ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS
BUSINESS ANALYTICS Master of Science

Analytics Practicum I (SAS Project)

M.Sc.: Business Analytics

LAKKAS-PYKNIS EVANGELOS
Registration number: f2822306
E-mail: eva.lakkaspyknis@aueb.gr

MESOLORA STAMATOULA-GERASIMOULA
Registration number:
f2822308E-mail:
sta.mesolora@aueb.gr

Contents

Case Study 1: Market Basket Analysis

Executive Summary

Our goal is to find a way to maximize the income that comes from the sales of Books related to the category of "Business Analytics" by finding which are the best cross-selling opportunities and exploiting them by optimizing the advertising and pricing policy of the books that belong in this category. To achieve the previously referred goal we evaluated the strength of multiple associations between the books, in simpler words we want to find which are the books that a customer is most likely to buy based on the knowledge which book(s) he previously bought. So, we used the Market Basket Analysis method to find and evaluate various combinations/associations of books based on a dataset that contained 56 distinct books and over 19,000 book purchases that were made in our store in the past. The basic decision support rule in our analysis was the Lift metric which expresses how more likely it is for a customer to buy a certain book that we want to sell knowing what books he/she already bought when compared to a customer picked at random without using a model. The calculations for our analysis were made with SAS Viya.

Indicatively we will refer to some results of our analysis to show why we strongly recommend that the MBA technique should become a main process in the decision-making of our organization and propose some tactics based on these results. We found out that customers who bought Managerial Analytics are 11.47 times more likely to buy Implementing Analytics and Web Analytics 2.0 than a random customer. So, our organization should advertise Implementing Analytics and Web Analytics 2.0 more to people who already bought Managerial Analytics. We will also analyze 3 cases more extensively below in our report.

Question 2:

To create the following Bar Chart, we used the SAS Visual Analytics tool. As we can see this is a horizontal bar plot that shows the sales of each distinct book in the Business Analytics category in units on the y-axis, we can find the name of each book and on the x-axis, we can see the sales (to find the sales of each book we counted the frequency of each title in the transactions data that we had available). Also, in order not to lose any information we put a label with the exact number of the sales of each book inside each bar. Indicatively we can see that the book with the most sales is Data Science and Business Analytics and the one with the least is Managerial Analytics.

(Please note that you can also find the bar chart in a more readable format in the corresponding PDF file that we have uploaded)

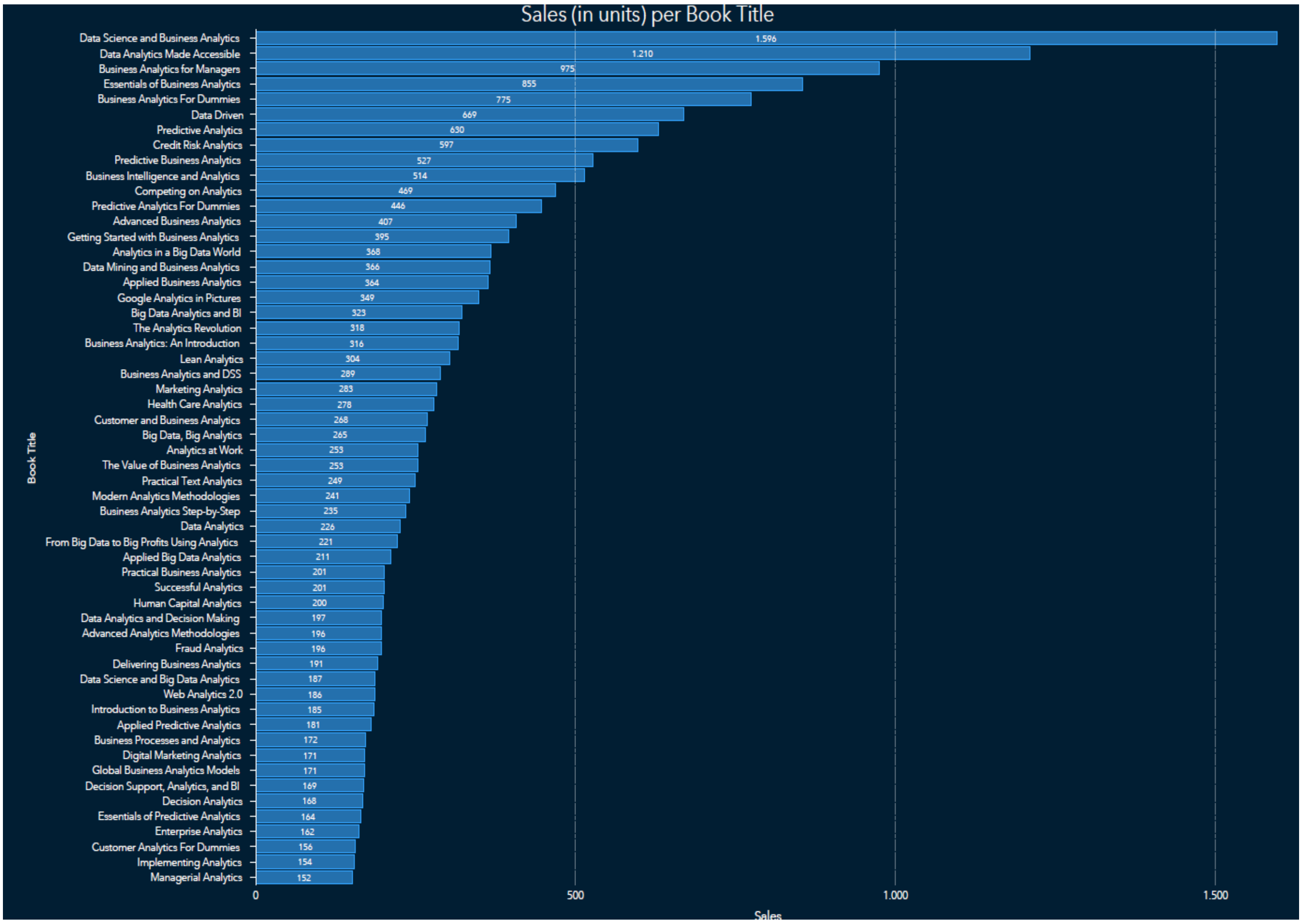


Figure 1: Sales of each book

Question 3:

Which two books should the store advertise to customers who bought/ are searching to buy only one of the following:

- Managerial Analytics
- Implementing Analytics
- Customer Analytics for Dummies
- Enterprise Analytics

We explored which are the two books that our store should advertise most to customers that we know that have bought one of the previous books alone. As we mentioned in the executive summary the support tool for this decision was Lift. So, in a technical manner we sorted all the association rules that contained only one of these books in the Left-Hand Side of the rule based on the Lift value in a descending order. So, we found the book combinations with the biggest lift we want to find the combinations with this metric value because we want to have as much better performance as possible from the results that we would have from luck and that is what lift counts. By following these steps, we created the table given below.

Book that the customer has already bought	Books that should be advertised	Lift	Association Rule
Managerial Analytics	Implementing Analytics, Web Analytics 2.0	11.47	Managerial Analytics => Implementing Analytics & Web Analytics 2.0
Implementing Analytics	Data Science and Big Data Analytics, Managerial Analytics	11.33	Implementing Analytics => Data Science and Big Data Analytics & Managerial Analytics
Customer Analytics for Dummies	Decision Analytics, Enterprise Analytics	11.19	Customer Analytics for Dummies => Decision Analytics & Enterprise Analytics
Enterprise Analytics	Customer Analytics for Dummies and Managerial Analytics	11.01	Enterprise Analytics => Customer Analytics for Dummies & Managerial Analytics

A way to explain lift better in a business manner is in a case that a customer who has already bought Managerial Analytics is 11.47 more likely to buy Implementing Analytics and Web Analytics 2.0 than a customer picked at random.

This insight can help us optimize the advertisement campaign for these two books (and for most books in general) by advertising them more intensely to the customers who already bought Managerial Analytics and less intensely to the rest of the customers which means less advertisement costs but better results. Similarly, customers who bought Implementing Analytics are 11.33 times more likely to buy Data Science and Big Data Analytics and Managerial Analytics than a random customer and customers who bought Customer Analytics for Dummies are 11.19 times more likely to buy Decision Analytics and Enterprise Analytics than a random customer. In the same way, we can interpret the lift for Enterprise analytics.

In a technical approach we can say that lift evaluates the strength of an association rule between two or more products, in our case the association rules that we evaluated contain 2 or 3 books, an association rule is a formulation that in our example can take for example this form Customer Analytics for Dummies => Decision Analytics & Enterprise Analytics and tries to find a potential opportunity of selling books based on the fact that a customer bought a specific book. Lift is also mathematically derived from the following formula:

$$Lift = \frac{Confidence}{Expected\ Confidence}$$

Where,

Confidence: the probability that a customer will buy the books that are found on the Right-Hand Side of the association rule (in the example that we wrote above these books are Decision Analytics & Enterprise Analytics) given that he has bought the book found on the Left-Hand side of the association rule (in our example Customer Analytics for Dummies)

Expected Confidence: the probability that a customer has the books that are found on the Right-Hand Side of the association rule.

In our Market Basket Analysis, we excluded association rules that had a confidence level greater or equal to 10% to find the most useful associations. As we stated earlier to find the more useful rules for the books that interested us in this question, we first sorted all of them in a descending order based on their lift and then we filtered the LHS of the rule to find the ones that had only one of the books that we want to analyze and pick the one with the biggest lift value.

Below we provide a set of pictures from the table rules that we derived from our analysis.

CASUSER.MBA_RESULTS Columns: 11 of 11 | Rows 1 to 200 (filtered)

Enter expression

LHS = 1 × | UPPER(ITEM1) = 'MANAGERI... ×

	RHS	COUNT	SUPP...	CONF	LIFT ↓	IT...	ITEM2	ITEM3	RULE
1	2	126	6.6455696203	82.894736842	11.472147522	Managerial Analytics	Implementing Analytics	Web Analytics 2.0	Managerial Analytics ==> Implementing Analytics & Web Analytics 2.0
2	2	127	6.6983122363	83.552631579	11.396819387	Managerial Analytics	Decision Analytics	Enterprise Analytics	Managerial Analytics ==> Decision Analytics & Enterprise Analytics
3	2	126	6.6455696203	82.894736842	11.389016018	Managerial Analytics	Enterprise Analytics	Essentials of Predictive Analytics	Managerial Analytics ==> Enterprise Analy Esse

Table 1: Top Association Rules for Customers Who Bought Managerial Analytics

CASUSER.MBA_RESULTS Columns: 11 of 11 | Rows 1 to 200 (filtered)

Enter expression

LHS = 1 × | UPPER(ITEM1) = 'IMPLEMEN... ×

	RHS	COUNT	SUPP...	CONF	LIFT ↓	IT...	ITEM2	ITEM3	RULE
1	2	127	6.6983122363	82.467532468	11.330321852	Implementing Analytics	Data Science and Big Data Analytics	Managerial Analytics	Implementing Analytics ==> Data Science and Big Data Analytics & Managerial Analytics
2	2	130	6.8565400844	84.415584416	11.271263947	Implementing Analytics	From Big Data to Big Profits Using Analytics	Managerial Analytics	Implementing Analytics ==> From Big Data to Big Profits Using Analytics & Managerial Analytics

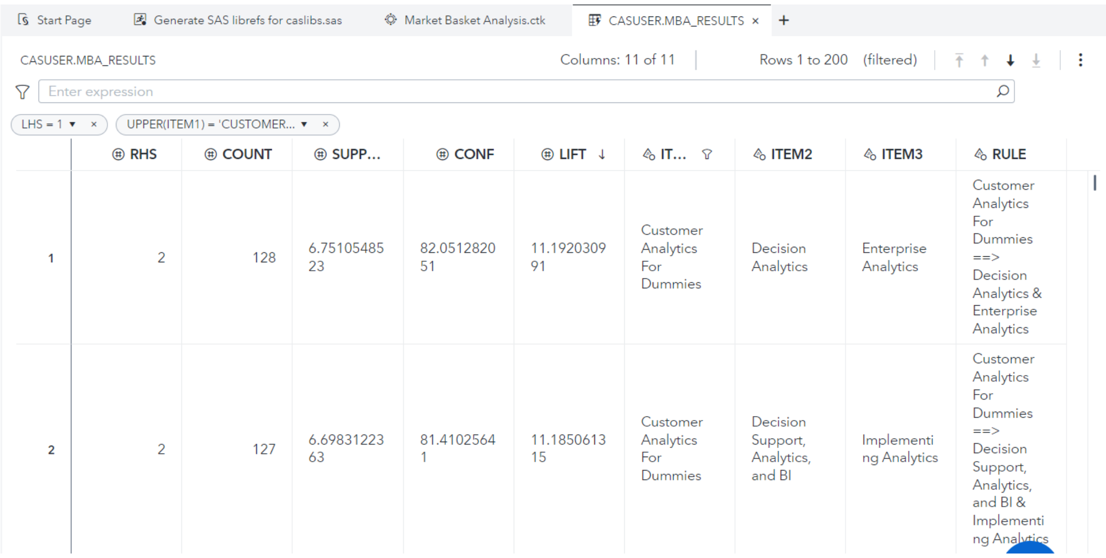
CASUSER.MBA_RESULTS Columns: 11 of 11 | Rows 1 to 200 (filtered)

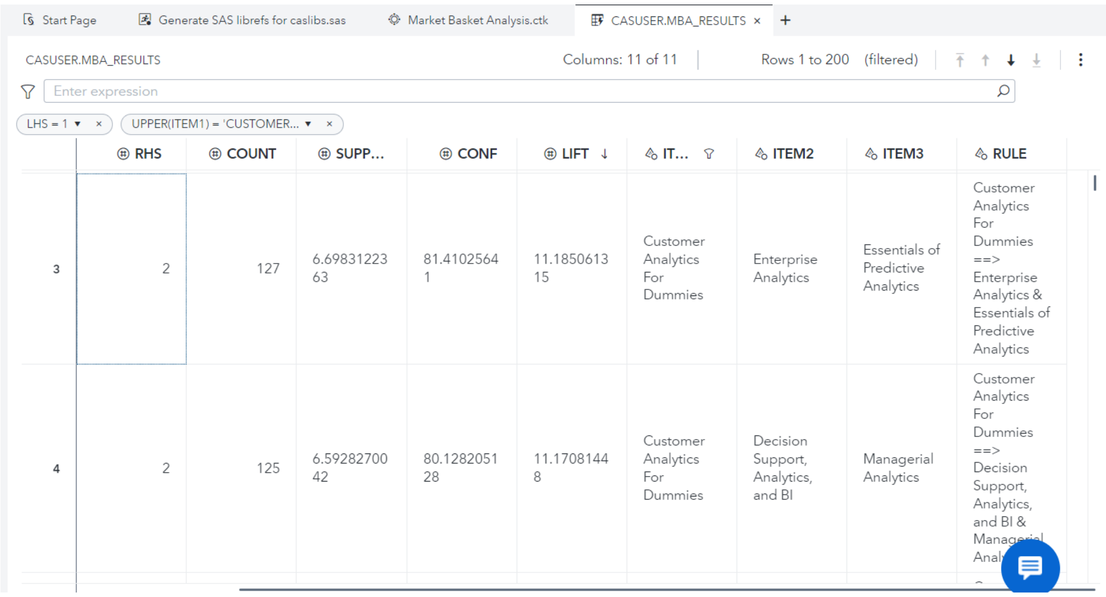
Enter expression

LHS = 1 × | UPPER(ITEM1) = 'IMPLEMEN... ×

	RHS	COUNT	SUPP...	CONF	LIFT ↓	IT...	ITEM2	ITEM3	RULE
3	2	119	6.276371308	77.272727273	11.26993007	Implementing Analytics	Customer Analytics For Dummies	Managerial Analytics	Implementing Analytics ==> Customer Analytics For Dummies & Managerial Analytics
4	2	128	6.7510548523	83.116883117	11.256400742	Implementing Analytics	Data Science and Big Data Analytics	Enterprise Analytics	Implementing Analytics ==> Data Science and Big Data Analytics & Enterprise Analytics

