best customers according to the probability that the Logistic Regression model gives them to be churners, the 35% of this 20% will be churners. As well as, if we solicit the 100% of the best customers according to the probability that the Logistic Regression model gives them to be churners, the 14.133% of this 100% will be churners.

We can observe that if we solicit the 20% of the best customers according to the probability that the Maximal Tree model gives them to be churners, the 32.589% of this 20% will be churners. As well as, if we solicit the 100% of the best customers according to the probability that the Maximal Tree model gives them to be churners, the 14.133% of this 100% will be churners.

For the fourth model we can observe that if we solicit the 20% of the best customers according to the probability that the Neural Network model gives them to be churners, the 23.333% of this 20% will be churners. As well as, if we solicit the 100% of the best customers according to the probability that the Neural Network model gives them to be churners, the 14.1333% of this 100% percent will be churners.

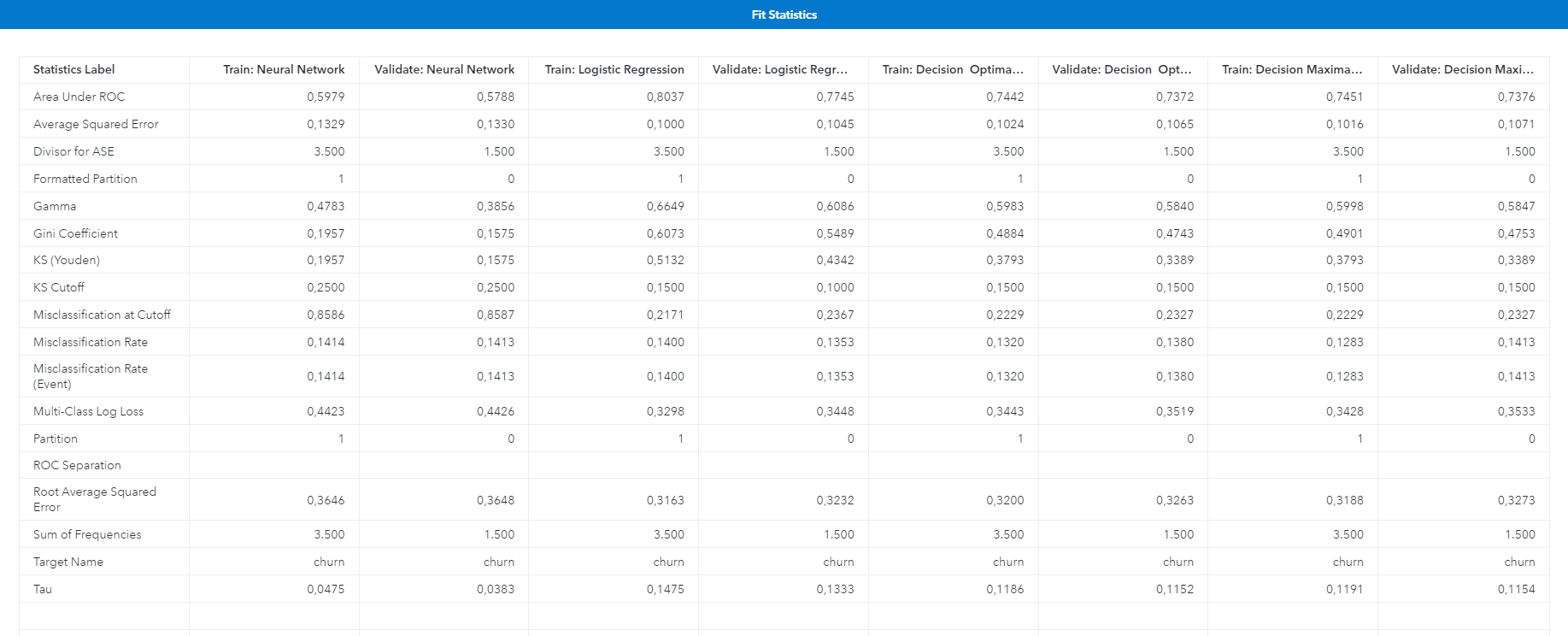
****

Figure 23: Train and Validate error for each model

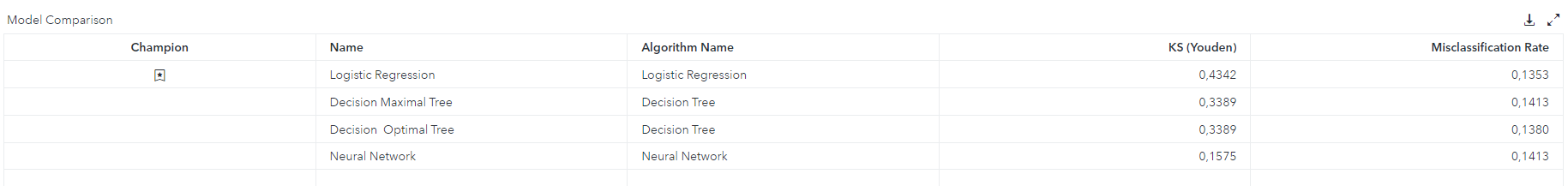
****

Figure 24: Misclassification Rate for each model

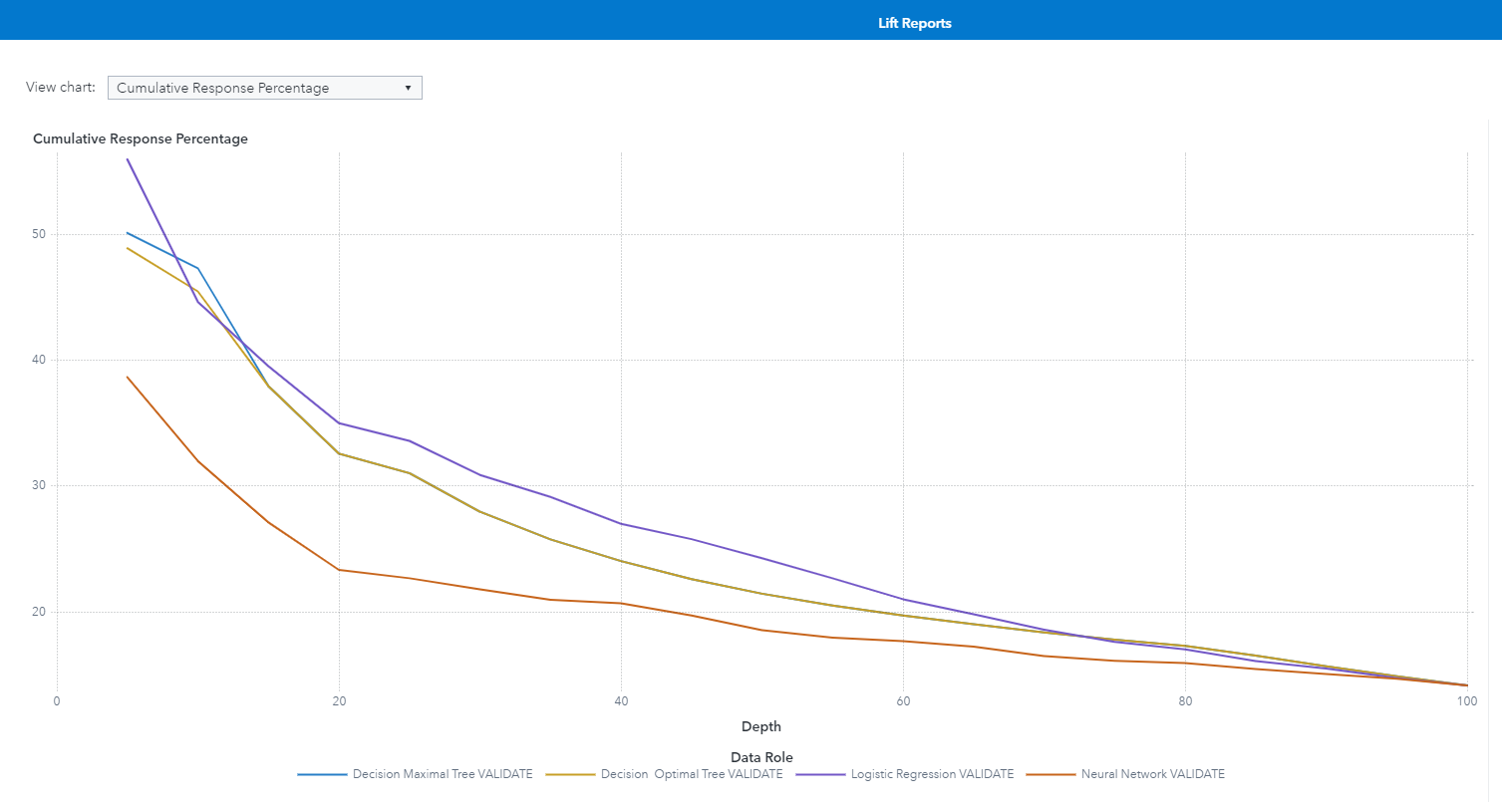
****

Figure 25: Cumulative Response Percentage

**15) Response Rate Analysis**

The  % response chart(e.g., Figure 26) for the validation set for our models is depicted. This graph is constructed by sorting the data according to their probabilities to be churners and as a next step by splitting them in buckets of equal size. In the x axis are represented those buckets using the sorted data. More specifically, we will examine the 5th bucket that is between 20%-25% of the best customers according to the constructed models.

So, if we solicit the fifth (5th) bucket (20%-25%) of the best customers according to the probability that the Optimal Tree model gives them to be churners, the 24.815% will be churners.

In addition, if we solicit the fifth (5th) bucket (20%-25%) of the best customers according to the probability that the Logistic Regression model gives them to be churners, the 28% of this bucket will be churners.

If we solicit the fifth (5th) bucket (20%-25%) of the best customers according to the probability that the Maximal Tree model gives them to be churners, the 24.815% of this bucket will be churners.

Finally, if we solicit fifth (5th) bucket (20%-25%) of the best customers according to the probability that the Neural Network model gives them to be churners, the 20% of this bucket will be churners.



Figure 26: Response Percentage

**16) Cumulative Lift Analysis**

The cumulative lift(e.g., Figure 27) for our created models is depicted. To be able to understand the graph an explanation for the 20% point in the x axis for each model will be provided.