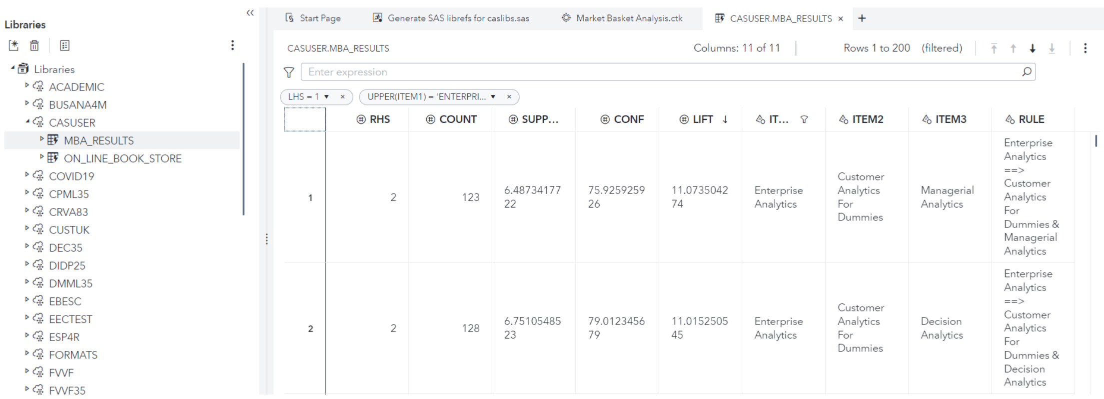
Table 2: Top Association Rules for Customers Who Bought Implementing Analytics



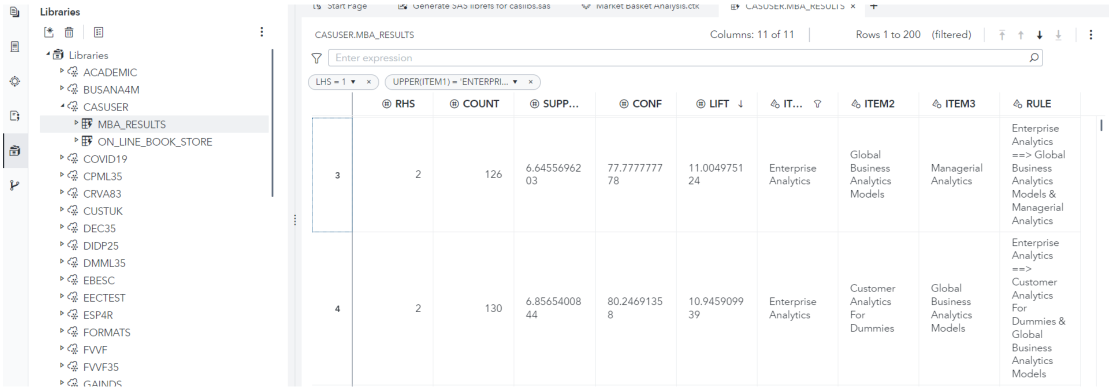


	RHS	COUNT	SUPP...	CONF	LIFT	IT...	ITEM2	ITEM3	RULE
1	2	128	6.7510548523	82.051282051	11.192030991	Customer Analytics For Dummies	Decision Analytics	Enterprise Analytics	Customer Analytics For Dummies ==> Decision Analytics & Enterprise Analytics
2	2	127	6.6983122363	81.41025641	11.185061315	Customer Analytics For Dummies	Decision Support, Analytics, and BI	Implementing Analytics	Customer Analytics For Dummies ==> Decision Support, Analytics, and BI & Implementing Analytics
3	2	127	6.6983122363	81.41025641	11.185061315	Customer Analytics For Dummies	Enterprise Analytics	Essentials of Predictive Analytics	Customer Analytics For Dummies ==> Enterprise Analytics & Essentials of Predictive Analytics
4	2	125	6.5928270042	80.128205128	11.17081448	Customer Analytics For Dummies	Decision Support, Analytics, and BI	Managerial Analytics	Customer Analytics For Dummies ==> Decision Support, Analytics, and BI & Manag[illegible] Anal[illegible]

Table 3: Top Association Rules for customers who bought Customer Analytics for Dummies



	RHS	COUNT	SUPP...	CONF	LIFT	IT...	ITEM2	ITEM3	RULE
1	2	123	6.4873417722	75.925925926	11.073504274	Enterprise Analytics	Customer Analytics For Dummies	Managerial Analytics	Enterprise Analytics ==> Customer Analytics For Dummies & Managerial Analytics
2	2	128	6.7510548523	79.012345679	11.015250545	Enterprise Analytics	Customer Analytics For Dummies	Decision Analytics	Enterprise Analytics ==> Customer Analytics For Dummies & Decision Analytics



	RHS	COUNT	SUPP...	CONF	LIFT	IT...	ITEM2	ITEM3	RULE
3	2	126	6.64556962 03	77.7777777 78	11.0049751 24	Enterprise Analytics	Global Business Analytics Models	Managerial Analytics	Enterprise Analytics ==> Global Business Analytics Models & Managerial Analytics
4	2	130	6.85654008 44	80.2469135 8	10.9459099 39	Enterprise Analytics	Customer Analytics For Dummies	Global Business Analytics Models	Enterprise Analytics ==> Customer Analytics For Dummies & Global Business Analytics Models

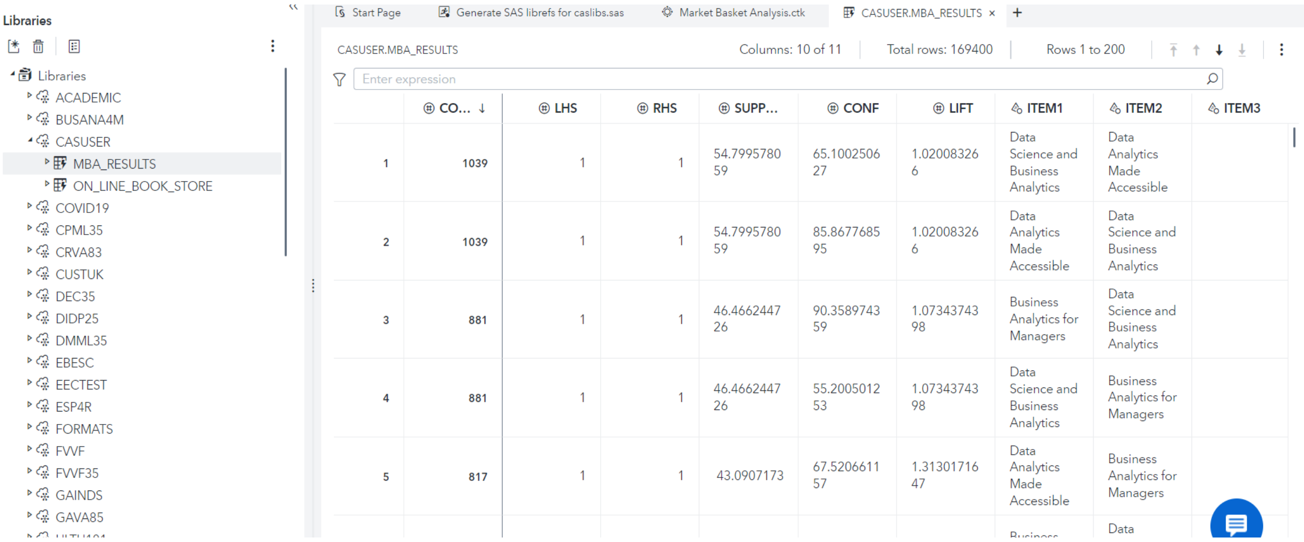
Table 4: Top Association Rules for customers who bought Enterprise Analytics

Question 4:

The set of 3 books that were bought together the most times was Data Analytics Made Accessible, Data Science and Business Analytics, and Business Analytics for Managers. These books were bought together 794 times, and the support metric of this association rule is 41.87. The support metric is the probability that a customer has bought all these 3 books together. It is calculated using the following formula:

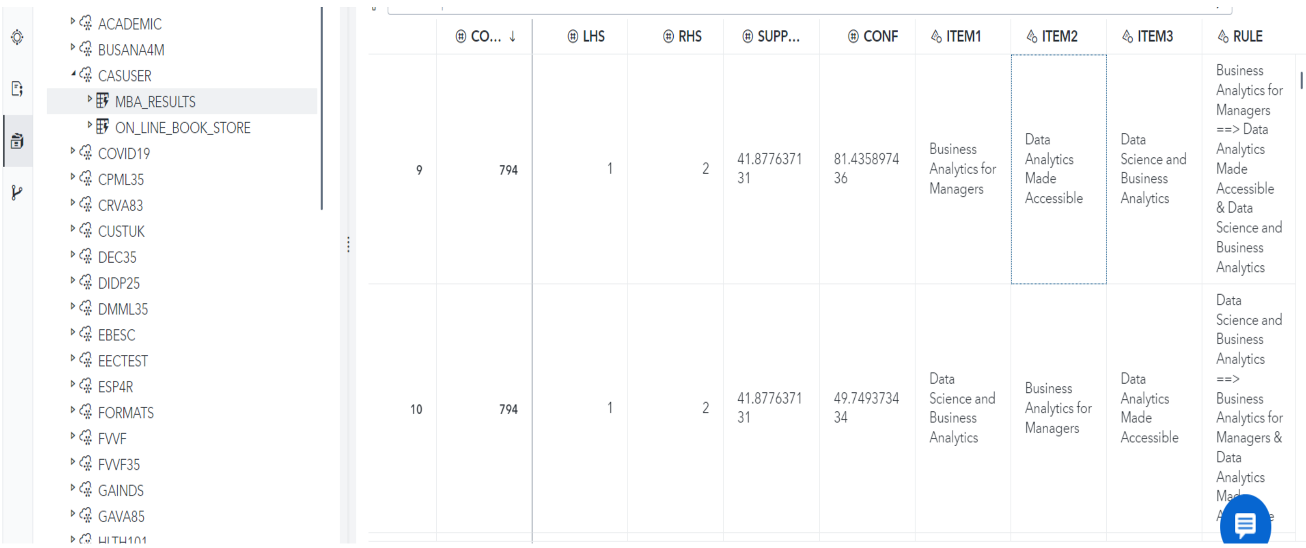
$$Support = \frac{\#\, transactions\ in\ which\ the\ 3\ books\ were\ bought\ together}{Total\ number\ of\ transactions}$$

So, we know that a customer has a probability approximately equal to 42% to buy these 3 books together. As for the technical part to find the answer to this question we sorted all the rules in descending order based on their count (i.e. how many times are found in the dataset) and we searched for combinations/associations that had 3 books.



	CO...	LHS	RHS	SUPP...	CONF	LIFT	ITEM1	ITEM2	ITEM3
1	1039	1	1	54.7995780 59	65.1002506 27	1.02008326 6	Data Science and Business Analytics	Data Analytics Made Accessible	
2	1039	1	1	54.7995780 59	85.8677685 95	1.02008326 6	Data Analytics Made Accessible	Data Science and Business Analytics	
3	881	1	1	46.4662447 26	90.3589743 59	1.07343743 98	Business Analytics for Managers	Data Science and Business Analytics	
4	881	1	1	46.4662447 26	55.2005012 53	1.07343743 98	Data Science and Business Analytics	Business Analytics for Managers	
5	817	1	1	43.0907173	67.5206611 57	1.31301716 47	Data Analytics Made Accessible	Business Analytics for Managers	

Table 5: Most common combinations with two books



	CO...	LHS	RHS	SUPP...	CONF	ITEM1	ITEM2	ITEM3	RULE
9	794	1	2	41.8776371 31	81.4358974 36	Business Analytics for Managers	Data Analytics Made Accessible	Data Science and Business Analytics	Business Analytics for Managers ==> Data Analytics Made Accessible & Data Science and Business Analytics
10	794	1	2	41.8776371 31	49.7493734 34	Data Science and Business Analytics	Business Analytics for Managers	Data Analytics Made Accessible	Data Science and Business Analytics ==> Business Analytics for Managers & Data Analytics Ma...

Table 6: Most common combinations with three books

Case Study 2: RFM Analysis

Executive Summary

The analysis conducted is a customer segmentation analysis and aims to help the company improve its marketing strategies by gaining deeper insight into its customer base and moreover improve its business performance. Over six years of transactional data was analyzed to effectively segment customers for targeted marketing initiatives. Utilizing SAS Viya, a powerful analytics platform, we conducted a Recency Frequency Monetary (RFM) segmentation analysis, employing clustering techniques. This method categorizes customers based on their recent purchase behavior, frequency of purchases, and their financial value. By grouping customers into discrete segments, important customer groups were identified, providing valuable information about purchasing behavior. SAS Viya's analytics tools facilitated the segmentation process, enabling the identification of unique customer segments. These segments will inform personalized marketing campaigns tailored to specific customer needs and preferences, ultimately optimizing marketing efforts, enhancing customer engagement, and driving revenue growth. The clustering results created five segments, “Value Driven Customers”, “New Customers”, “Churners”, “Lapsed Costumers “ and “Best Customers”.

Question 2:

The provided data was of very good quality. In particular, there were no missing values in the variables used for clustering, which were Recency, Frequency and Monetary. The only necessary transformation involved addressing the distribution and outliers of certain variables. Specifically, for Recency, a moderate positive skewness was observed, indicating that while most customers had made recent purchases, a tail of customers had not made purchases in a longer timeframe (skewness of recency: 0.58). Similarly, Frequency showed a slight positive skewness, suggesting that while most customers made a low number of purchases, some made a higher number (skewness of frequency: 0.34). On the other hand, Monetary displayed a near symmetrical distribution, indicating relatively balanced spending among customers, although not perfectly normal (skewness of monetary value: 0.04). To address these characteristics, logarithmic transformations were applied before clustering as logarithmic transformations can help normalize positively skewed data, making it more symmetrical and easier to analyze. Additionally, outliers in these variables were identified and excluded during the transformation process. The lower Whisker of Recency is 1 and the upper Whisker is 30, for Frequency limits are 1 to 10 and for Monetary is 27 to 161,33 respectively. These adjustments ensured the robustness of the clustering analysis.

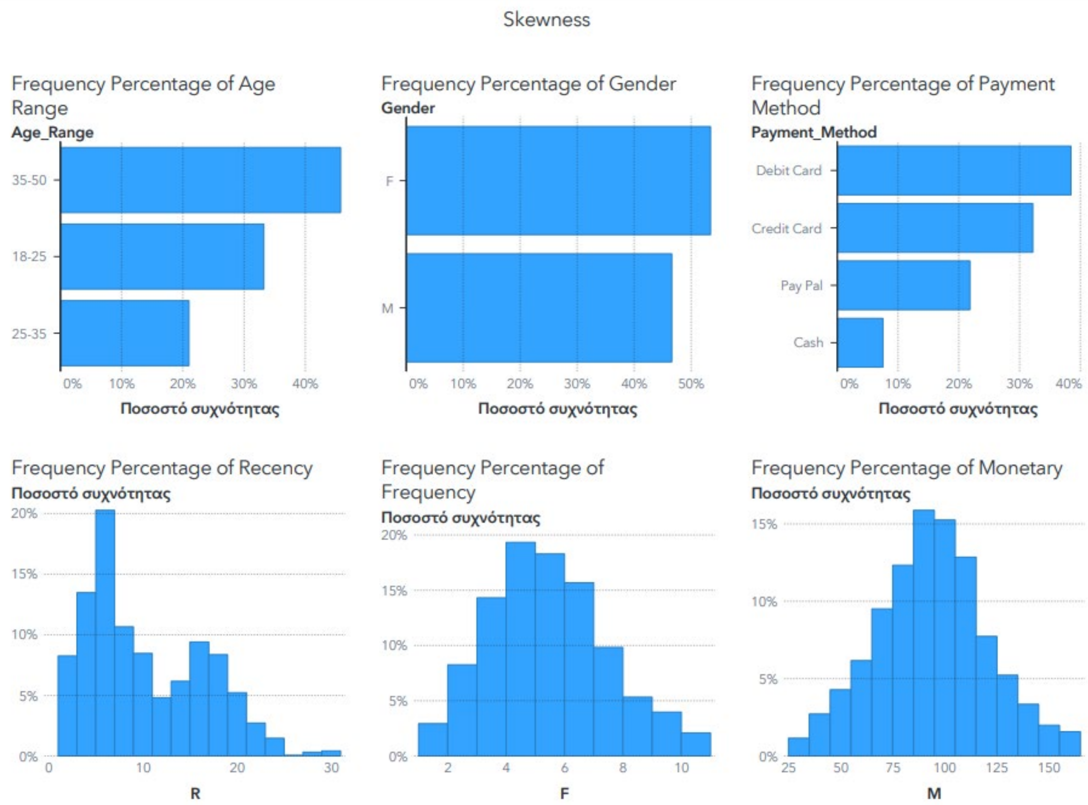


Figure 2: Distribution of variables

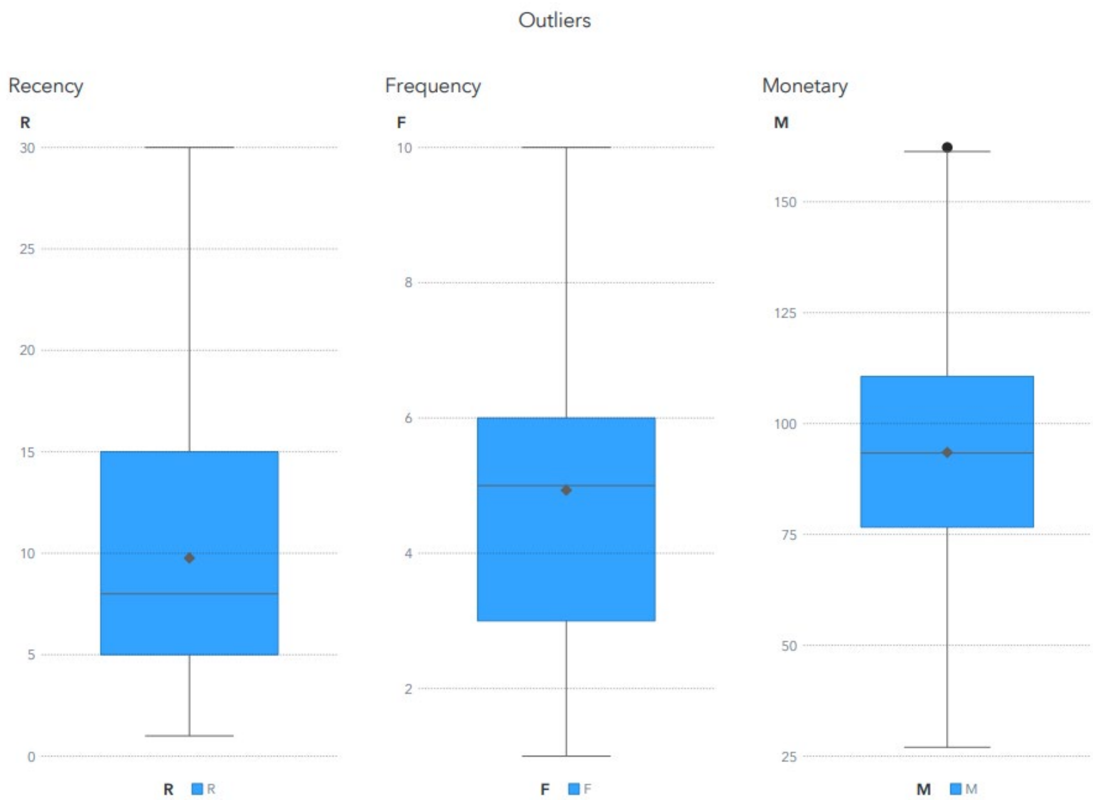


Figure 3: Outliers

Question 3:

The clustering analysis revealed the existence of distinctive customer segments, each characterized by unique patterns of purchasing trends. These segments were created as follows:

- Value Driven Customers:
  Customers who have not interacted with the company for 14 months and in the analysis period in specific, have interacted 3 to 4 times, but spend €115 each time.
    - Recency: The recency of 14 months indicates that these customers have not engaged with the company for an extended period, which is higher than the average recency of 9,7 months.
    - Frequency: With a frequency of 3,8, these customers have made fewer interactions compared to the average frequency of 4,9. This further supports the notion that they are less engaged with the company.
    - Monetary: The monetary value of €115 per interaction is higher than the average of €93,48. While this indicates that these customers tend to spend more per transaction, their overall contribution to revenue may be limited due to their lower frequency of interactions.

- New Customers:
  Customers who have not interacted with the company for 4 to 5 months, in the analysis period in particular, have interacted 3 to 4 times and spend €66,64 each time.
    - Recency: Their recency of 4,7 months is lower than the average recency of 9,7 months. This confirms that they are relatively new to the company or have recently re-engaged after a period of inactivity.
    - Frequency: While their frequency of 3,9 interactions is slightly below the average of 4,9, it indicates that they engage with the company relatively frequently for their recent time frame. This level of interaction aligns with the behavior of new customers who are still exploring the brand and its offerings.

  Monetary Value: Their spending of €66,64 per transaction is lower than the average of €93,48. New customers often start with smaller purchases as they familiarize themselves with the products or services offered by the company. This spending level is typical for new customers who are testing the waters before making larger purchases.

- Churners:

  Customers who have not interacted with the company for 14 to 15 months, in the analysis period in particular, have interacted 5 to 6 times and spent €75.34 each time.

  - Recency: Their recency of 14,6 months is lower than the average recency of 9,7 months. While this might seem positive at first glance, it's important to note that "Churners" typically exhibit a higher recency, indicating a longer period of inactivity. However, in this case, the recency is still relatively high at 14,68, which suggests that these customers have not engaged with the company for an extended period, contributing to their classification as "Churners."
  - Frequency: Their frequency of 3,9 interactions is slightly above the average of 4,9. While this might seem contradictory for "Churners," it's possible that these customers exhibited higher engagement levels in the past before becoming inactive. This higher frequency could indicate a previous pattern of regular interaction with the company before the onset of churn behavior.
  - Monetary Value: Their spending of €66,64 per transaction is lower than the average of €93,48. This aligns with the typical behavior of "Churners," who may decrease their spending or engagement levels before eventually ceasing interaction with the company altogether.

- Lapsed Customers:

  Customers who have not interacted with the company for 15 to 16 months, in the analysis period in particular, have interacted 1 to 2 times and spend €43,31 each time.

  - Recency: Their recency of 15,9 months is higher than the average recency of 9,7 months. This indicates that these customers have not engaged with the company for an extended period, contributing to their classification as "Lapsed Customers." The higher recency suggests a longer period of inactivity, which is characteristic of lapsed customers.
  - Frequency: Their frequency of 1,9 interactions is lower than the average of 4,9. This aligns with the behavior of "Lapsed Customers " who typically exhibit a decrease in engagement levels before eventually ceasing interaction with the company altogether. The lower frequency indicates a decline in their interaction over time.
  - Monetary Value: Their spending of €43,31 per transaction is lower than the average of €93,48. This further supports their classification as "Lapsed Customers," as they may have decreased their spending

levels before becoming inactive. The decrease in spending is often a precursor to disengagement from the company.

- Best Customers:
  Customers who have not interacted with the company for 4 to 5 months, in the analysis period in particular, have interacted 6 to 7 times and spend €99.81 each time.
  - Recency: Their recency of 4,8 months is lower than the average recency of 9,7 months. This indicates that these customers have engaged with the company relatively recently, which is a positive sign. "Best Customers" typically exhibit lower recency, as they are actively engaged with the company.
  - Frequency: Their frequency of 6,5 interactions is higher than the average of 4,9. This indicates that these customers engage with the company frequently, making them valuable in terms of their repeat business. "Best Customers" are characterized by their high frequency of interaction with the company.
  - Monetary Value: Their spending of €99,81 per transaction is higher than the average of €93,48. This indicates that these customers spend more per transaction compared to the average customer, which is another characteristic of "Best Customers." They contribute significantly to the company's revenue through their higher spending levels.

Profiling

Cluster ID	Segment Names	Συχνότητα	Ποσοστό συχνότητας	R	F	M
1	High Value customers	296	30.93%	14.043918919	3.8141891892	115.05286626
2	New customers	129	13.48%	4.7906976744	3.976744186	66.641454411
3	Churners	153	15.99%	14.68627451	5.2352941176	75.342107584
4	Worst customers	45	4.70%	15.933333333	1.9777777778	43.312962963
5	Best customers	334	34.90%	4.8083832335	6.5508982036	99.813032269
Άθροισμα		**957**	**100.00%**	**9.7648902821**	**4.9320794148**	**93.486284852**

Table 7: Profiling of Customers

The largest proportion of customers, comprising 36%, falls within the category of "Best Customers", representing those who exhibit high levels of engagement and contribute significantly to the revenue stream. Following closely behind, "Value Driven Customers" account for 30% of the customers, showcasing a substantial commitment to obtaining value from their interactions with the company. Notably, only a minimal 4,7% of customers are classified as 'lapsed,' indicating a relatively low rate of disengagement within customer segments.