To begin with, if we solicit the 20% of the best customers according to the probability that the Optimal Tree model gives them to be churners, we will capture 2.3058 times more churners than if we did the same job without the model i.e., at random.

In addition, if we solicit the 20% of the best customers according to the probability that the Maximal Tree model gives them to be churners, we will capture 2.3058 times more churners than if we did the same job without the model i.e., at random.

Furthermore, if we solicit the 20% of the best customers according to the probability that the Logistic Regression model gives them to be churners, we will capture 2.4764 times more churners than if we did the same job without the model i.e., at random.

Finally, if we solicit the 20% of the best customers according to the probability that the Neural Network model gives them to be churners, we will capture 1.6509 times more churners than if we did the same job without the model i.e., at random.

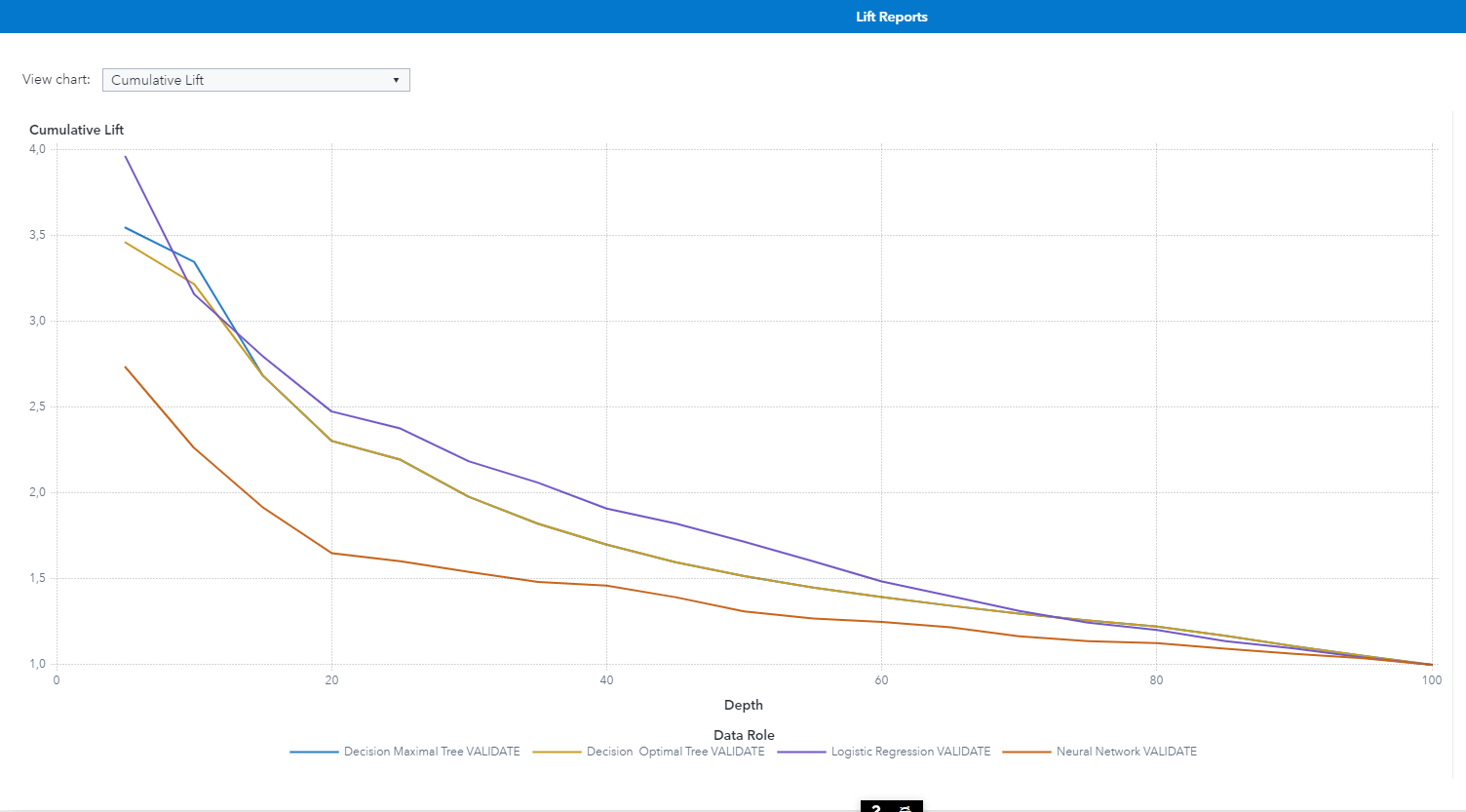


Figure 27: Cumulative Lift

**17) Captured Response Analysis**

The cumulative % captured response (e.g., Figure 28) for our models is depicted. To be able to understand the graph an explanation for the 40% point in the x axis for each model will be provided.

Firstly, if we solicit the 40% of the best customers according to the probability that the Optimal Tree model gives them to be churners, we will capture the 68.021% of all the churners of the whole validation data set.

In addition, if we solicit the 40% of the best customers according to the probability that the Maximal Tree model gives them to be churners, we will capture the 68.021% of all the churners of the whole validation data set.