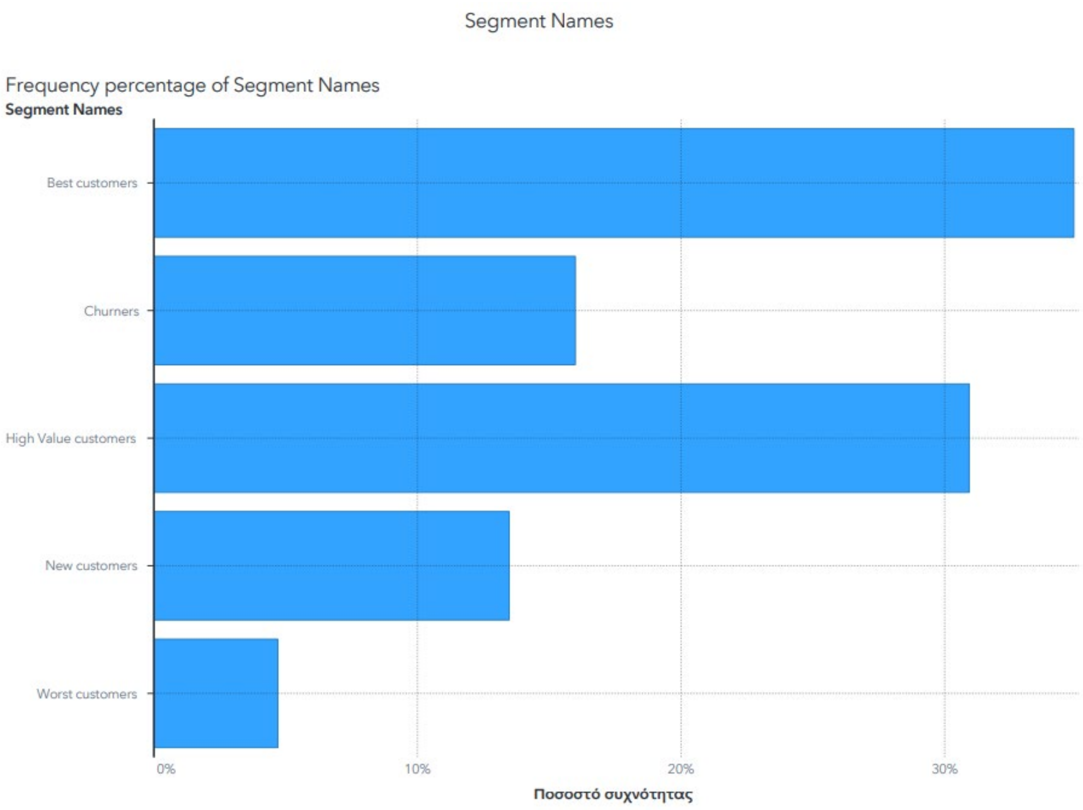


Figure 4: Segment Names

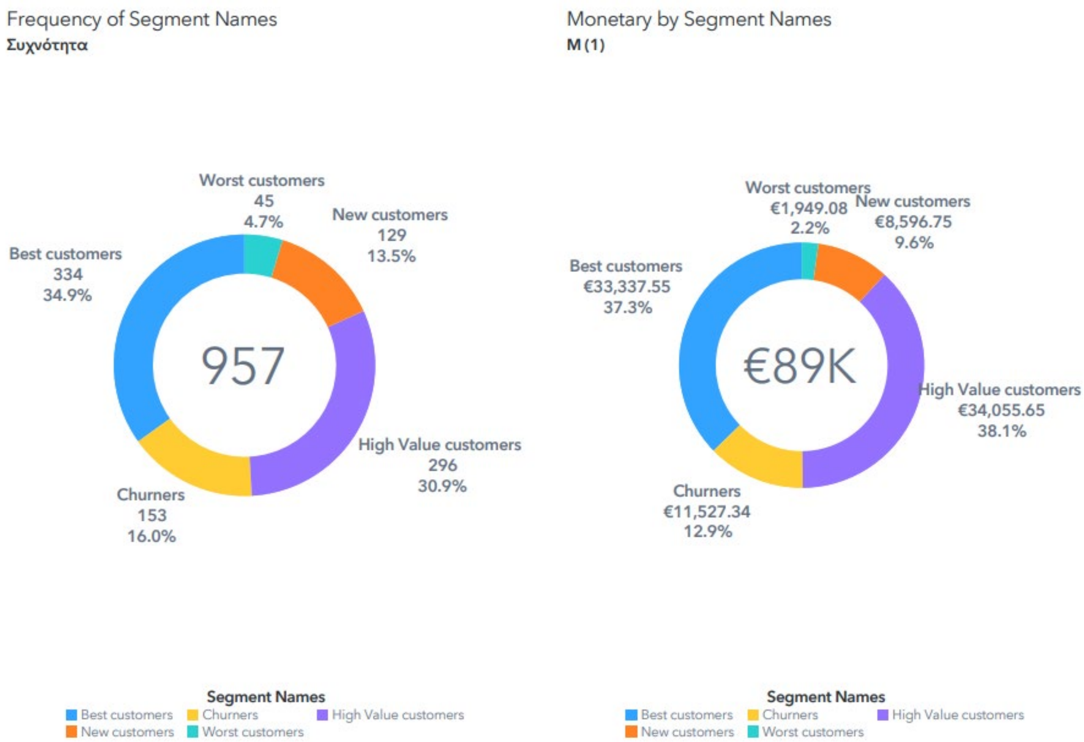


Figure 5: Frequency & Monetary by Segments

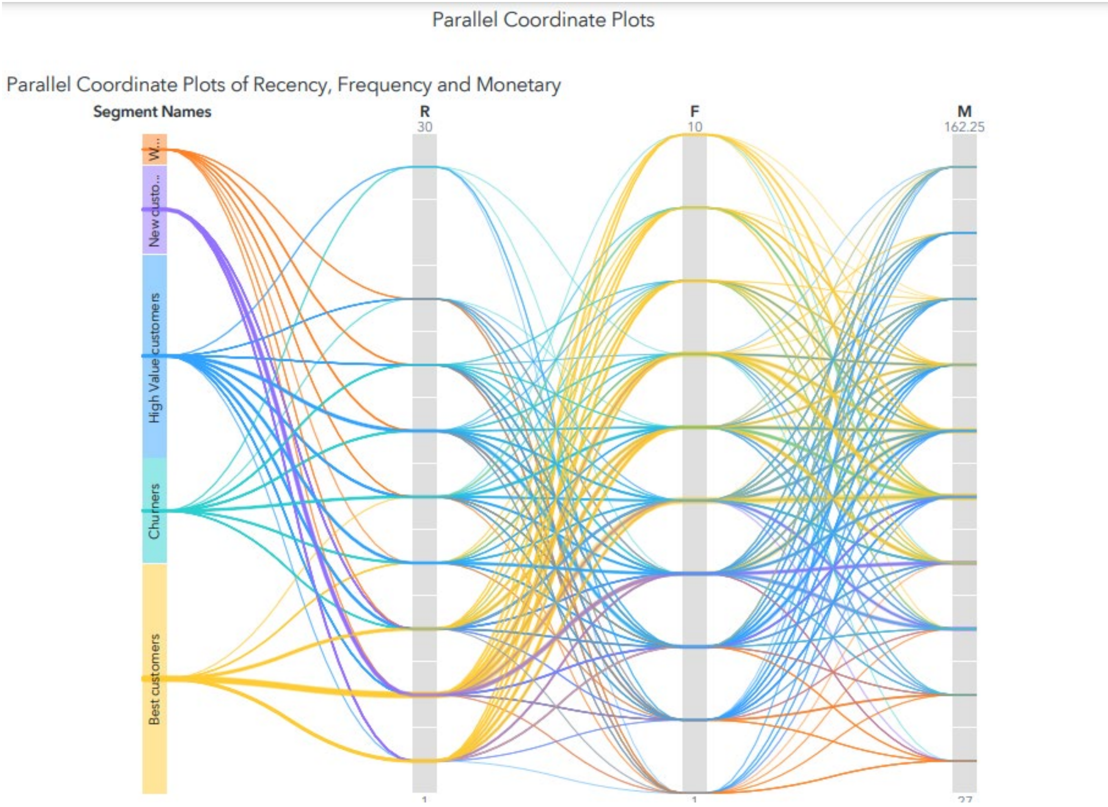


Figure 6: RFM Parallel Coordinate plot

Question 4:

In addition to the Recency, Frequency, and Monetary variables utilized for the customer segmentation analysis, supplementary data regarding customers' demographics and preferences were also accessible. This included information such as age range, gender, and most common payment method.

Based on the findings, customers within the age range of 25 to 35 years have exhibited the most recent interactions with the stores and have demonstrated the highest level of engagement with the company during the analysis period. Additionally, this age range appears to contribute significantly to the revenue stream, indicating their considerable value to the business.

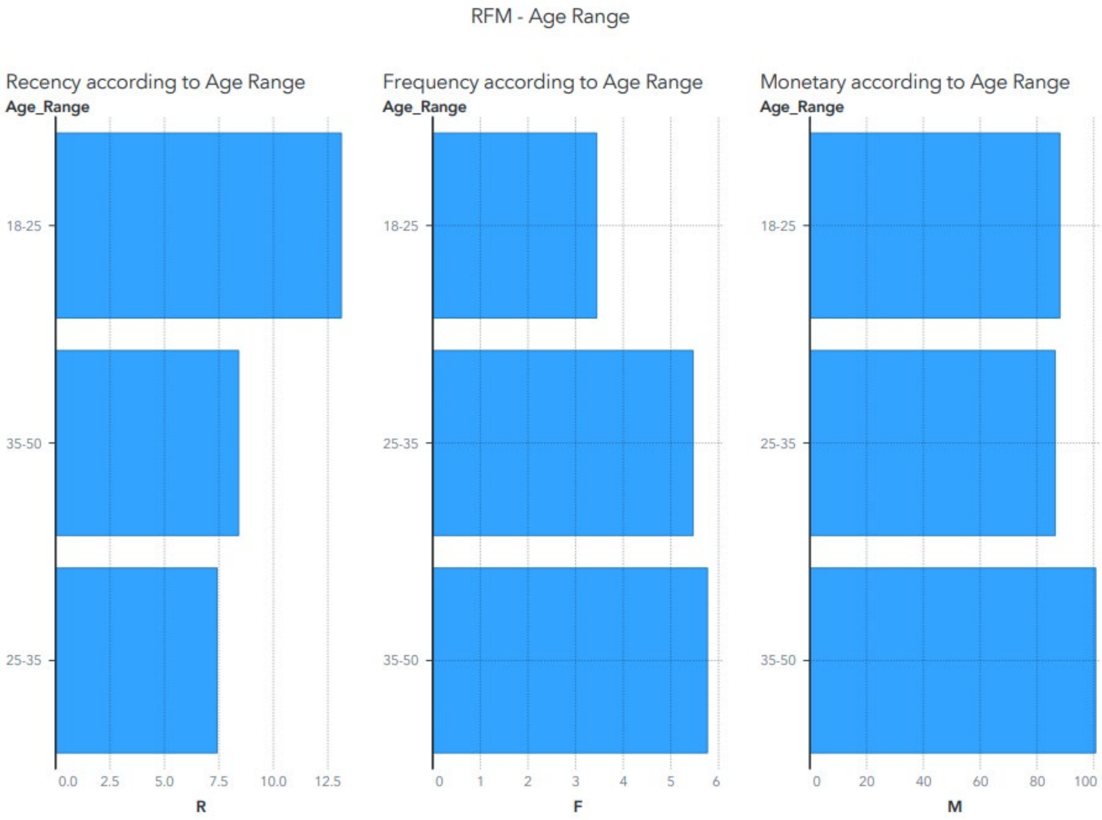


Figure 7: RFM by Age range

Our analysis reveals a notable trend wherein women emerge as more valuable customers compared to men. This is evident from their higher frequency of interactions with the company and their tendency to spend more with each purchase.

Question 5:

Below there are appropriate marketing actions for each segment along with the rationale behind them.

- Best Customers:
  A great idea would be to implement a loyalty award program to incentivize commitment and continuous shopping from valuable customers. Rewarding them for their loyalty can reinforce the brand and encourage them to remain loyal in the long run. Still, it would be worth providing early access to new products for the best customers as a way to show appreciation for their loyalty. This can also create a sense of community among these valued customers.

- Value-Driven Customers:

Since this category of customers prioritize getting value from their purchases, offering relevant products or packages can improve their overall shopping experience and increase their satisfaction. In addition, cross-selling strategies could be used to introduce complementary products or services.

- New Customers:

  For new customers, welcoming offers or introductory discounts could be implemented as an incentive to make their first purchase and have a positive initial experience with the brand. This can help build customer loyalty and encourage repeat purchases in the future. Furthermore, the implementation of a personalised onboarding process to facilitate new customers' navigation of the brand's products would significantly contribute to their initial experience. This may include welcoming emails, product presentations or personalized recommendations based on their preferences.

- Churners:

  For churners, it's urgently necessary to launch a targeted reactivation campaign to win back customers who have recently left or have shown signs of disengagement. Offering special pricing or personalized incentives can revive their interest in the brand and encourage them to reconnect with the company. At the same time, it is essential to conduct exit surveys to get feedback from customers who have left and understand the reasons for their departure. This information can be used to improve the brand's products and services and prevent future churn.

- Lapsed Customers:

  To address the challenge of lapsed customers, we propose a multi-faceted approach focused on reactivation and reconnection. Firstly, we recommend launching a reactivation campaign tailored to individual customers, leveraging personalized offers based on their past purchase history and preferences. By enticing lapsed customers with exclusive discounts or rewards, we aim to reignite their interest and incentivize them to return. Additionally, we propose conducting a feedback survey to gather insights into why customers have lapsed and their expectations moving forward. Following up with personalized outreach to address their feedback and concerns, we can demonstrate our commitment to listening and improving, ultimately fostering trust and encouraging lapsed customers to reconnect with the brand.

By targeting customers with relevant and personalized strategies, the company can effectively foster relationships, drive engagement, and ultimately, maximize customer lifetime value.

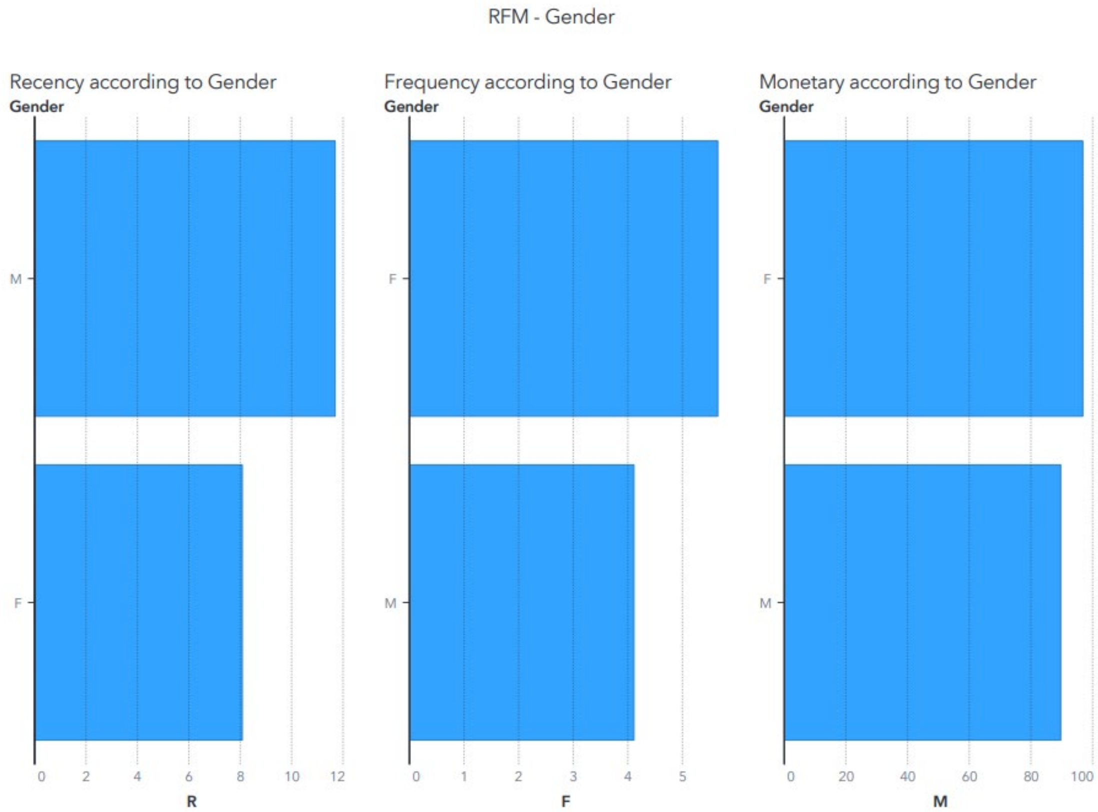


Figure 8: RFM by Gender

Our analysis reveals a noteworthy trend: customers who made the most recent purchases predominantly used credit cards, whereas those with the highest interaction frequency during the analysis period and who spent the most opted for debit cards. This correlation aligns with the common practice of using debit cards for larger transactions.

Case study 3: Churn Prediction

Executive Summary

In this analysis we want to find a competent model that will predict which customers are most likely to churn from our services. By doing this the company will be able to find effectively which customers will hop to another telecommunication provider and develop a strategy which will ensure that XYZ will maintain its market share in a competitive environment. We tried out different predictive methods to find the most suitable for our case such decision trees, logistic regression, and a neural network. We found out that logistic regression is the best model for the churn prediction (we used misclassification rate for the evaluation of the models). The tools used for the building and the evaluation of the models was SAS Viya and SAS Visual Analytics. Finally, after implementing the models we found out that the most important characteristics for keeping the customers that are likely to churn is if they have an international plan, a voice mail plan, how many calls they have made to the customer services, how many minutes they have talked on international calls and how many minutes they have spent during the day.

Question 2:

To understand what the cost of a customer will be turning into a churner, what will be the cost of a campaign that will motivate the customers that are likely to churn to stay at our company, and how the proper churn prediction will help the company to invest effectively in a campaign to maximize its profits knowing which customers are likely to become churners based on the predictions of a model. To do the above we interpreted the following profit matrix provided to us by the management of the XYZ company.

		Business action given the predicted Customer Status	
		Motivate to Stay	**Do Nothing**
Actual Customer Status	**Churner**	1000	-1500
	Non-Churner	-500	0

This table shows what will be the profits of the company when comparing the actual customer status (i.e. if the customer will leave our company or not) with the profits or losses resulting from business policies that are based on the prediction of customer status. The negative numbers are indicating losses, and the positive numbers are indicating profits. We can analyze this matrix further in the following manner:

- If a customer is likely to churn and XYZ gives him the motivation to stay, then the company will gain a profit of 1000€ since it will continue to receive the price of the services that will be provided (the 1000€ amount also includes some expenses because of the motivation that the company will give to the churners).
- If a customer is not likely to churn and XYZ gives him the motivation to stay then we will face a loss of 500€ because the cost of the campaign will have no return on the company. After all, this customer would have stayed without needing to make him/her a better offer.
- If a customer is likely to churn and we do not give him a reason to stay then the company will face a loss of 1500€ since it will lose one customer and won't receive the price of the services that will be provided.
- Similarly, if a customer is not likely to churn and the company does not offer him/her a better deal then no loss or profit will be derived because the customer will continue to buy from us and the company will not have any cost since it does not provide him any additional motivation.

From the interpretation of this profit matrix it is obvious that the development of a reliable predictive machine-learning model is crucial because wrong predictions can lead XYZ to losses or wrong decisions that will not allow it to extract all available profits from the market.

Question 3:

Given that the development of our model will be based on the calculation of probabilities the optimal cut-off point based on which a customer should be classified as a churner is 16% or 0.16. This point is derived from the profit matrix provided above with decision theory. Given that the event we want to model is the customers that are likely to churn we will represent the probability of a customer becoming a churner as p and the probability of a customer not being such as 1-p because the event that a customer won't be a churner is complementary with a customer being a churner. Also given these probabilities we want the benefits that derive from giving churners motivation to stay to be bigger than the benefits that derive from simply doing nothing and letting the current state be as it is. So given all the above we constructed and solved the following mathematical inequality:

$1000\times p - 500\times(1-p) > -1500\times p + 0\times(1-p) \Rightarrow 2500\times p > 500\times(1-p) \Rightarrow 2500\times p > 500 - 500\times p \Rightarrow 3000\times p > 500 \Rightarrow p > \frac{500}{3000} \Rightarrow p > 0,16$

Question 4:

Before starting training various predictive models, we partitioned the historical data set in training and validation using a 70%-30% split. This step must be done because the training and the evaluation of the model must be done in two distinct datasets. The training set is used to fit the model's parameters, essentially teaching it to recognize patterns in the data. The reason that the majority of the historical dataset is used for training purposes is that the more data we have available for training, the better the model can learn these patterns. The part of the data that is used as a validation set serves as an independent dataset to evaluate the model's performance. By testing the model on data it hasn't seen during training, we can assess honestly its ability to generalize to new, unseen data.

Stratified sampling is a method that divides the dataset into homogeneous subgroups and then draws random samples from each subgroup. So, in the context of partitioning data for machine learning, stratified sampling ensures that the distribution of classes remains consistent across the training and validation sets. So in our case, we know that the proportion of non-churners is approximately 86% and that of churners 14% so the stratified sampling ensures that both the training and validation sets maintain this 86-14 ratio between these two classes.

Question 5:

After exploring the dataset using SAS Visual Analytics tools, we found that none of the variables of the dataset has missing values. This fact is also shown in the screenshots below where some descriptive statistics can be found for each variable along with the number of missing values.

Figure 9:Number of NAs and descriptive stats for Account Length

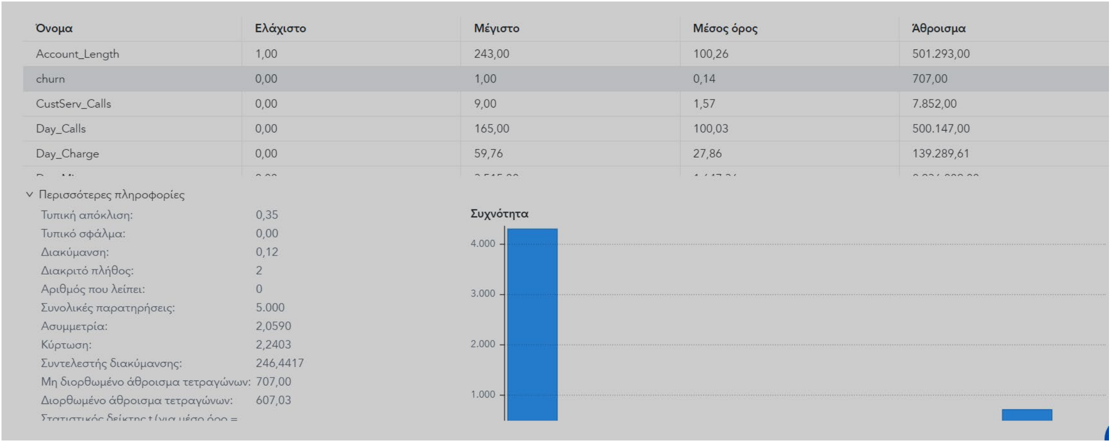


Figure 10: Number of NAs and descriptive stats for churn (target variable)

Figure 11: Number of NAs and descriptive stats for Calls made to Customer Service