Furthermore, if we solicit the 40% of the best customers according to the probability that the Logistic Regression model gives them to be churners, we will capture the 76.415% of all the churners of the whole validation data set.

Finally, if we solicit the 40% of the best customers according to the probability that the Neural Network model gives them to be churners, we will capture the 58.491% of all the churners of the whole validation data set.

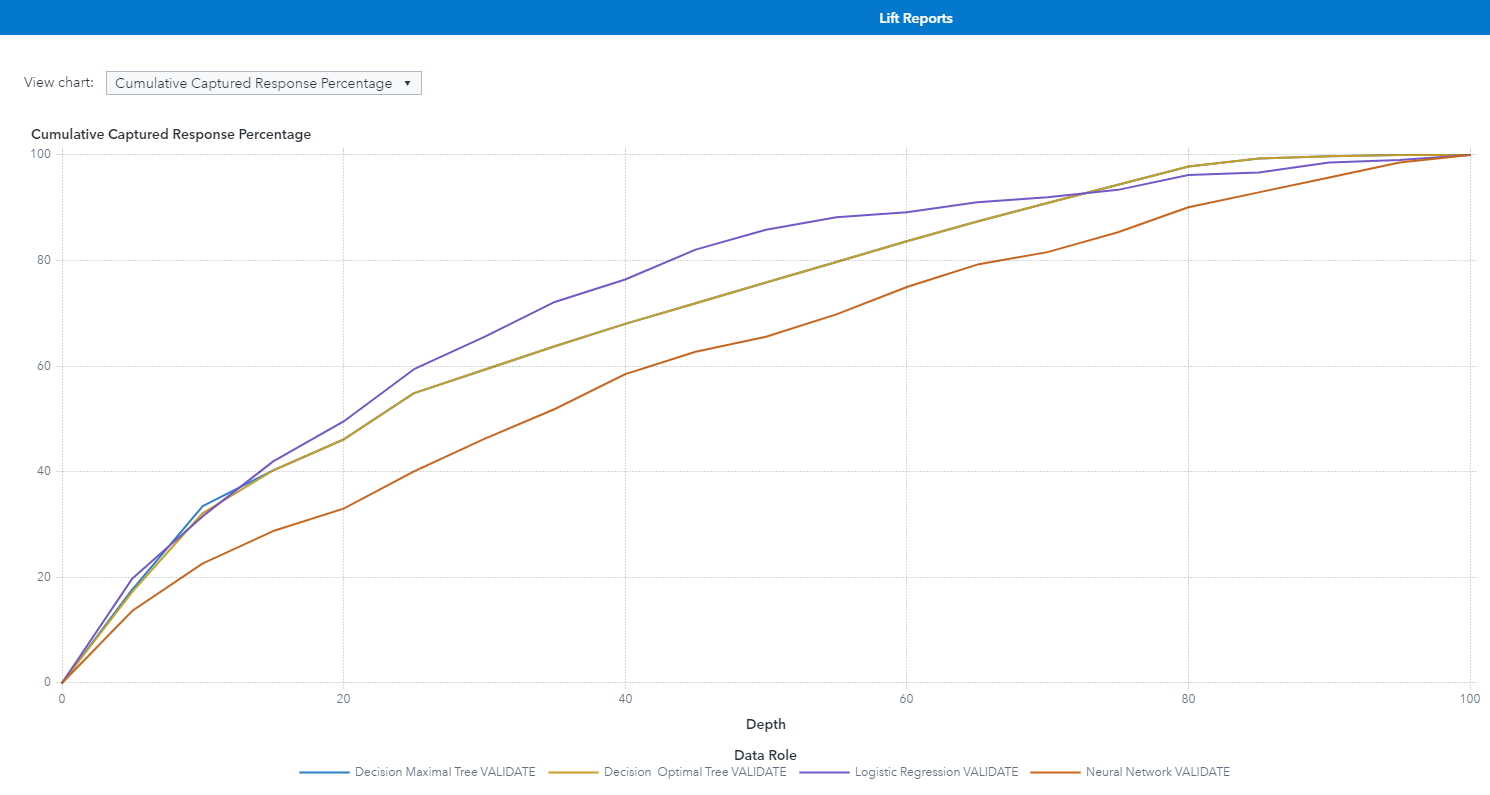


Figure 28: Cumulative Captured Response Percentage

We selected as the final model, the model with the best performance among the Maximal tree, Optimal Tree, Logistic Regression and Neural Network, the Logistic Regression model. This preference had to do with the fact that the logistic regression model had the lowest misclassification rate (0.135). Also, from the previous plots we understood that it performed better from the other models. Next, the process flow (e.g., Figure 29) is depicted that was created in the SAS software.