Figure 12: Number of NAs and descriptive stats for Day Calls

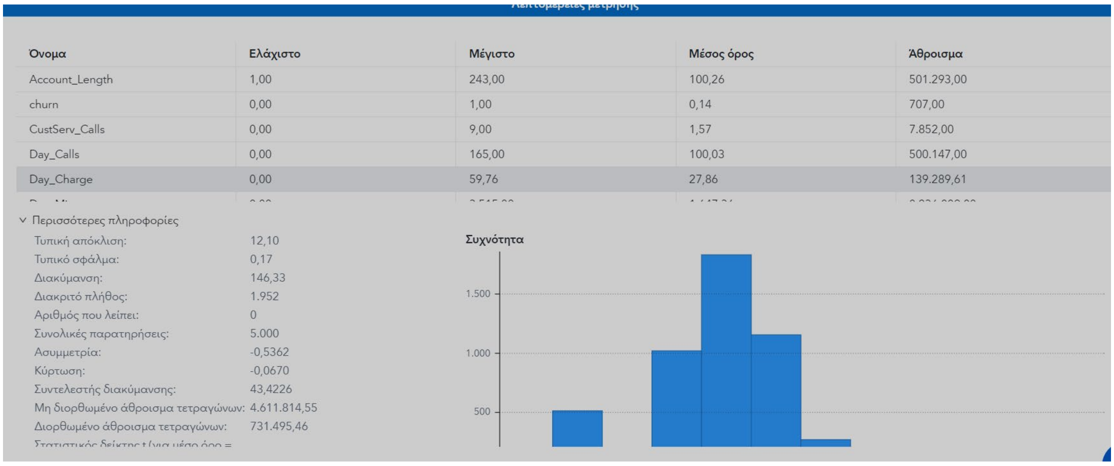


Figure 13 : Number of NAs and descriptive stats for Day Charge

Figure 14: Number of NAs and descriptive stats for Day Minutes

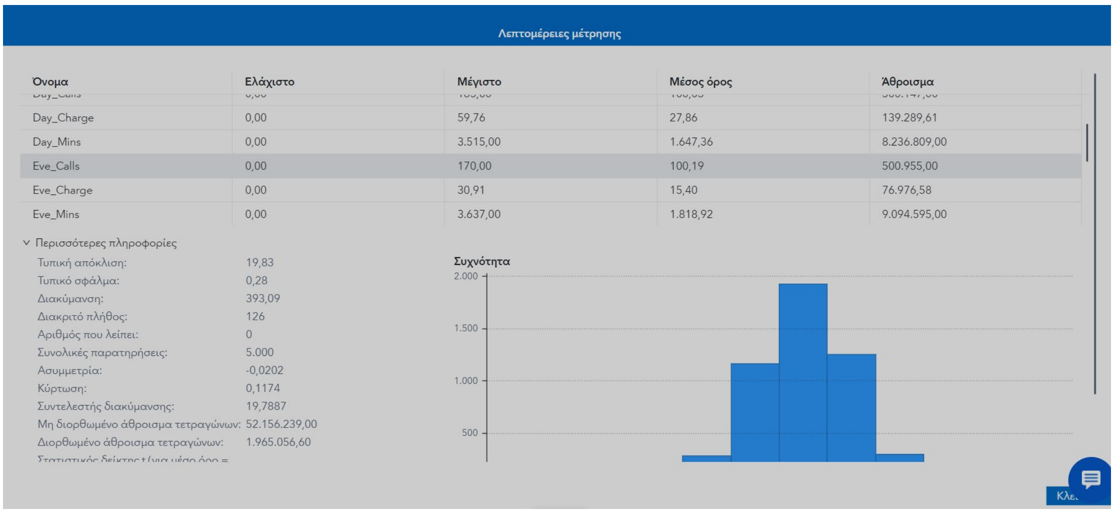


Figure 15: Number of NAs and descriptive stats for Evening Calls

Figure 16: Number of NAs and descriptive stats for Evening Charge

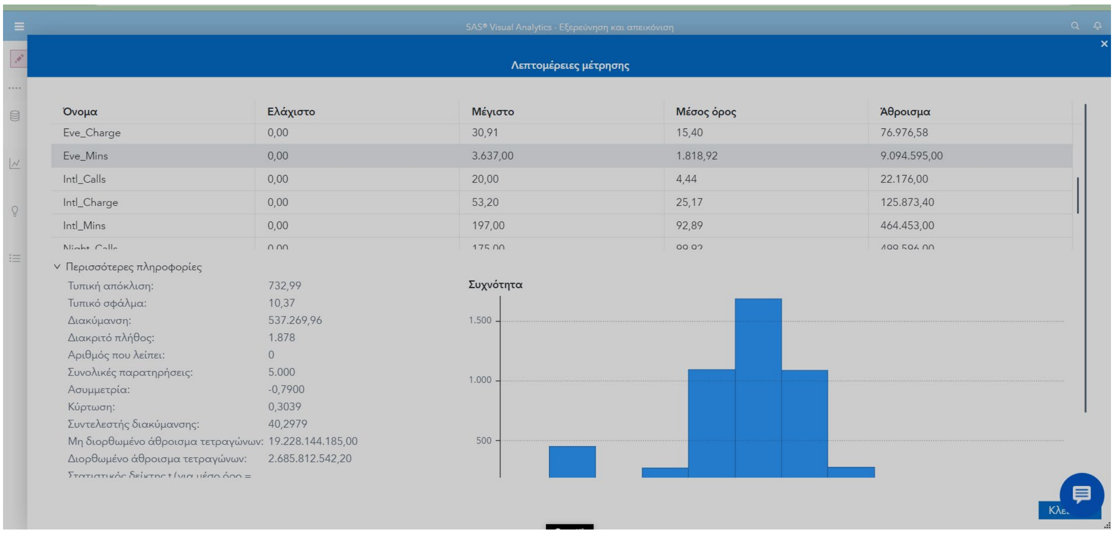


Figure 17: Number of NAs and descriptive stats for Evening Minutes

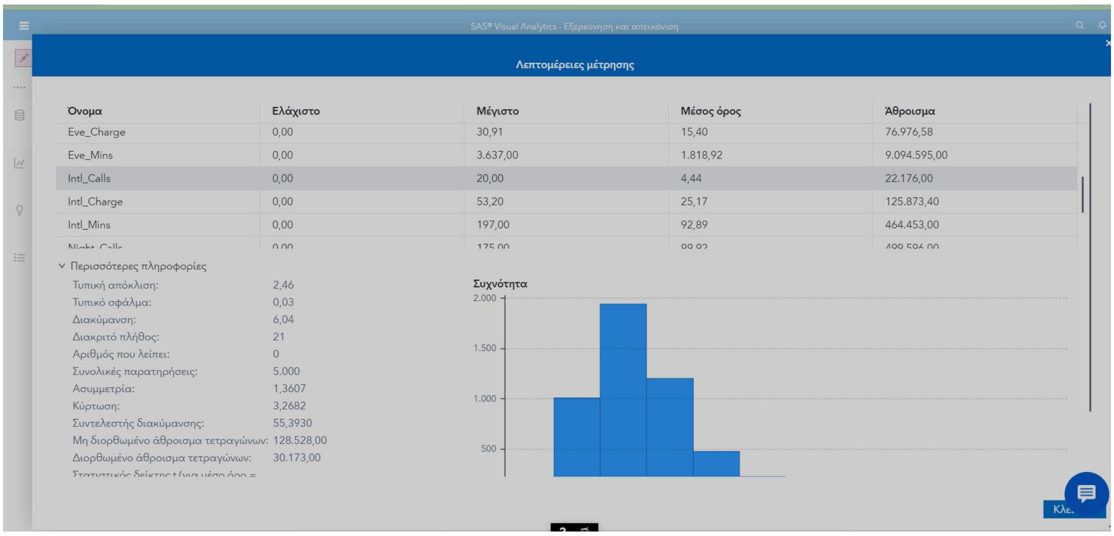


Figure 18: Number of NAs and descriptive stats for International Calls

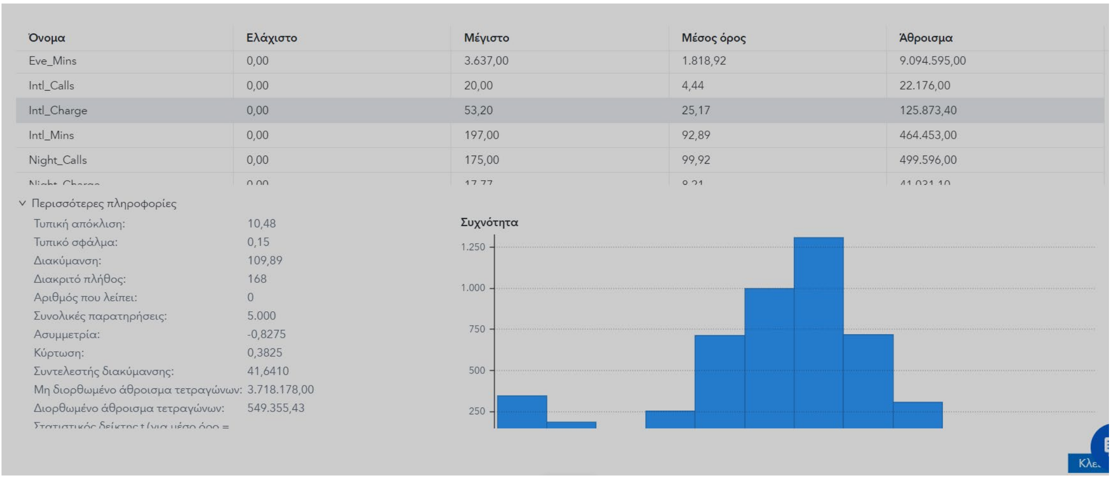


Figure 19: Number of NAs and descriptive stats for International Charge

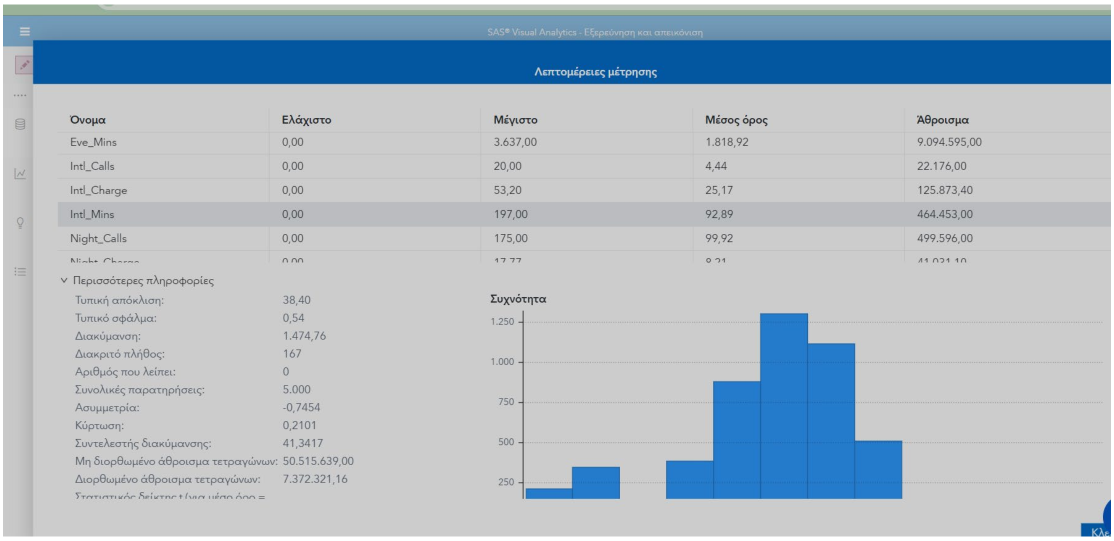


Figure 20: Number of NAs and descriptive stats for International Minutes

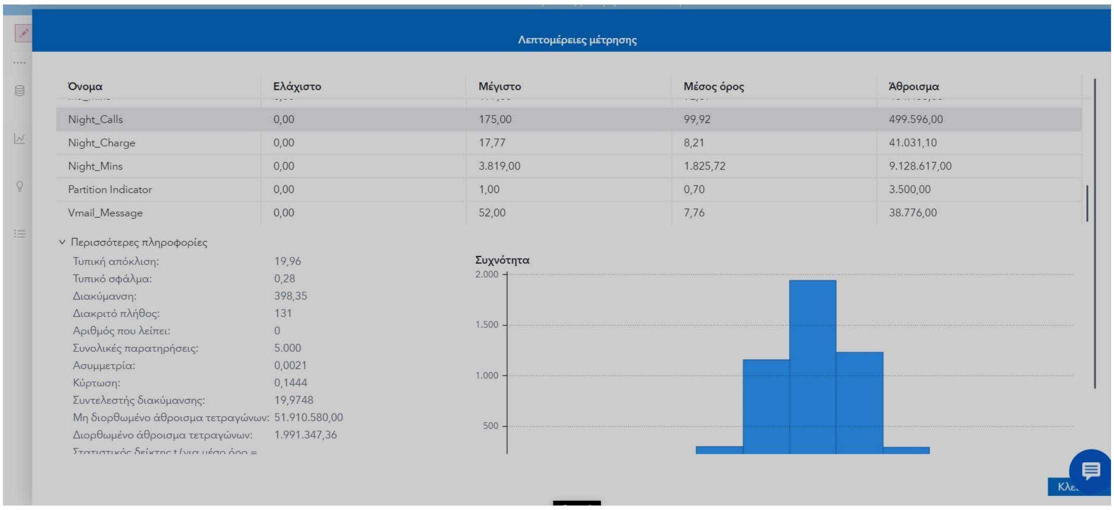


Figure 21: Number of NAs and descriptive stats for Night Calls

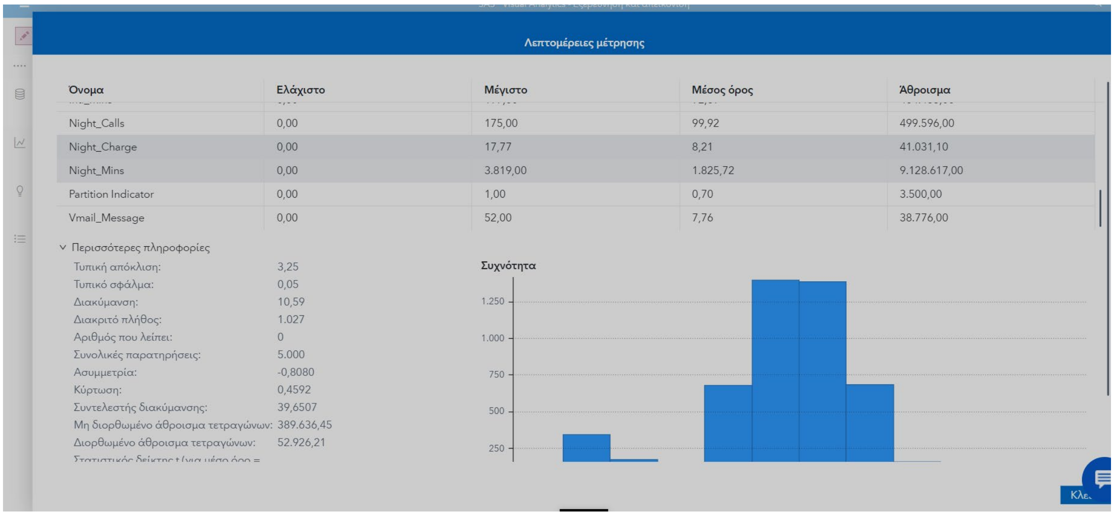


Figure 22: Number of NAs and descriptive stats for Night Charge

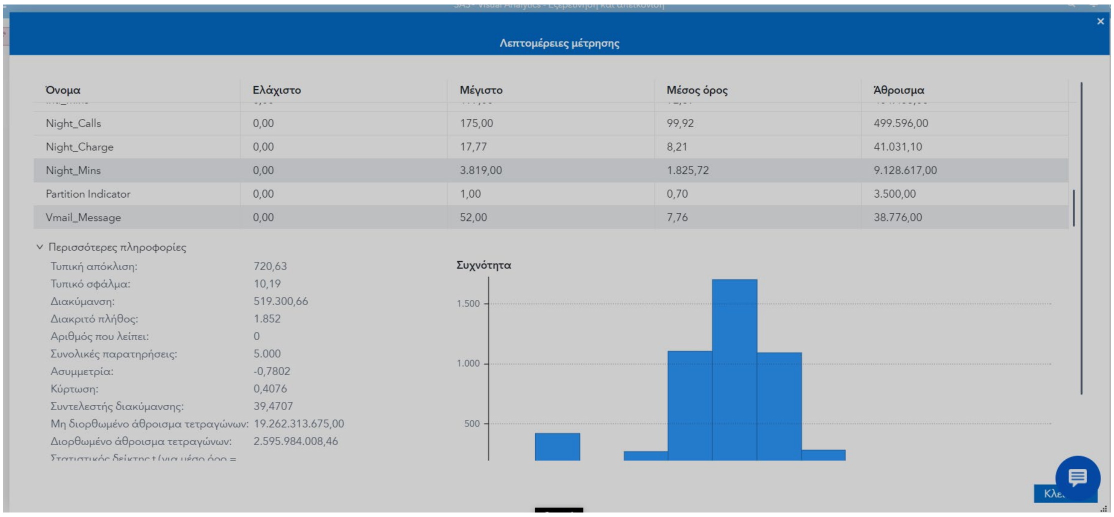


Figure 23: Number of NAs and descriptive stats for Night Minutes

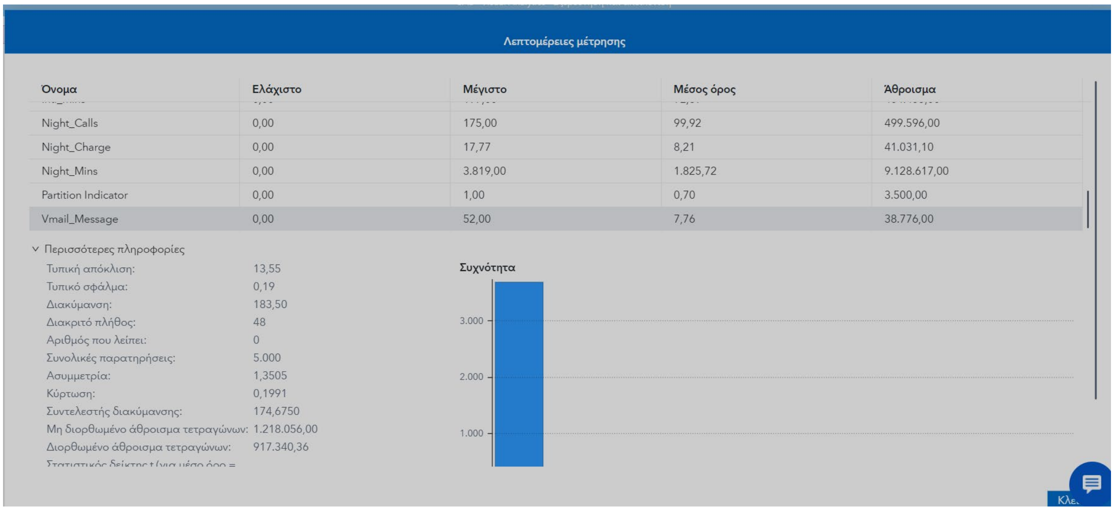


Figure 24: Number of NAs and descriptive stats for Voice Mails messages sent

In the pie chart below we can see how many customers there are in total in the dataset in the center of the pie (5000), and we can see the proportions (in percentages) of churners and non-churners in the historical data set and also the absolute numbers of the two customer classes.

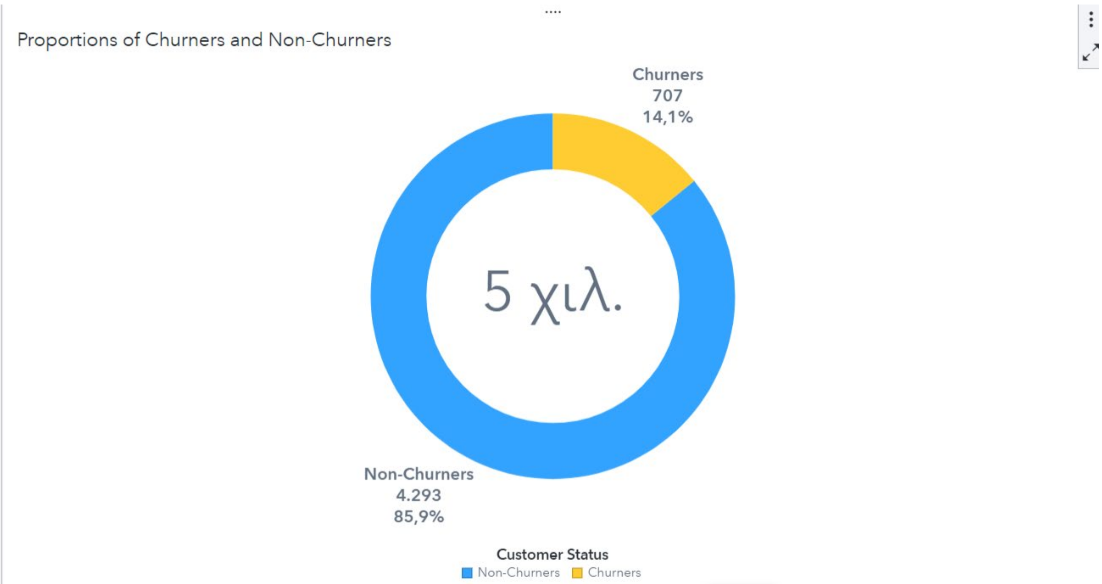


Figure 25: Churners & Non-Churners

Question 6:

If the proportion of churners and non-churners was 3% - 97% respectively then we would have a case of an extremely imbalanced dataset that could lead machine learning models to poor predictions because of the very rare occurrence of the churners' class. In that case we should balance our data with some technique to ensure a satisfactory performance for our models.

So, in a case where the ratio of the two classes for the target variable was 3% and 97% to give the dataset a more balanced form, we should increase the size of the churners' class and for that we would perform oversampling and undersampling for the majority class (non-churners). This way, we would randomly select a subset of the data from the churners until the ratio between the non – churners and the churners becomes more balanced. Then, by reducing the number of instances in the majority class, the dataset becomes more balanced, allowing machine learning algorithms to learn from both classes more effectively. As said above, the percentage of churners is less than 5% so we would prefer to apply balanced sampling and get a share equal to 50% for each category. There is also an additional step, the adjustment of balanced sampling, which refers to additional steps taken after under-sampling to ensure that the resulting dataset maintains enough information from both classes. This step is essential, as we ensure that the retained data from the majority class are still representative of its overall distribution. Another technique that is widely used for the above-mentioned processes is SMOTE which generates new observations for the minority class with characteristics close to these of the already existing observations and leads to a more balanced dataset. SMOTE technique has also the advantage that avoids overfitting which is a common danger for oversampling techniques (because when we use oversampling the model seems the same observations for a class many times and this can lead to overfitting)

Question 7:

From the pie chart below we notice that there are totally 51 customers with 6 or more calls to customer service and the vast majority (62,7%) of them are churners. Given that we should include the customer service department of XYZ in the effort to decrease the number of churners by recording which customers make 6 or more calls and find a way for this department to serve customers more efficiently and quickly.

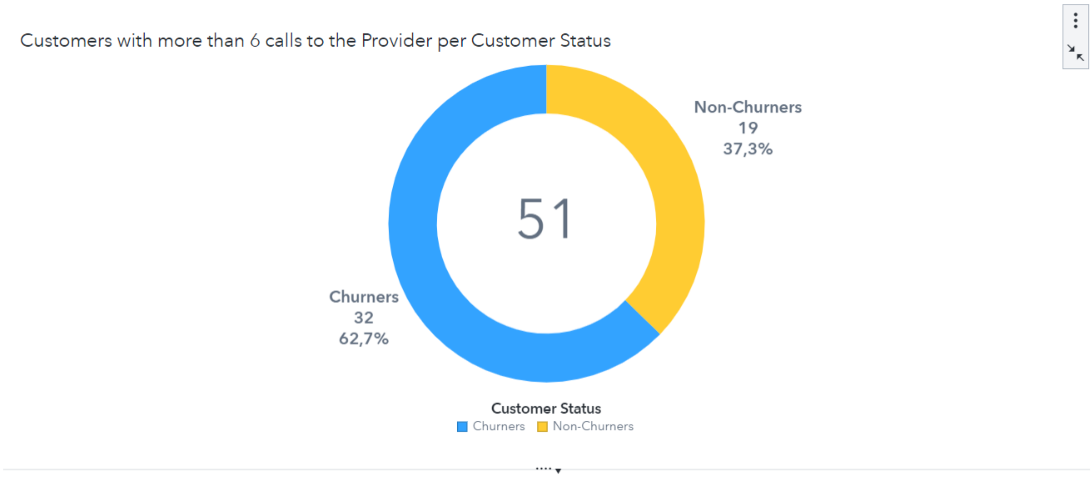


Figure 26: Customers with more than six calls to the provider per customer status

Question 8:

In the bar chart below we can see the average amount of minutes that the customers speak on their phone during the day for churners and non-churners. We observe that churners speak on average approximately 300 minutes more than non-churners. This difference may indicate that there are customers that are not satisfied with the package of services that they currently have, and they may need to have the opportunity to talk for more minutes at a better price.

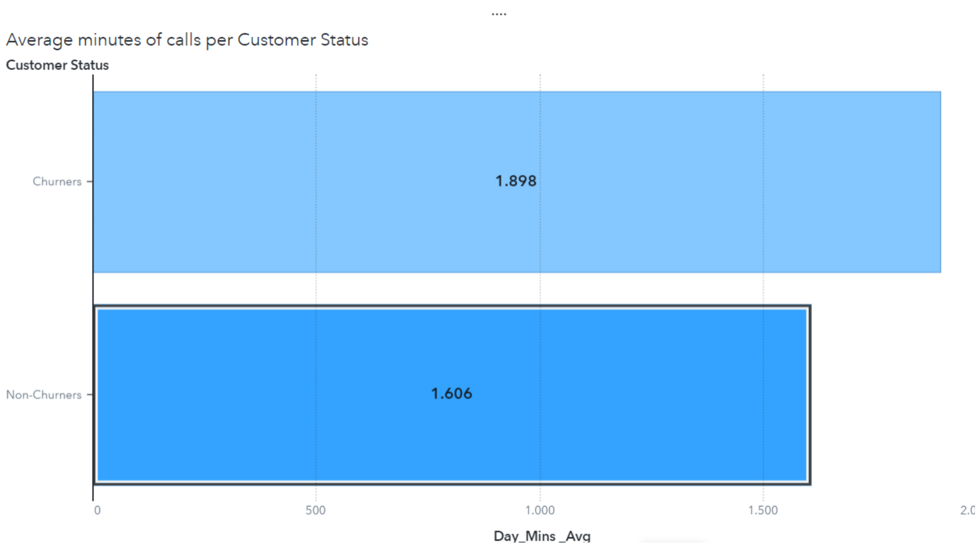


Figure 27: Average minutes of calls per Customer Status

Question 9:

The variable used for the first in the decision tree model (i.e. the most important variable) is International Plan which is a binary variable that indicates if the customer has an international plan or not. If the customer has an international plan, then we are directed to the left split and if not then we are directed to the right one. In this way, we can interpret the whole tree and not only the first split and state that if a customer meets the condition that is posed on a node, then he/she is directed on the left otherwise he/she is directed to the right part of the split. Missing values are directed in the left branch of a split, but we remind that in our case we do not have any missing values in the dataset, so this function of the tree is not crucial. However, a decision tree has the advantage that can handle missing values if it is needed and make predictions despite the presence of NAs.

The international plan variable is selected for the first split using the logworth statistic which evaluates the predictive power of each variable individually. The decision tree is constructed using the CHAID algorithm which exploits the chi-square statistical test. Logworth is calculated using the Chi-square test so using this test evaluates whether there is a significant difference in the distribution of the target variable across the categories or values of the predictor variable. A higher logworth indicates for better predictive ability of the candidate variable and a lower logworth indicates for worse predictive ability. International plan is selected for the first split because it has the highest logworth among all variables. The variable selection for the other splits will become again according to their logworth so variables with bigger logworth will be placed higher in the tree and those with lower logworth will be placed lower.

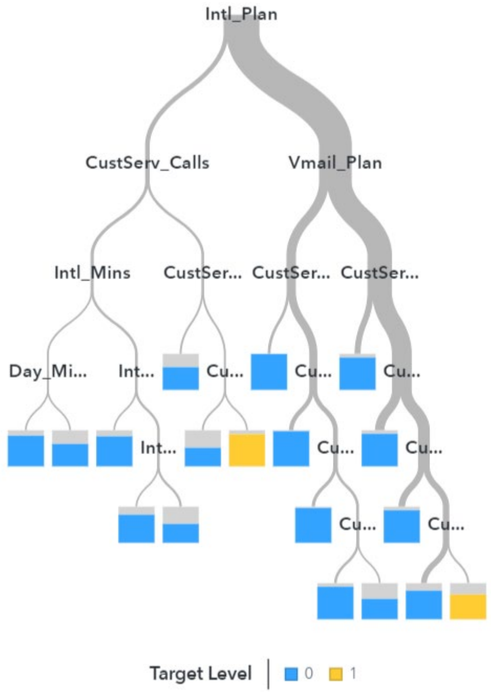


Figure 28: Optimal Decision Tree

Question 10:

The second decision tree that we created is called Maximal and, in our case, has 22 terminal leaves. This tree is called maximal because no pruning method has been applied and it is the "largest" possible tree.

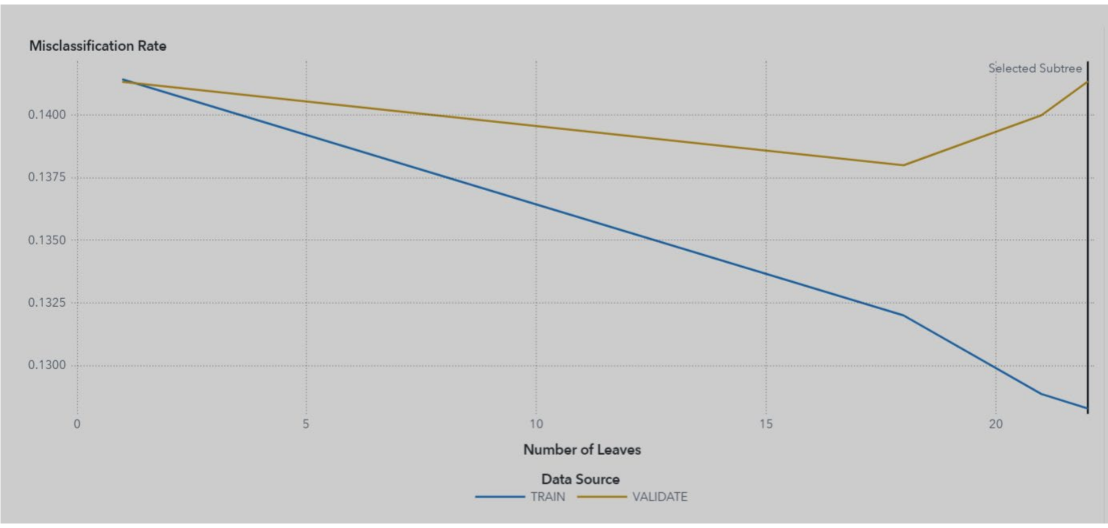


Figure 17: Subtree Assessment Plot

In the above error plot, we can see how the tree model with different number of leaves each time performs on the training and validation data sets (blue and orange lines respectively) using Misclassification Rate as a criterion. We see that the blue line (training data) is always descending which is expected because a predictive model performs better on the training data with more leaves. However, we are not interested in evaluating the model in the training data set but in the validation data set which consists of data that the model "sees" for the first time. We see that the orange line is descending for the most part but as the number of leaves increases suddenly the misclassification rate increases too. This phenomenon where the decision tree performs significantly better in the training data set compared to the validation is called overfitting. This is because the excessive number of leaves has led the tree to get too dependent on the training data and not be able to perform well and generalize its predictions in new data. To avoid overfitting that exists in this model we must prune the tree at the point where the line for the validation data set (orange line) presents a global minimum this action will give us the optimal number of leaves and the optimal decision tree.

An image of the maximal tree that we have constructed is given below:

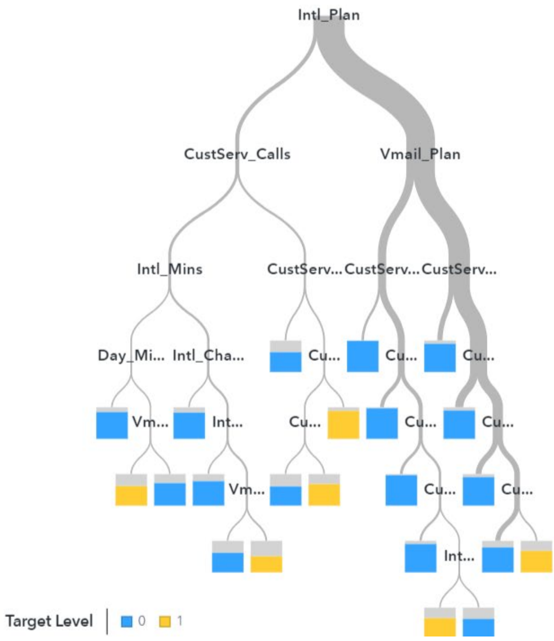


Figure 18: Maximal Decision Tree

Question 11:

Below we display the assessment plot for the optimal decision tree. We notice that the subtree that has been selected as optimal has 18 terminal leaves and the number of leaves for the optimal tree is selected at the point where the assessment line in the validation data set meets a global minimum as we said earlier.

This fact indicates that the optimal decision will have the best possible performance on newly seen data.

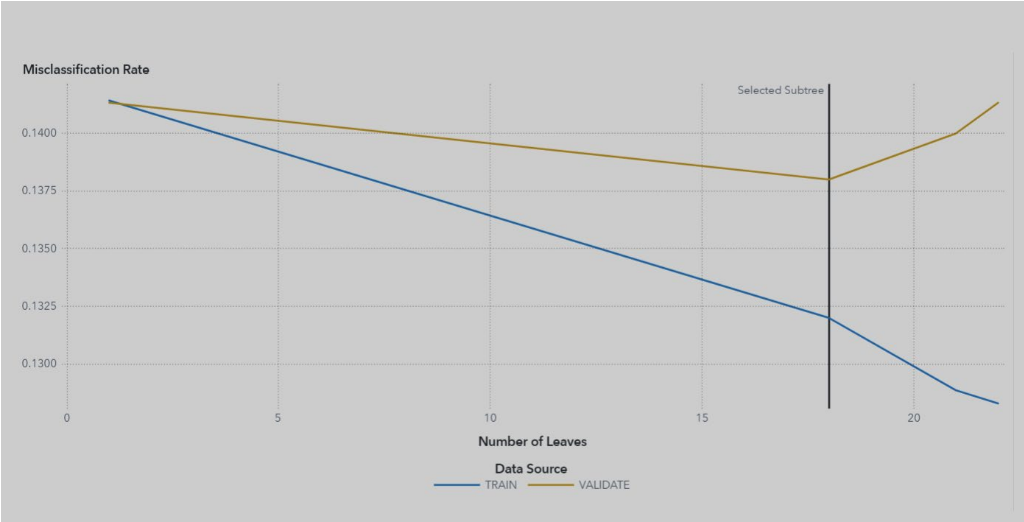


Figure 29: Optimal Decision Tree - Assessment plot

The optimal tree (18 leaves) is once again shown below:

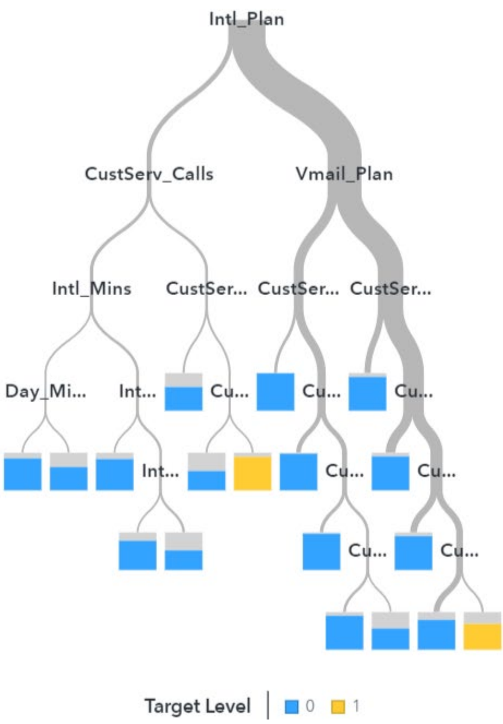


Figure 30: Optimal Decision Tree

Question 12:

As we said before the decision tree model was constructed through CHAID algorithm which uses the chi-square statistical test to recursively separate the data into distinct groups for the target variable (churn) the chi-square test tests to identify significant associations between predictor variables and the churn variable these associations are considered significant if they have a small p-value (< 0.05) or a large logworth (which is calculated based on the p-value and has an inverse relationship with it). Another important point for the construction of the model is the cut-off point which was found as described in question 3 and is equal to 0.16. Customers with predicted probabilities above this threshold are assigned to the churner's category, while those below are characterized as non-churners. The decision tree model is interpreted as we explain below for 5 terminal leaves indicatively.

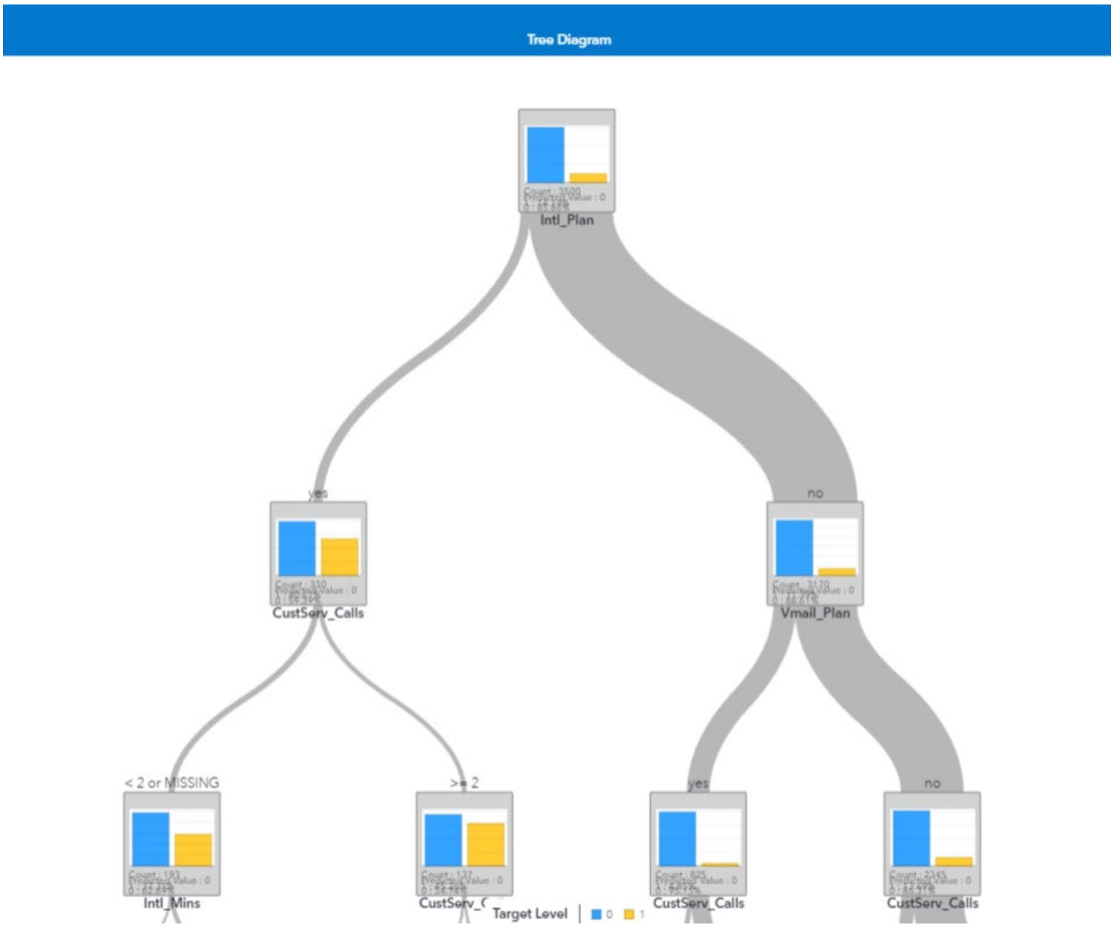
Leaf ID	Posterior Prob.	Decision	Rule
20	p1=88% p0=12%	p1>16%=>Churner	If Int.Plan=Yes & Cust.Serv.Calls>=5 => Churner
21	p1=2% p0=98%	p1<16%=>Non-Churner	If Int.Plan=No & V.Mail Plan=Yes& Cust.Serv.Calls<2 or Missing => Non-Churner
34	p1=68% p0=32%	p1>16%=>Churner	If Int.Plan=No & V.Mail Plan = No & Cust.Ser.Calls>=5 => Churner
13	p1=10% p0=90%	p1<16% => Non-Churner	If Int.Plan=No & V.Mail Plan = No & Cust.Serv.Calls<1 or Missing => Non-Churner
27	p1=3% p0=97%	p1<16%=>Non-Churner	If Int.Plan=No & V.Mail Plan=Yes & Cust.Ser.Calls>=2 & Cust.Ser.Calls <3 or Missing => Non-Churner

Question 13:

- If a customer does not have an international plan nor a voice mail plan and has made five calls or more to the customer service during the period in examination, according to the predictions of the decision tree he will become a churner.

- If a customer does not have an international plan nor a voice mail plan and has made three calls or more, but not five to the customer service during the period in examination or the number of these calls is missing, according to the predictions of the decision tree he will become a churner.
- If a customer does not have an international plan nor a voice mail plan and has made two calls to the customer service during the period in examination or the number of these calls is missing, according to the predictions of the decision tree he will not become a churner.
- If a customer does not have an international plan nor a voice mail plan and has made one call to the customer service during the period in examination or the number of these calls is missing, according to the predictions of the decision tree he will not become a churner.
- If a customer does not have an international plan nor a voice mail plan and has not made any calls to the customer service during the period in examination or the number of these calls is missing, according to the predictions of the decision tree he will not become a churner.

Our decision tree analysis reveals several key factors that differentiate churners from non-churners in our customer base. Firstly, customers who do not have an international plan or a voice mail plan and have made five or more calls to customer service during the examined period are highly likely to churn. Similarly, those who have made three or more calls to customer service but not five, or where the number of calls is missing, are also at risk of churning. On the other hand, customers who do not have an international plan or a voice mail plan and have made two calls to customer service or where the number of calls is missing are unlikely to churn. Furthermore, those who have made only one call to customer service or where the number of calls is missing are even less likely to churn. Finally, customers who do not have an international plan or a voice mail plan and have not made any calls to customer service during the examined period or where the number of calls is missing are the least likely to churn.