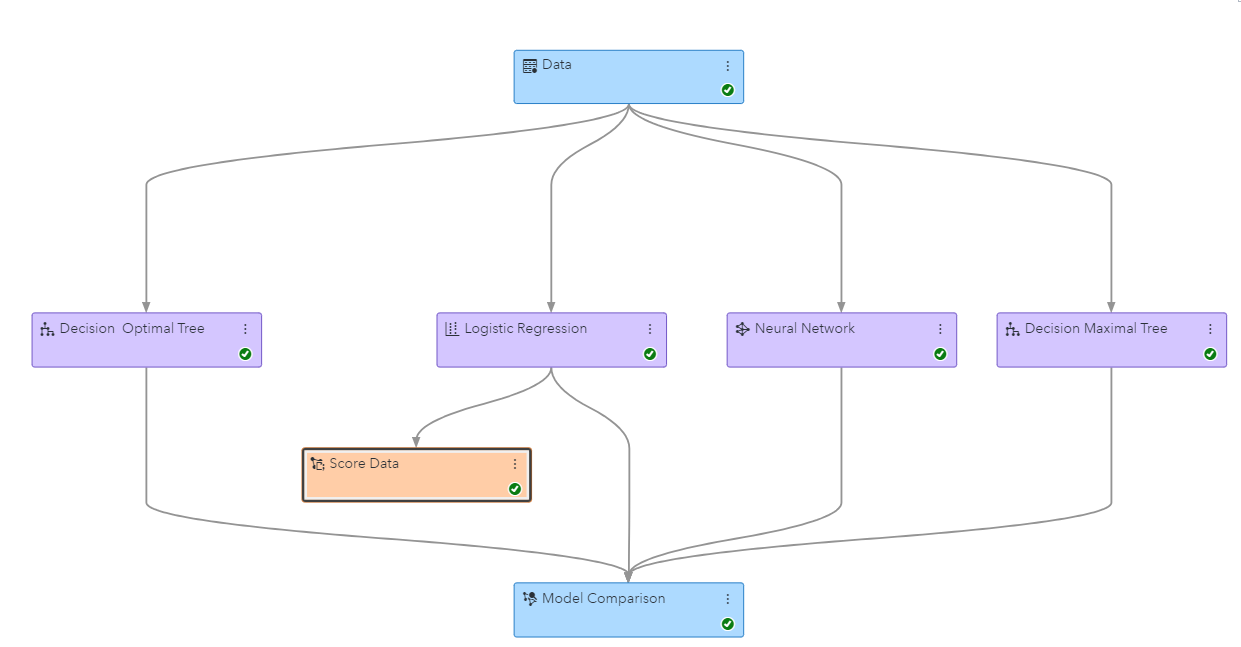


Figure 29: Process Flow

**18) Post-Deployment Churn Prediction Results**

The total number of customers in the “telco\_data\_apr\_sep” is equal to 1,884. From those customers the 1,365 of them are predicted by the best model as non-churners and 519 of them as churners. Below is presented a bar chart (e.g., Figure 30) using SAS visual analytics to show the prediction of customers to be churners and non-churners.

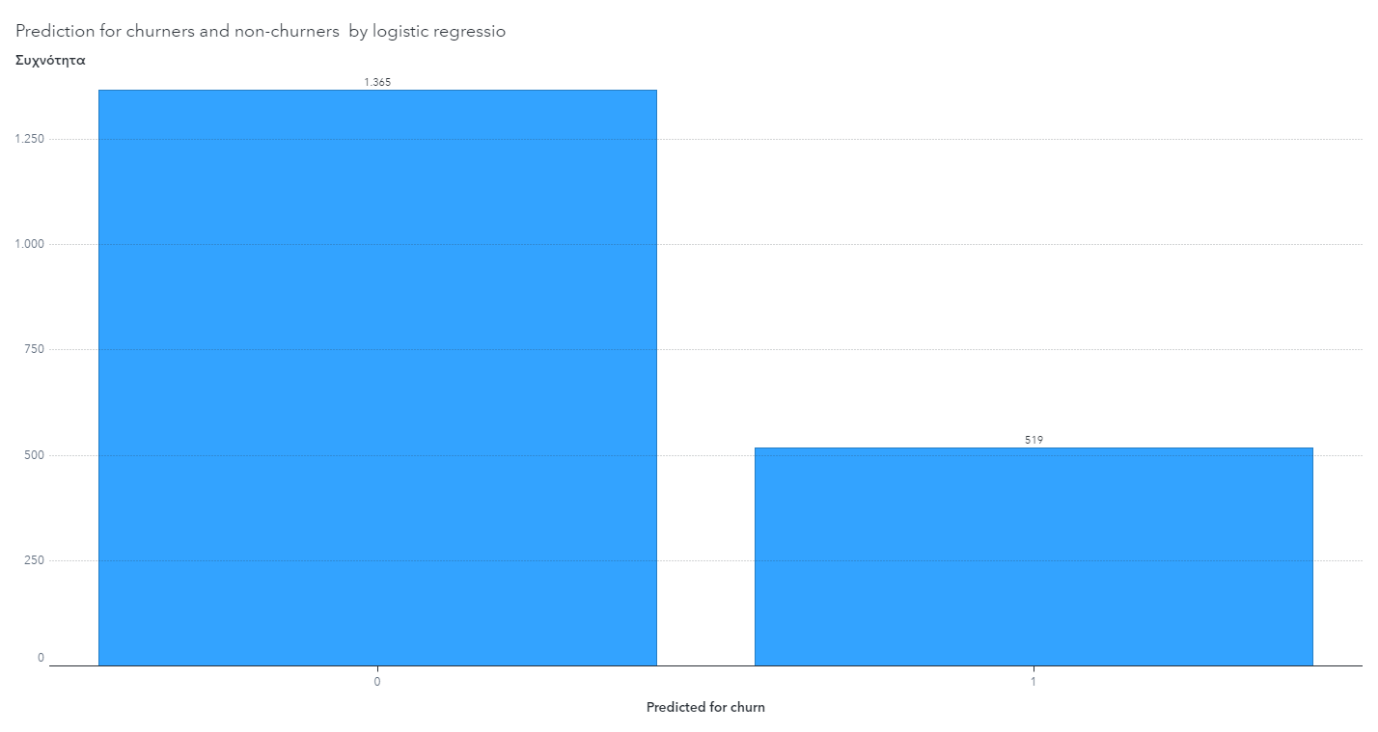


Figure 30: Barchart for prediction of churners and non-churners

**19) Range of Predicted Churn Probabilities**

Based on our model the biggest probability of being a churner assigned to a customer is 0.99 or 99%, whereas the smallest probability that a customer is assigned by the model as a churner is 0.166 or 16.6%.

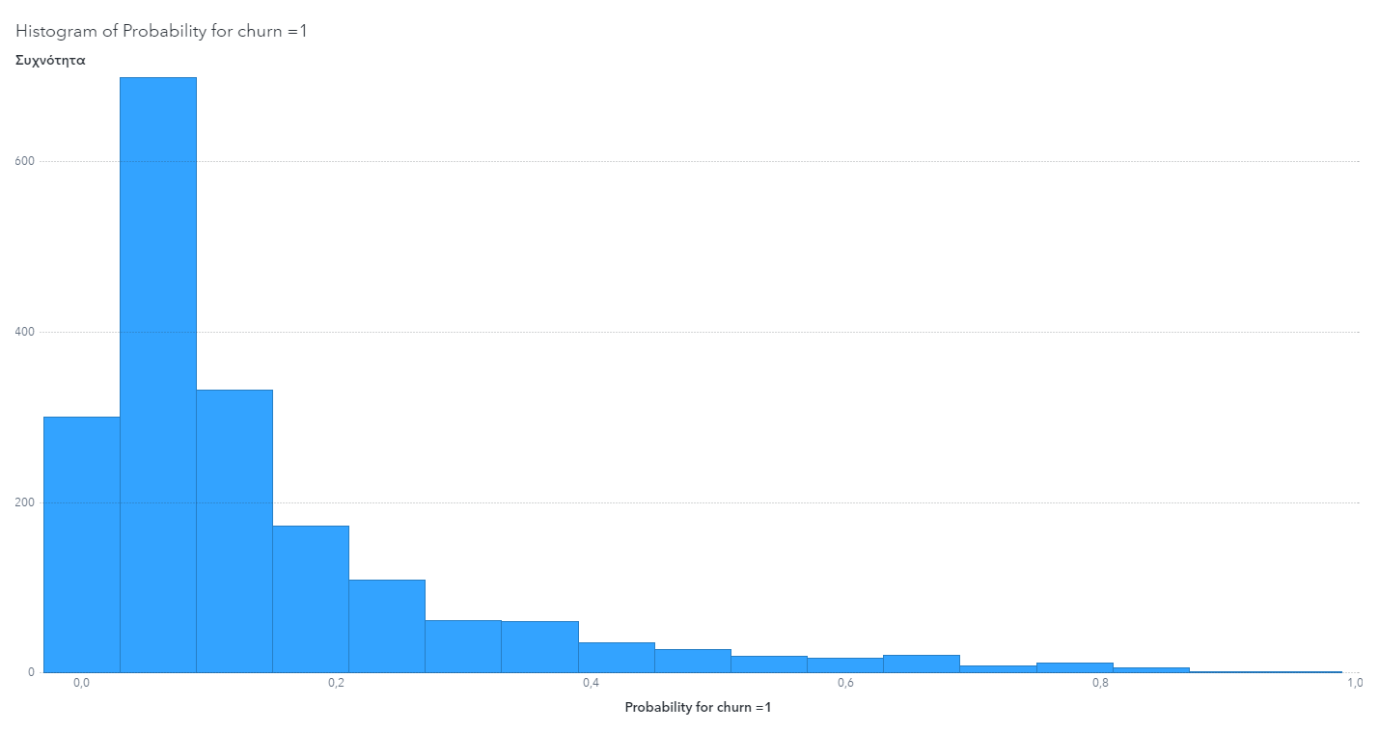
****

Figure 31: Histogram of probability of being a churner

**20) Case Study of Predictive Outcome for an Individual Customer**

Our first choice was the customer with Phone number 371-7191. This customer has a probability of being a churner based on the logistic regression model equal to 0.04 (e.g., “Probability for churn = 1” column in the score data set), which is lower than the threshold (cut-off point) that is 0.166, so the customer is classified by the model as non-churner.

Our second choice was the customer with Phone number 375-9999. This customer has a probability of being a churner based on the logistic regression model equal to 0.519 (e.g., “Probability for churn = 1” column in the score data set), which is higher than the threshold (cut-off point) that is 0.166, so the customer is classified by the model as churner.

**Appendix**

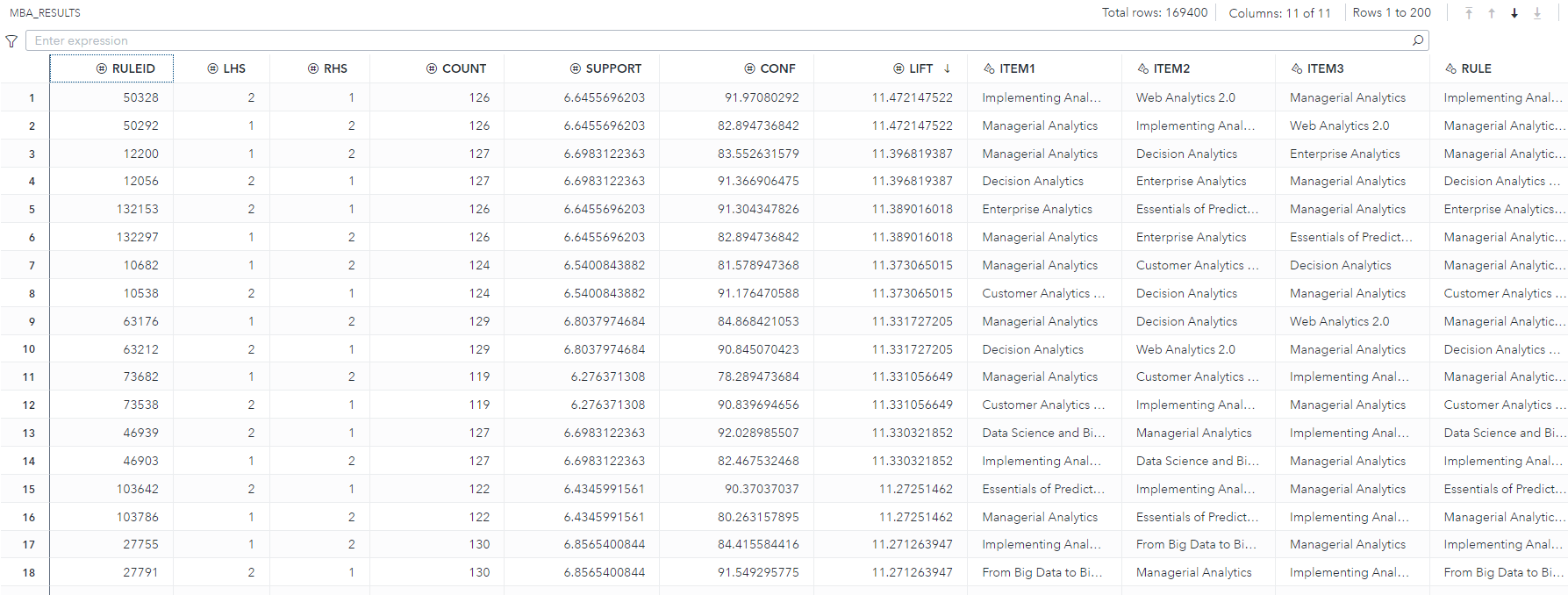
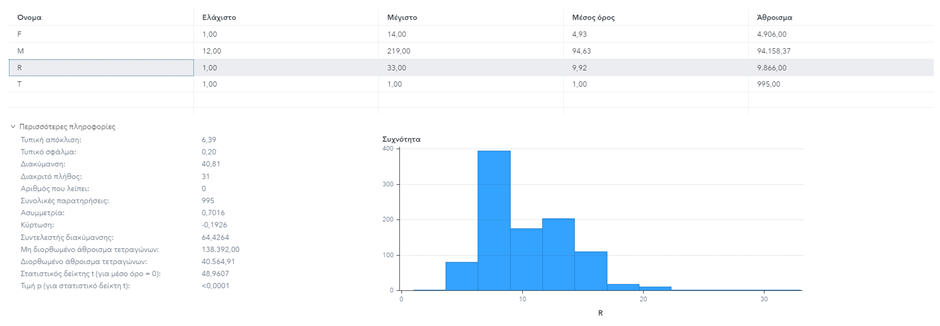
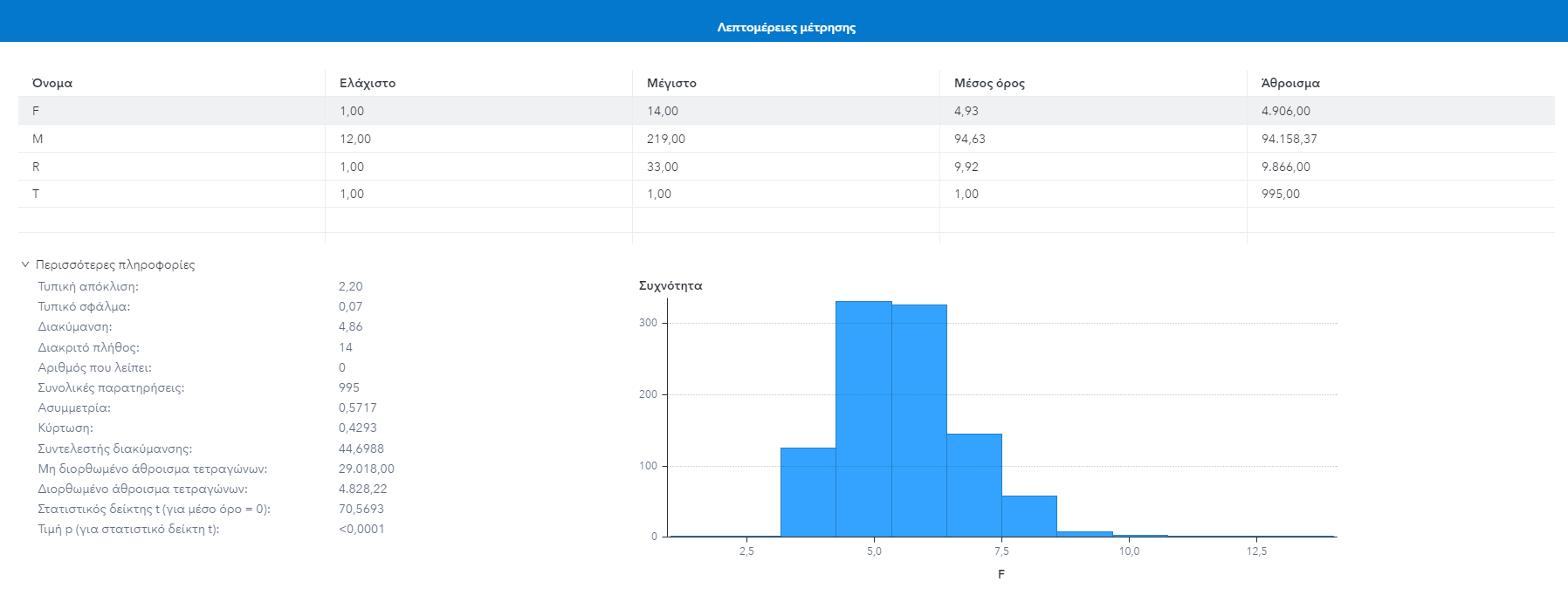
****

Table 3: Market Basket Analysis



****

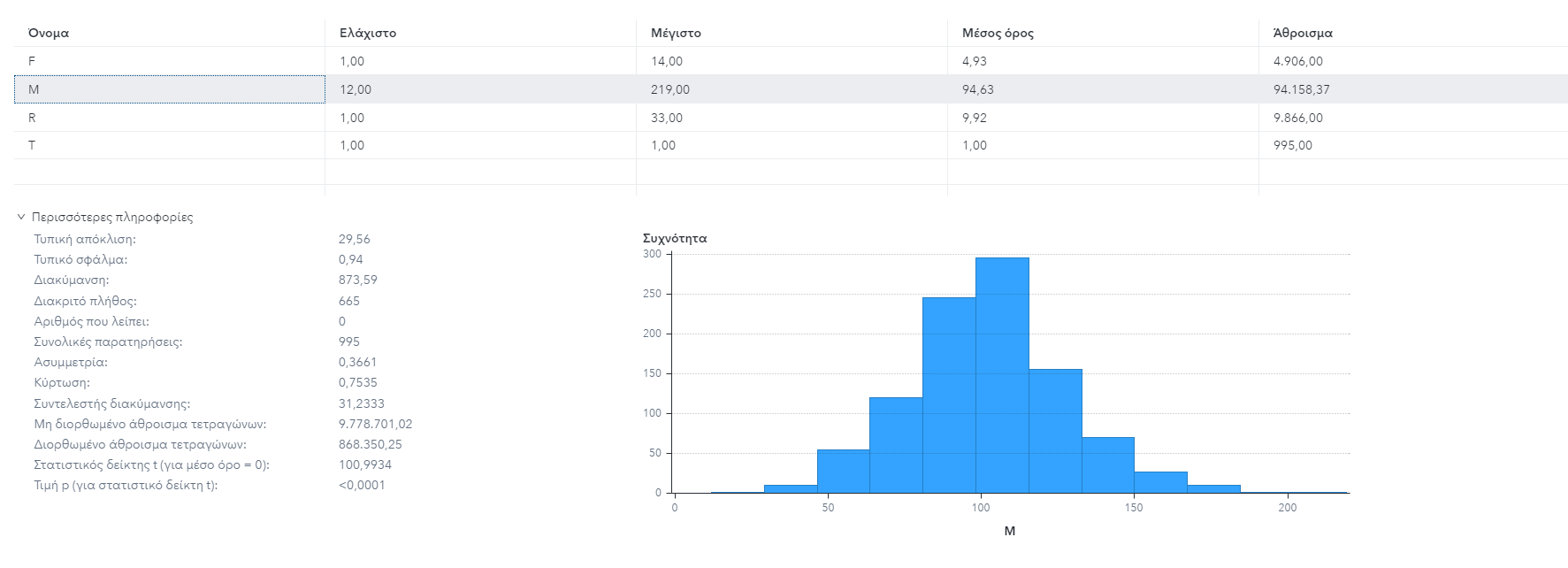
****

Figure 4: RFM histograms with metrics details

****

Table 5: Segment Names by Age Range by Payment Method