Tree Diagram
Intl_Plan
yes
no
CustServ_Calls
Vmail_Plan
< 2 or MISSING
>= 2
yes
no
Intl_Mins
CustServ_Calls
CustServ_Calls
Target Level
0
1

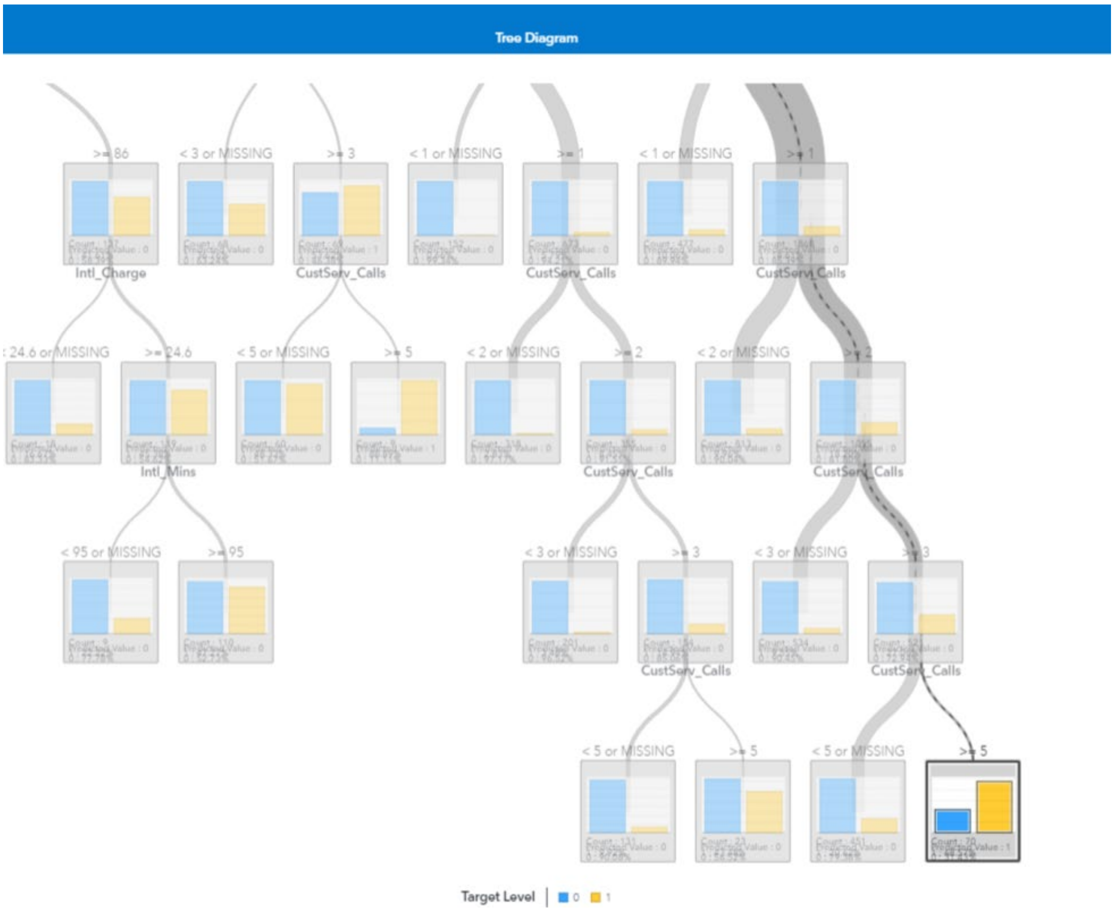


Figure 31: Optimal Decision Tree - 5 terminal leaves

Question 14:

The best model is the logistic regression model. According to this model and considering the cumulative response (%), if we focus on the top 20% of the customers who have the highest probabilities (calculated by the logistic regression model) 35% of these customers will churn. Similarly, if we target the 100% of the customers and sort them in descending order according to the probabilities calculated by the logistic regression model, 14,13% of these customers will churn.

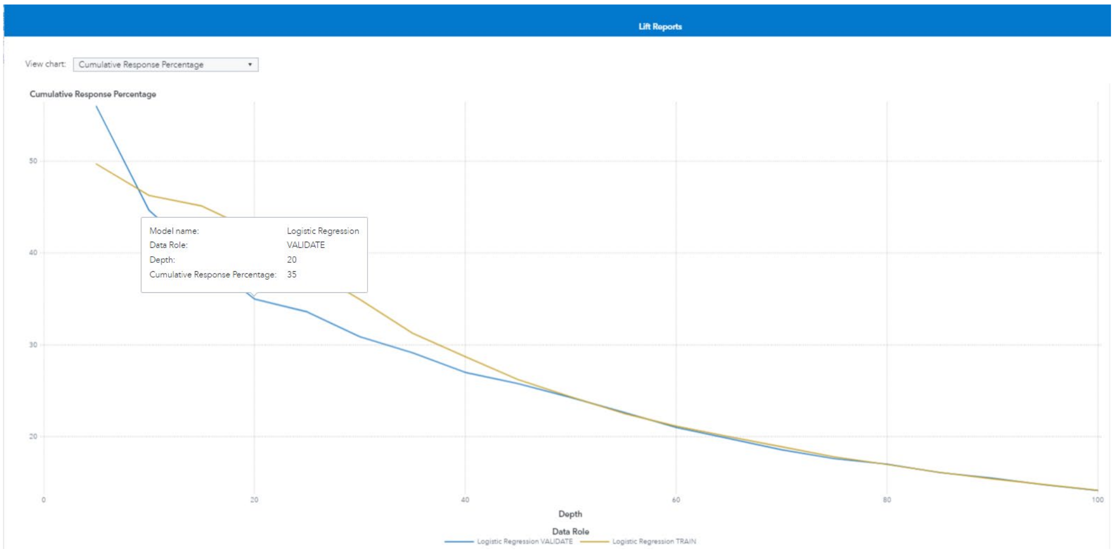


Figure 32: Cumulative Response (20%)

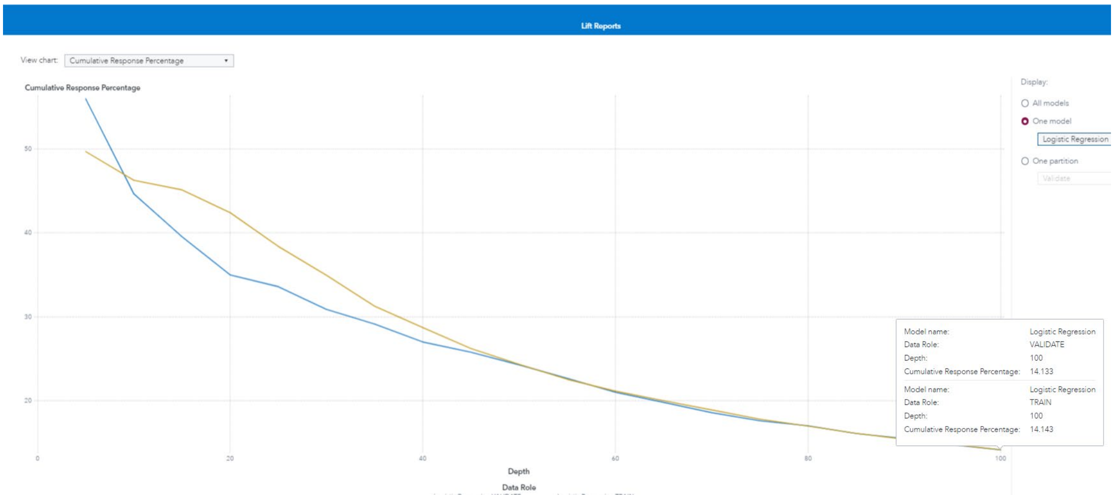


Figure 33: Cumulative Response (100%)

Question 15:

The % response chart for the validation dataset is constructed by plotting the cumulative percentage of true positives/responses (i.e. the customers that are correctly predicted as churners by the model) against the cumulative percentage of the population, sorted by the predicted probability of the positive class. More specifically, Depth represents the cumulative percentage of the sorted (according to the calculated probabilities) customers who are separated in buckets (i.e. in which portion/bucket of customers we will focus on).

Now, focusing on the 25% point on the x-axis, it If we target the fifth bucket (20% - 25%) of the customers with the highest probabilities (calculated by the logistic regression model), the 28% of this bucket will be churners.

Figure 34: Response (20%) Chart

Question 16:

If we target the 20% of the customers who have the highest probabilities (calculated by the logistic regression model), we will capture 2,47 times more churners than we would if we did the same job at random.

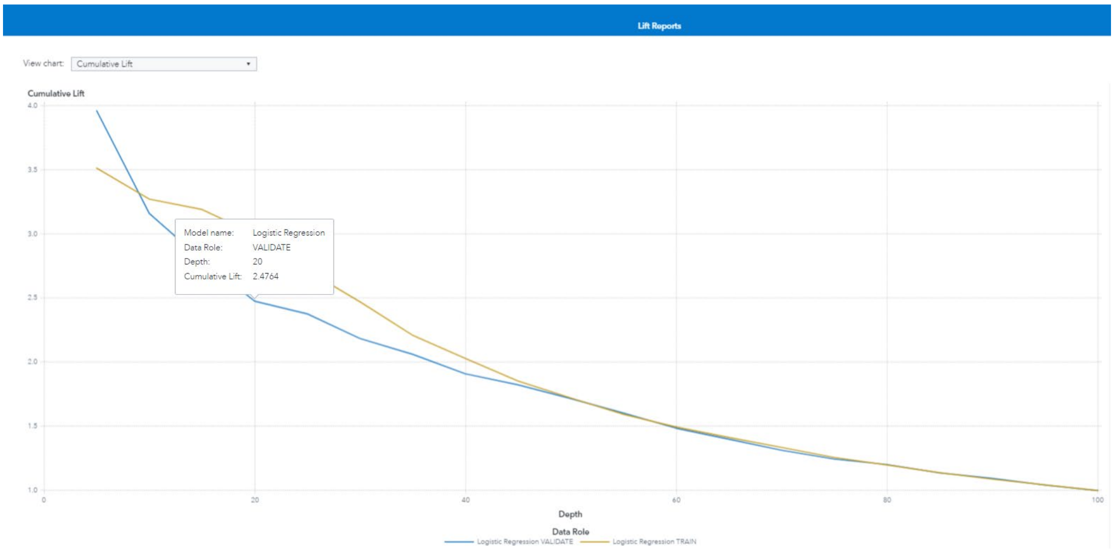


Figure 35: Cumulative Lift Chart (20%)

Question 17:

If we target the 40% of the best customers who have the highest probabilities (calculated by the logistic regression model), we expect to capture 76.41% of all churners of the validation dataset.

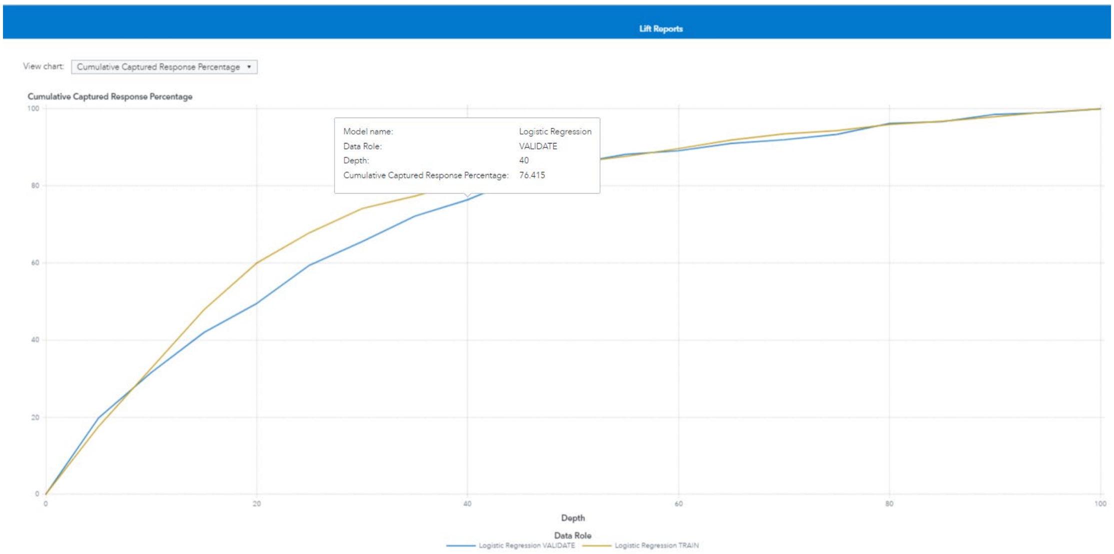


Figure 36: Cumulative (40%) Captured Response

Question 18:

In our dataset "telco_data_apr_sep" there are 1.884 customers in total. As we can see from the below plot, 1.365 of them have been predicted as non-Churners and 519 have been predicted as Churners.

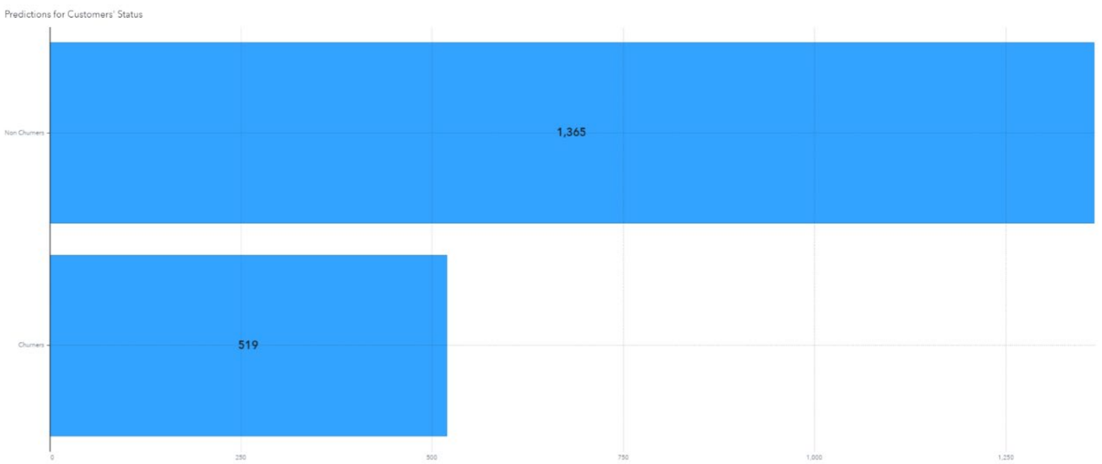


Figure 37: Predictions for Customers' Status

Question 19:

According to the best model, which is the logistic regression model, the smallest probability of a customer to be a churner is 16% (as expected, because the cut-off point in 16%) and the biggest corresponds to 98%.

Partition Indicator	Phone	State	Vmail_Message	Vmail_Plan	churn	Into: churn	Predicted: churn=1	P_churn0	Probability for churn =1 ↓	Predicted for ch...
1	388-6479	SC	46	yes	0	1	0.1604681966	0.8395318034	0.1604681966	1
1	368-8964	DC	0	no	0	1	0.1603658002	0.8396341998	0.1603658002	1
1	419-3427	ND	0	no	0	1	0.160130703	0.839869297	0.160130703	1
1	354-4045	HI	34	yes	0	1	0.1600704431	0.8399295569	0.1600704431	1
1	395-6149	PA	0	no	0	1	0.1600007882	0.8399992118	0.1600007882	1
1	396-5534	PA	0	no	0	0	0.1597352272	0.8402647728	0.1597352272	0

Figure 38: Smallest Probability for Churn prediction

Partition Indicator	Phone	State	Vmail_Message	Vmail_Plan	churn	Into: churn	Predicted: churn=1	P_churn0	Probability for churn =1 ↓	Predicted for ch...
0	416-2778	MN	20	yes	1	1	0.9838642908	0.0161357092	0.9838642908	1
1	376-7004	NJ	0	no	1	1	0.9727380118	0.0272619882	0.9727380118	1
1	384-4938	NC	0	no	1	1	0.958946922	0.041053078	0.958946922	1
1	406-5281	KS	0	no	1	1	0.9534646169	0.0465353831	0.9534646169	1
1	379-7890	TX	0	no	1	1	0.9384179896	0.0615820104	0.9384179896	1
0	382-4771	WV	0	no	1	1	0.9214242565	0.0785757435	0.9214242565	1

Figure 39: Biggest Probability for Churn prediction

Question 20:

The software assigns customers to Churners and non-Churners (1/0) based on the “Probability of churn” column. If the probability is higher than the cut of point, which is 0,16, the customer is predicted to be Churner, otherwise they are characterized as non-Churner. Taking customer with phone number 352-5412 as an example, their probability of churn is equal to 0,03, which is less than 0.16. As expected, they seem to be non-Churner, assigned with 0 in column “Predicted for churn”.

Partition Indicator	Phone ↑	State	Vmail_Message	Vmail_Plan	churn	Into: churn	Predicted: churn=1	P_churn0	Probability for churn =1	Predicted for ch...
1	352-5118	NM	15	yes	1	1	0.3411644202	0.6588355798	0.3411644202	1
0	352-5393	AL	0	no	0	0	0.0732639824	0.9267360176	0.0732639824	0
1	352-5412	CO	0	no	0	0	0.0394059741	0.9605940259	0.0394059741	0
1	352-5466	NV	0	no	1	1	0.227627136	0.772372864	0.227627136	1

Figure 40: Selected Customer

List of Tables and Figures

Figures:

Tables: