Business Analytics Practicum I

Group Assignment

Professor Name: Zaras Andreas

Team Members:

Mandyli Chrysoula – Charikleia F2822106

Taklakoglou – Chidiroglou Argyrios F2822114

SAS Account Credentials:

Username: f2822114@aueb.gr

Password: Argiris4321.

Περιεχόμενα

Case Study 1

Executive Summary

We are analysts hired by the on-line store “Buy-books-on-line.com”, that sells books about science and information technology. The sales department of the store wants to exploit cross selling opportunities to sell as many books as possible. To achieve this, we used the software SAS Visual Data Mining and Machine Learning on SAS Viya, to identify association rules between transactions. Our analysis was based on a dataset with 19,805 past sales transactions related to the “Business Analytics” book category, gathered by the on-line store. After importing the historical data, I have used SAS software to learn more about those associations by performing a market basket analysis. The main benefit of conducting market basket analysis is that it uncovers hidden purchasing patterns by customers i.e. which products sell together well, so retailers can run specific campaigns/promotions to cross-sell the items (bundling of two items). Thus, we ended up that customers interested in “Managerial Analytics” should receive advertising proposals for “Web Analytics 2.0 & Implementing Analytics , those interested in Implementing Analytics should receive proposals for Managerial Analytics & Data Science and Big Data Analytics .Moreover, to those that search for Customer Analytics For Dummies we must advertise Enterprise Analytics & Decision Analytics and for those that like or search for Enterprise Analytics we should advertise Managerial Analytics & Customer Analytics For Dummies . Analyzing the association rules, we have identified that the 3 books most bought together by customers are `Data Analytics made accessible`, `Data Science and Business Analytics` and `Business Analytics for managers` . So it is wise to apply an appropriate shelving and promote to users that proceed to the purchase of one of those 3 books, the pair of the other two.

2) In Figure 1, we can see the number of sales in units for each book. On the y-axis we have all the 56 books in descending order according to their sales. On the x-axis we have the count of the sales. Next to each book's bar we can see the exact number of sales for the specific book. As we can see from the graph, the book that sold the most copies is the "Data Science and Business Analytics" with 1596 sales, and the one with the least sales is the "Managerial Analytics" that sold 152 copies.

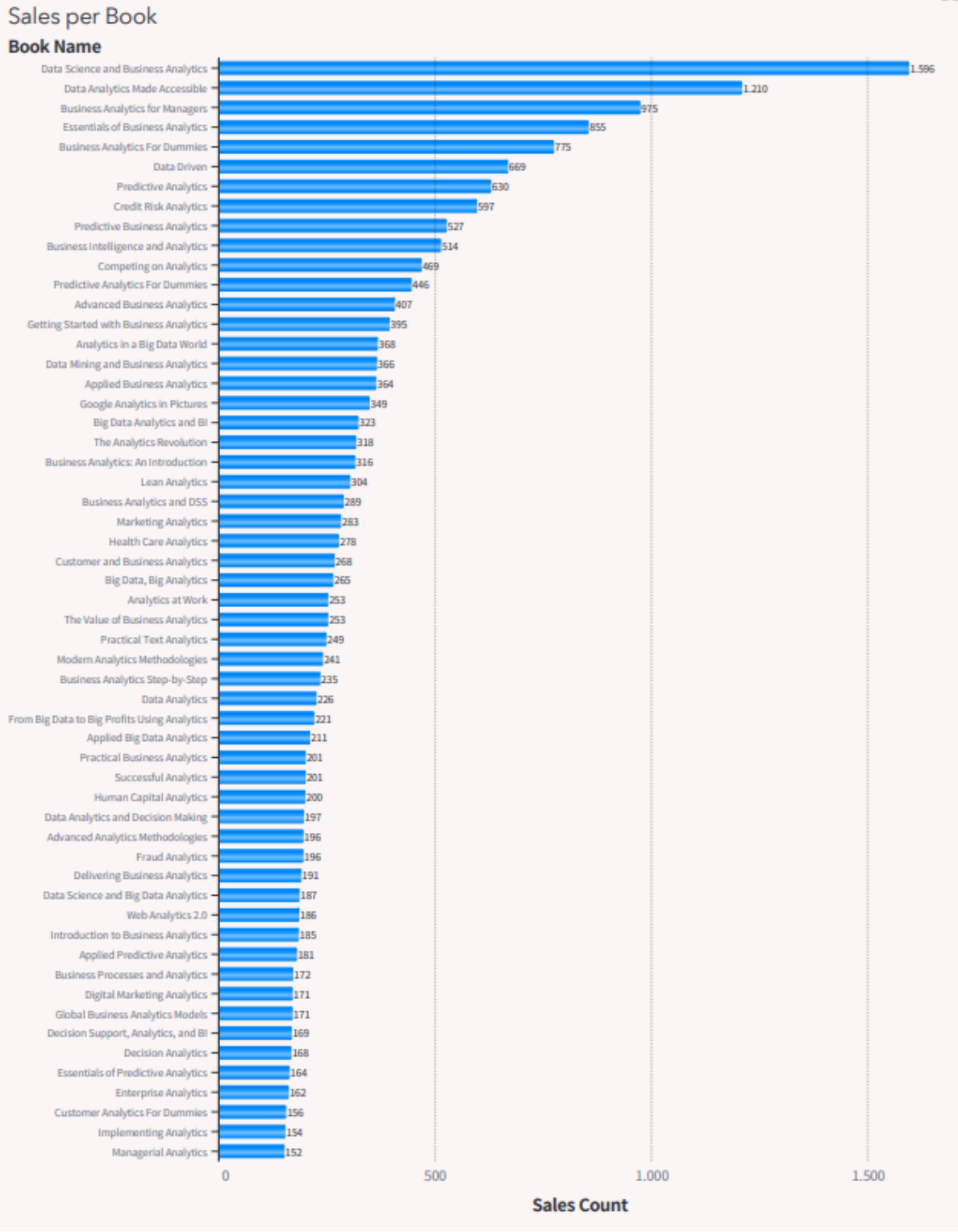


Figure 1: Number of Sales per Book

3) After implementing our analysis on book transactions, we focused on the following 4 books:

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

Specifically, we identified which two books we should propose to every user that is interested on buying one of these analytics books.

First of all, for customers that are interested in "Managerial Analytics" it would be wise to recommend 'Web Analytics 2.0" and "Implementing Analytics". The lift of this rule is equal to 11.47. This means that if a customer buys the "Managerial Analytics" book it is 11.47 times more likely to buy both "Web Analytics 2.0" and "Implementing Analytics", from a random customer i.e. that has not bought it.

Regarding "Implementing Analytics" book, the two books that should be proposed to customers that are interested in it, are "Managerial Analytics" and "Data Science and Big Data Analytics". This rules lift is equal to 11.33. So, if a customer buys the "Implementing Analytics" book it is 11.33 times more likely to buy both "Managerial Analytics" and "Data Science and Big Data Analytics", from a random customer i.e. that has not bought it.

Furthermore, to a customer that is interested in "Customer Analytics for Dummies" we should advertise "Enterprise Analytics" and "Decision Analytics". The lift is 11.19 thus, if a customer buys the "Customer Analytics for Dummies" book it is 11.19 times more likely to buy both "Enterprise Analytics" and "Decision Analytics", from a random customer i.e. that has not bought it.

Finally, a customer that is searching to buy "Enterprise Analytics" should receive a recommendation for the books "Managerial Analytics" and "Customer Analytics for Dummies". This rule has a lift that is equal to 11.07. This means that if a customer buys the "Enterprise Analytics" book it is 11.07 times more likely to buy both "Managerial Analytics" and "Customer Analytics for Dummies", from a random customer i.e. that has not bought it.

4) By setting the maximum items in a rule to 3, we can identify that the 3 books most bought together by customers are "Data Science and Business Analytics", "Data Analytics Made Accessible" and "Business Analytics for Managers". After performing an association-rule analysis, we saw that the combination of these three books appeared at 794 transactions. In Figure 2 we can see all the possible rules that are created with these three items.

	LHS	RHS	COUNT	SUPPORT	ITEM1	ITEM2	ITEM3	RULE
9	2	1	794	41.877637131	Data Analytics Made Accessible	Data Science and Business Analytics	Business Analytics for Managers	Data Analytics Made Accessible & Data Science and Business Analytics ==> Business Analytics for Managers
10	1	2	794	41.877637131	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
11	1	2	794	41.877637131	Data Analytics Made Accessible	Business Analytics for Managers	Data Science and Business Analytics	Data Analytics Made Accessible ==> Business Analytics for Managers & Data Science and Business Analytics
12	2	1	794	41.877637131	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
13	2	1	794	41.877637131	Business Analytics for Managers	Data Science and Business Analytics	Data Analytics Made Accessible	Business Analytics for Managers & Data Science and Business Analytics ==> Data Analytics Made Accessible
14	1	2	794	41.877637131	Data Science and Business Analytics	Business Analytics for Managers	Data Analytics Made Accessible	Data Science and Business Analytics ==> Business Analytics for Managers & Data Analytics Made Accessible

Figure 2: Association Rules for the 3 books most bought together

Support metric is the percentage of transactions that contain all of the items in an itemset. In our itemset the items are “Data Science and Business Analytics”, “Data Analytics Made Accessible” and “Business Analytics for Managers”. For that combination of books the software gives us a support metric of 41.87% . This number is calculated via the mathematical formula **support = Freq(X,Y)/N**. Freq(X,Y) is the number of purchases that contain those 3 books , where X is the left hand’s rule products and Y the right hand’s rule. In our case is the number in the count column. N denotes the number of customer purchases in total. If we calculate it we get 794/1896=0.4187 which is the same number that the software gave us. The higher the support metric , the more frequently the itemset occurs. Rules with a high support are preferred since they are likely to be applicable to a large number of future transactions.

Case Study 2

Executive Summary

We are Marketing Analytics consultants hired by Sports-OnLine.com, an on-line retailer that sells sport clothes and shoes. The management team of the store wants to exploit the electronic data captured the previous years to better understand the market, so we were asked to perform a customer segmentation analysis. We decided that the most suitable technique for our case was a Recency Frequency Monetary (RFM) analysis. To achieve that we used the machine learning software SAS Visual Data Mining and Machine Learning in SAS Viya. Our analysis was based on some historical data gathered by the on-line store, during the period October 2001 – December 2006. These records came from 995 customers that have done 4906 sales transactions. After importing the data, we have used an unsupervised method called k-means for clustering. K-means identifies clusters/partitions that refer to a collection of data points aggregated together because they have certain similarities. Thus, we saw that there are 4 groups of customers with similar shopping activities. Each cluster depicts a segment with clients considered as good, bad, churners and new customers. After our analysis we have identified that actions like reactivation programs, contact customers for feedback, cross sell activities or special promotions are proposed. Chiefly, we should attract first time clients, with some special promotions as they count for 26,3% of total customers and 20% of total customer monetary. Moreover, we must focus on churners, as they are 30,6% of total customers and 36,7% of monetary and that is a significant percentage. It could be wise to enable a reactivation program or some special promotions, and feedback to identify mistakes and make our business strategy more efficient. The percentage of the bad customers is 20.9% of the total customers. This is a quite high percentage so it would be wise to try making these customers better buyers. Lastly, loyalty credits and cross selling or upselling, should be provided to good customers as they count for 34,3% of total monetary. Detailed proposals for each segment can be found inside the main body of our case study.

We began our analysis by importing the data to the software. Firstly, we wanted to create some plots to help us visualize our data and understand them better.

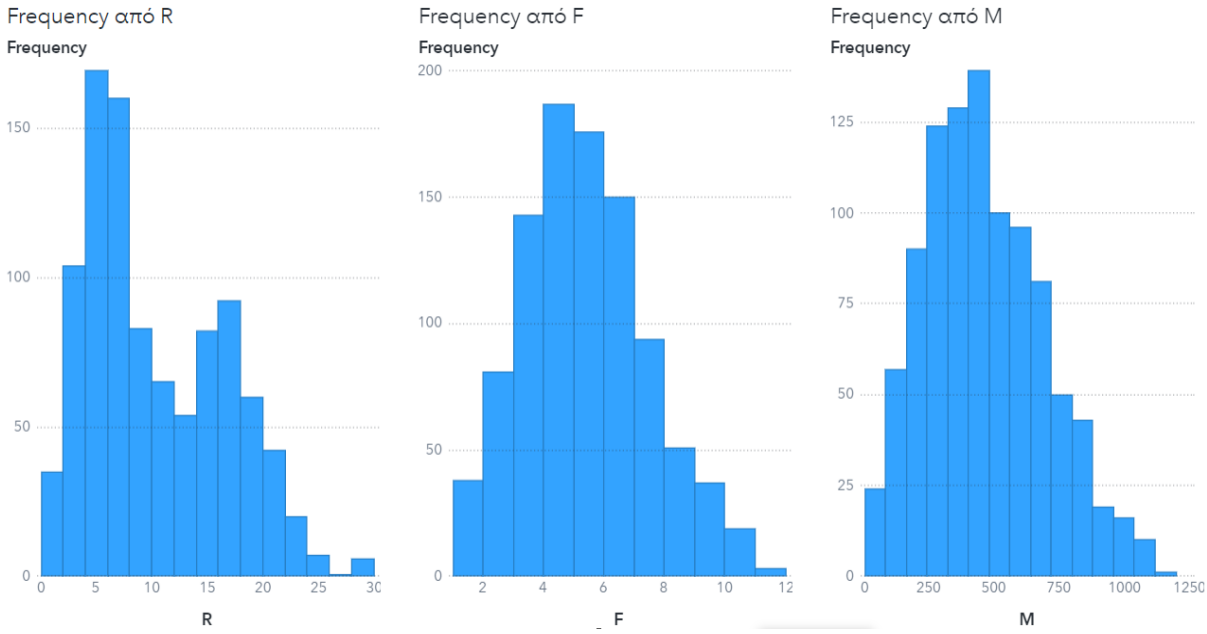


Figure 3: Histograms for R, F, M

In Figure 3 we see some histograms for recency, frequency and monetary. The first one is for Recency and we can comprehend that it has 2 peaks. The 1st peak depicts that the majority of our customers have 2-7 months to make a purchase from our company. The 2nd peak depicts that many customers have 15 or more months to make a purchase. The second graph is for Frequency. From this graph we understand that the majority of the customer base have purchased 3-7 times from our company. Finally, the third one is for Monetary. It shows that the majority of the customer base has given to the company 200-800€.

After that we wanted to ckeck if there are outliers in our data so we made some boxplots for the variables R,F,M. As we can see in Figures 4,5,6 our data have some outliers. These outliers are depicted in the plots as dots above the whiskers. The median for the variable R is 8 months, for the variable F is 5 times of purchases and for the variable M the amount of 435€.

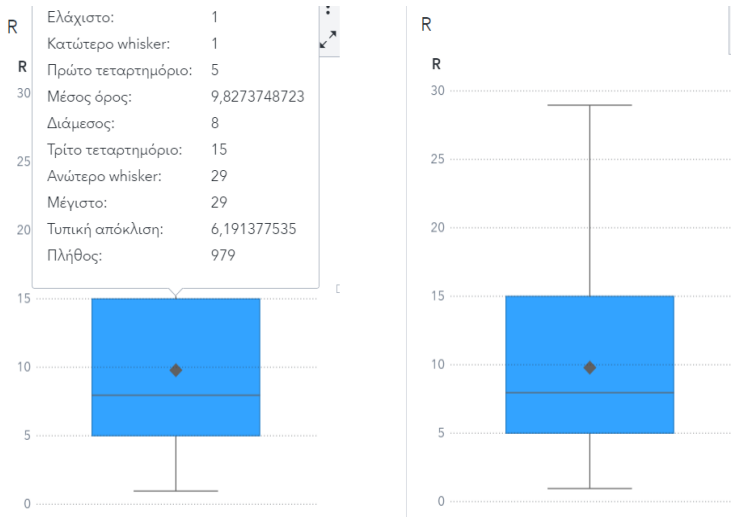
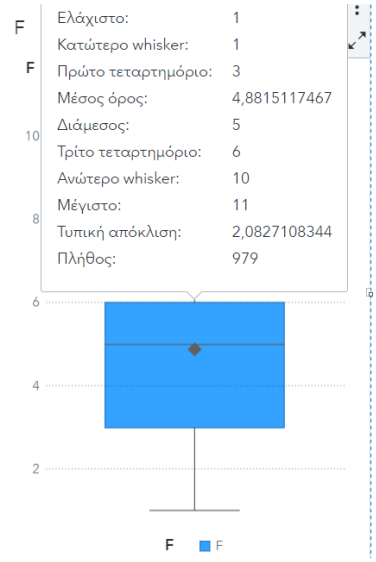


Figure 4: Boxplot for R variable



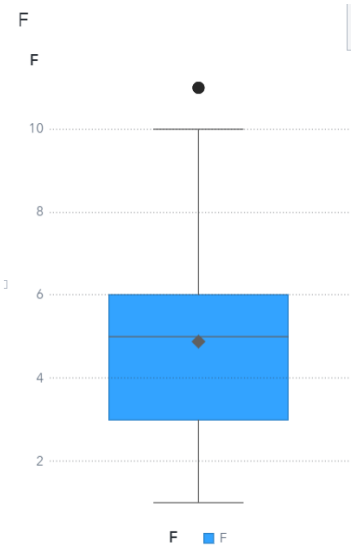
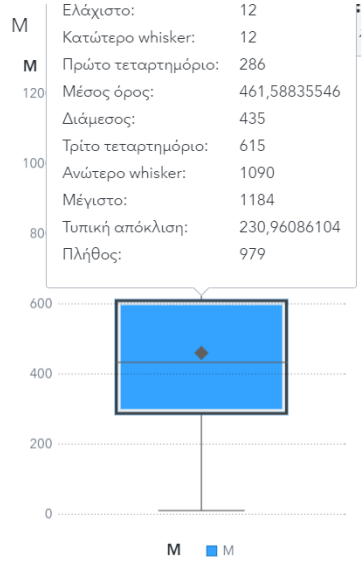


Figure 5: Boxplot for F variable



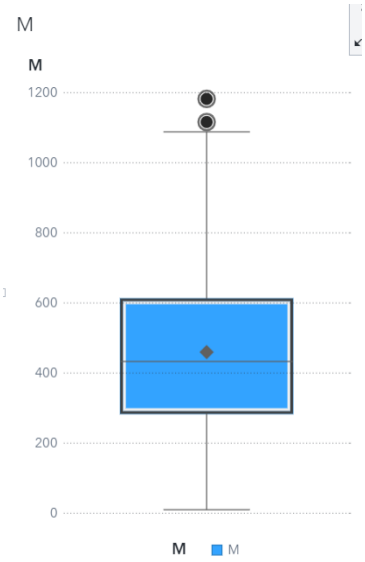


Figure 6: Boxplot for M variable

Based on the results that we conclude from the graphs, we performed some preprocesses to make our data more useful. Specifically, we did some filtering to exclude the observations that were outliers from our data, and performed a log-transformation to the input variables. Initially we had 995 customers. The filtering action excluded 16 observations, so we ended up having 979 customers to use for our analysis.

We used the k-means algorithm for our clustering creation. Our purpose was to make clusters that are as similar as possible concerning the observations inside the segment, and observations between clusters to be as dissimilar as possible. As we can see on the Figure 7 the clustering model identified 4 segments.

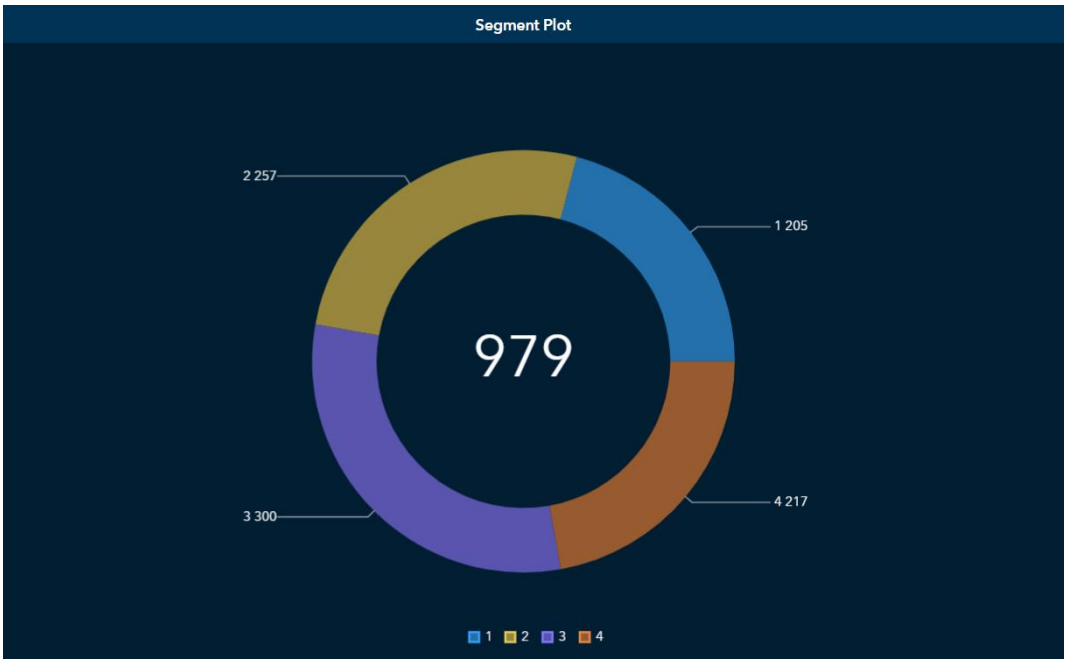


Figure 7: Segment pie chart

In Table 1 we can see some details about the 4 clusters that were created, and the Recency, Frequency and Monetary values of these clusters. We did this, in order to have a better knowledge of our customers. Also, we can find insights about each cluster, and we can do different marketing actions and promotions in each cluster based on the company's needs. The values on this table are referring to the mean characteristics of the ideal client of each cluster. At the bottom of the table there is a row that depicts the mean characteristics of a customer generally.

- The column Frequency holds the information about how often the cluster's customer purchases items from our company.
- The column Monetary embody the information about how much money the ideal customer from each segment has given to the company.
- The column Recency contain the information about the months that have passed from the last purchase of the typical customer of each cluster.

Cluster ID	Frequency	Ποσοστό συχνότητας	R	F	M
1	205	20,94%	15,717073171	2,3512195122	196,62439024
2	257	26,25%	5,2217898833	4,1089494163	352,29961089
3	300	30,64%	13,813333333	5,53	552,87
4	217	22,17%	4,2073732719	7,2903225806	715,13824885
Άθροισμα	979	100,00%	9,8273748723	4,8815117467	461,58835546

Table 1: Mean characteristics of each cluster's ideal customer

We wanted to name each segment, based on the performance of their ideal customer in the three main categories, compared to the mean characteristics of an individual in our data. We did that by evaluating the values in the R, F, M columns for each cluster. We coloured the values in these columns with green and red, based on whether they are better from the mean value that appears in the last row or worse, respectively.

Regarding the recency (R), we want it to be lower that the mean recency. On the other hand, we wish that the frequency and monetary are higher. Based on this, we gave proper names to each group as we can see on table 2.

Cluster ID	Segment Names	Frequency	Ποσοστό συχνότητας	R	F	M
1	Bad Customers	205	20,94%	15,717073171	2,3512195122	196,62439024
2	First Time Customers	257	26,25%	5,2217898833	4,1089494163	352,29961089
3	Churners	300	30,64%	13,813333333	5,53	552,87
4	Good Customers	217	22,17%	4,2073732719	7,2903225806	715,13824885
Άθροισμα		979	100,00%	9,8273748723	4,8815117467	461,58835546

Table 2: Clusters names and mean characteristics

We named the first segment as 'Bad Customers' because the Frequency is lower than the average Frequency of all segments. This means that they haven't bought many times items from our company. The Monetary is also lower than the average Monetary of all segments. This means that when they buy items from our company, they don't spend much money. Lastly, the Recency is higher than the average Recency of all segments. This means that many months have passed from the last time they did a purchase from our company. To be more accurate, the customer that belongs to this category has purchased approximately 2.3 times, the total money that has given to the company is 196.62 € and the last time he made a purchase was 15.7 months before.

We named the second Segment as 'First Time Customers' because the Frequency is lower than the average Frequency of all segments. This means that they haven't bought many times items from our company because they are new customers. The Monetary is also lower than the average Monetary of all segments. This means that they haven't spent much money, because they haven't made so many purchases as they are new customers. Lastly, the Recency is lower than the average Recency of all segments. This means that few months have passed from the last time they did a purchase from our company. To explain it more, the customer that belongs to this category has purchased approximately 4.1 times, the total money that has given to the company is 352.3 € and the last time he made a purchase was 5.2 months before.

We named the third Segment as 'Churners' because the Frequency is higher than the average Frequency of all segments. This means that they have bought many times items from our company because they used to be good customers. The Monetary is also higher than the average Monetary of all segments, because they used to spend a lot of money. Lastly, the Recency is higher than the average Recency of all segments. This means that many months have passed from the last time they did a purchase from our company. Moreover, we see that the typical customer that belongs to this category has purchased approximately 5.5 times, the total money that has given to the company is 552.9 € and the last time he made a purchase was before 13.8 months.

We named the fourth Segment as 'Good Customers' because the Frequency is higher than the average Frequency of all segments. This means that they have bought many times items from our company. The Monetary is also higher than the average Monetary of all segments. This means that they spend a lot of money on our items. Lastly, the Recency is lower than the average Recency of all segments. This means that few months have passed from the last time they did a purchase from our company. From the Table 2 we can see, that the customer that belongs to this category have purchased approximately 7.3 times, the total money that has given to the company is 715 € and the last time he made a purchase was before 4.2 months.

After naming and describing the segments, we continued our analysis by making some pie charts. Our goal as consultants, is to propose marketing actions appropriate for each segment and help the organization gain more money. Thus, our proposals will be based on those charts. The first graph is about the size of each segment, and the proportion of each group to the customer base. The second one is referring to how the total monetary is separated in the 4 clusters that we created. We can see these pie charts in Figures 8 and 9 respectively.

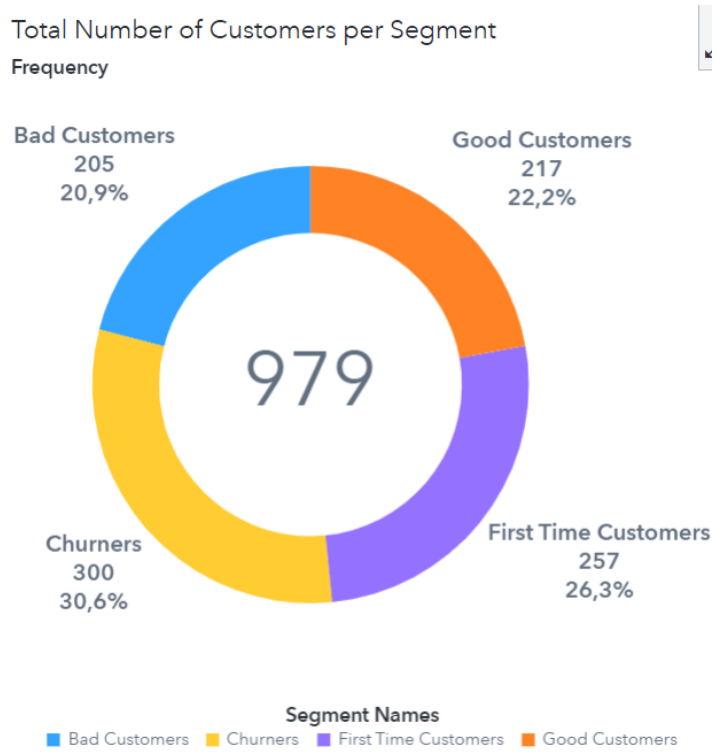


Figure 8: Total number of Customers pe Segment

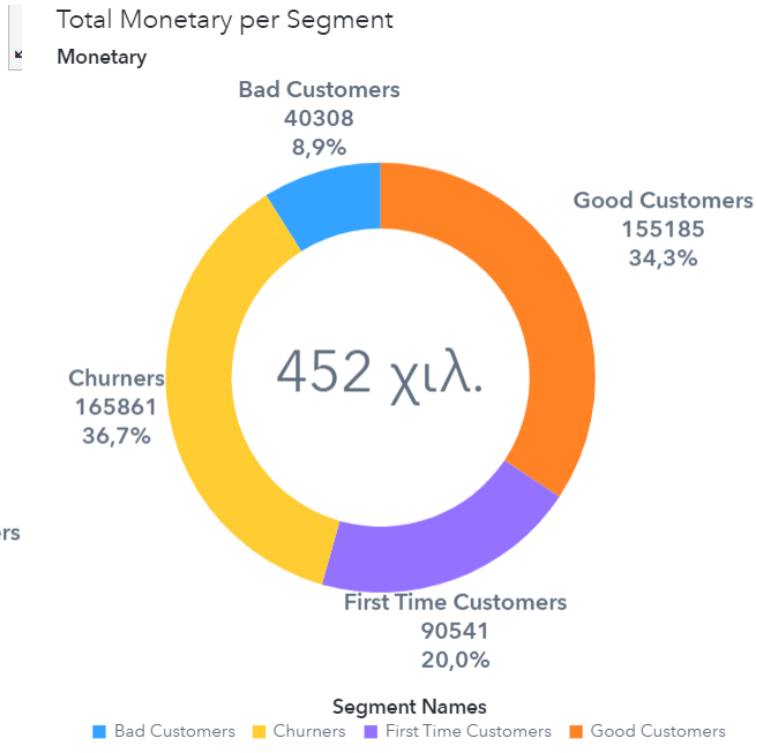


Figure 9: Total Monetary per Segment

After combining information coming from both charts, we can come to some proposals based on our analysis results. Thus, regarding these segments we could come to the following business wised actions:

- **Segment 1 – Bad Customers:** We have 205 customers which comprise Bad Customers cluster, and they constitute the 20.9% of the customer base. Besides, the total amount of money that these 205 customers have given to our company is 40308 € that corresponds to 8.9% of the total company's income. The aim is to make them more active customers if possible. So, an appropriate marketing actions for this segment will be to contact them and have some feedback about how they see the company and what do they not like. Also, some promotions could also be helpful for our purpose.

- **Segment 2 – First Time Customers:** We have 257 customers which comprise First Time Customers cluster, which constitute the 26.3% of the customer base. Besides, the total amount of money that these 257 customers have given to our company is 90541 € that corresponds to 20% of the total company's income. It's obvious that the purpose is to make these customers become good buyers and avoid becoming "one-time buyers" because the total amount of money they have given to the company is more than twice as much as Bad Customers have given us. If they become good buyers, the frequency of the purchases would be higher and as a result the total money they spent in buying athletic clothes and items for our company would be much more. So, an appropriate marketing actions for this segment will be to make frequently special offers to the customers belonging to these segment (a small

discount to some specific products that most of the customers in this cluster may be interested in)

- **Segment 3 – Churners:** We have 300 customers which form Churners cluster, which make up the 30.6% of the customer base. Furthermore, the total amount of money that these 300 customers have given to our company is 165861 € that corresponds to 36.7% of the total company's income. So, we should place great importance on making Churners be active customers again because a great amount of the company's income came from the customers belonging to this cluster when they were good customers. Specifically, we should try to make them good customers as they used to be. An appropriate marketing action for this segment , could be to make frequently special offers to these customers. In depth, we could make bigger or smaller discounts (based on the percentage of the profit we have for each item) for every item we sell for the next 3 purchases in a specific period of time. In addition, we should contact these customers not only via email but also via call in order to take honest feedback about the company. Like this, we can understand why they have stopped being good buyers and we may make them feel important for us.

- **Segment 4 – Good Customers:** We have 217 customers which form the Good Customers cluster. This cluster is the 22.2% of the customer base. Furthermore, the total amount of money that these 217 customers have given to our company is 155186 € that corresponds to 36.7% of the total company's income. Thus, undeniably, the customers belonging in this segment along with the churners, are the most important for the company because the total income from these two categories account for 70,1% of the total company's income. So, we should make them feel important. In order to keep them loyal, we should send them frequently emails with items that are in sale and also to send them a card or a barcode in order to have a standard small discount in checkout. Also, we should enhance some cross selling and upselling activities.

Case Study 3

Executive Summary

We are machine learning engineers hired by the insurance organization XYZ to help the management department with it's current campaign. The target is to identify segments of customers who are likely to purchase a variable annuity. Specifically, our task is to aid the insurance organization develop a machine learning based customer response model, that will predict whether a customer will prove to be a buyer of the insurance product or not if he/she is solicited. The software that we used to perform our statistical analysis was SAS Visual Data Mining and Machine Learning on SAS Viya. The model was built under some historical data that were created from a random sample of this year's customers who were solicited. The aim was to apply the model to the rest of the customer's database next year to predict whether the rest of the customers will buy the product or not. Firstly, we built a variety of models, specifically two decision trees, a logistic regression and a neural network model, in order to find the optimal one. By comparing the models, we found out that the best model for our analysis was a decision tree. After applying our model to the score data, we came in the conclusion of which customers are predicted to be buyers and which are not. The customers that were predicted to buy the insurance product after solicitation were 629 out of the 3013 that the dataset contained.

2) Our aim is to find the minimum cost after the solicitation of the customers. It is not wise to solicit all the customers, because each solicit has a cost. The management team of the marketing department provided to us the following profit matrix.

		Prediction	
		Contact – Solicit	**Ignore**
Actual	**Responder**	800	0
	Non-Responder	-300	0

Figure 10: Profit Matrix

From the above matrix we conclude that a successful solicitation i.e. if we solicit the customer and he ends up to be a responder, the company will have a profit of 800 €. On the other hand if we send a solicitation to a customer and he does not respond, the company will have a cost of 300 €. If we ignore a customer, there is neither a profit nor a cost for the company no matter if the customer is a buyer or a non-buyer. So, it is very important for the marketing department to choose wisely which customers will receive a solicitation. We want to target the customers that are more likely to respond positively to the new year's campaign, in order to maximize our profit for the company.

3) In order to calculate the cut - off point we should evaluate the expected profit if the insurance organization will sent a solicitation, and the expected profit if the insurance organization won't do that. The probability for a customer to be a responder is p1 and to be a non-responder is 1-p1.

Expected profit if a customer (possible buyer or possible non-buyer) will be solicited = 800p1 – 300(1-p1)

Expected profit if a customer (possible buyer or possible non-buyer) will be ignored = 0p1 + 0(1-p1) = 0

In order to do a solicitation, the expected profit if a customer will be solicited should be greater than expected profit if a customer will be ignored.

- ⇨ 800p1 – 300(1-p1) > 0p1 + 0(1-p1)
- ⇨ 800p1 – 300 + 300p1 > 0
- ⇨ 1100p1 > 300
- ⇨ p1 > 0.2727

So, the minimum probability (cut - off point) that a customer should have in order to be considered as a buyer, and hence to be considered for solicitation is **0.2727**.

After importing the data, we perfomed an EDA process. We set the data types of our variables according to the types that were given for the project. Then, we partitioned the data and created some charts to understand them better.

4) We partitioned the historical data set to training and validation using the 70% - 30% rule.

The training dataset is the set of data that is used to train the model and make it learn the features-patterns in the data.

The validation set is a set of data that is used to validate our model performance during training. The model is trained on the training set, and, simultaneously, the model evaluation is performed on the validation set. Splitting the dataset into training and validation sets helps to prevent our model from overfitting.

This process must be done because the model should be trained in as many as possible scenarios so as to be able to predict any unseen data sample that may appear in the future. At the same time is used to validate our model performance during training. We used the 70% - 30% rule to avoid overfitting. Moreover, we set the sampling in the data partition to be stratified. Stratified split, ensuring that both training and validation datasets will contain the same proportion of observations for buyers and non-buyers. In other words, we used stratification to ensure the same distribution of classes on both training and validation datasets. So, the training and validation datasets will contain observations corresponds to 70% for non-buyers and 30% for buyers. Also, we used the Misclassification Rate (Event) as the performance criterion and we changed the cut-off point to the one, we calculated previously.

5) In Figure 11 we can see that the dataset "insurance_campaign_history'' doesn't contain any missing values.

Variable Name ↑	Missing
AcctAge	0,0000
Age	0,0000
ATM	0,0000
ATMAmt	0,0000
Branch	0,0000
CashBk	0,0000
CC	0,0000
CCBal	0,0000
CCPurc	0,0000
CD	0,0000
CDBal	0,0000
Checks	0,0000
CRScore	0,0000
cust_id	0,0000

IRABal	0,0000
LOC	0,0000
LOCBal	0,0000
LORes	0,0000
MM	0,0000
MMBal	0,0000
MMCred	0,0000
Moved	0,0000
MTG	0,0000
MTGBal	0,0000
NSF	0,0000
NSFAmt	0,0000
Partition_Indicator	0,0000
Phone	0,0000
POS	0,0000

DDA	0,0000
DDABal	0,0000
Dep	0,0000
DepAmt	0,0000
DirDep	0,0000
HMOwn	0,0000
HMVal	0,0000
ILS	0,0000
ILSBal	0,0000
InArea	0,0000
Income	0,0000
Ins	0,0000
Inv	0,0000
InvBal	0,0000
IRA	0,0000

POSAmt	0,0000
Res	0,0000
Sav	0,0000
SavBal	0,0000
SDB	0,0000
Teller	0,0000

Figure 11: Variables' names and missing values

The proportion of buyers and non-buyers in the data set is 70% non-buyers and 30% buyers as we can see in the pie charts in Figures 12 and 13. The total number of the customers that are non-buyers is 21089 and for the buyers is 9039.

Figure 12: Frequency of non-buyers

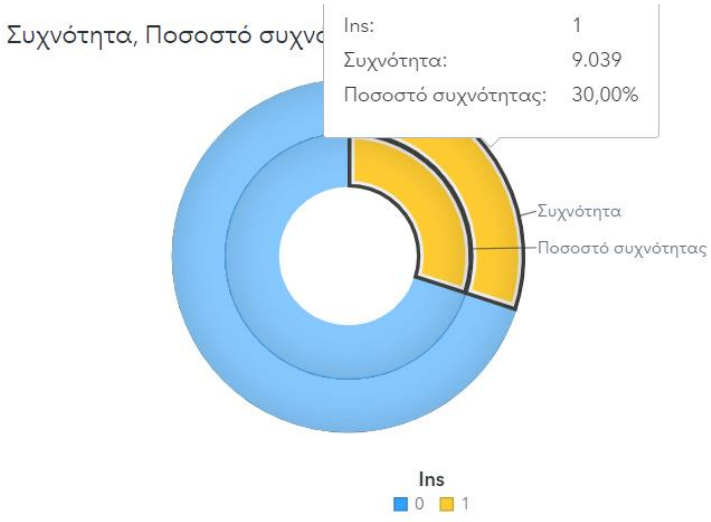


Figure 13: Frequency of buyers

6) According to the proportion of the buyers/non-buyers in our historical data set, we end up that our data are balanced. But if the proportion was 10%-90% for buyers/non-buyers respectively, the data would be imbalanced and specifically we would have an event of interest that is between 5-10%. In this case we would follow an empirical rule to do undersampling to our data. According to this rule we create a separate sample. In this new sample we take the 100% of the buyers i.e. all the cases of the event of interest, and an amount of non-buyers, such that the separate sample will contain 70% non-buyers and 30% buyers.

7) The proportion of buyers and non-buyers for those customers that have purchased more than 5 times credit cards is 100% buyers and 0% non-buyers. By looking the Figure 14, we can assume customers with more purchases of the credit card, are more likely to be buyers of a new insurance product.

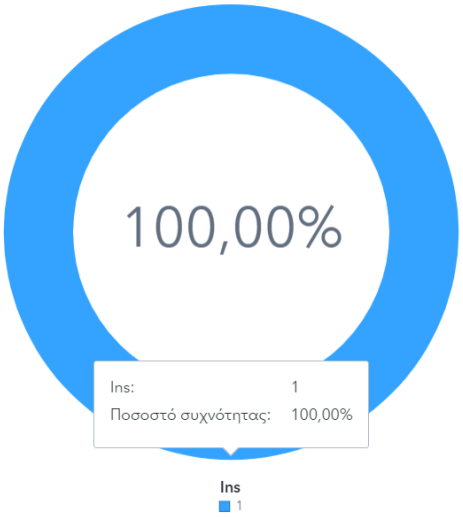


Figure 14: Proportion of buyers and non-buyers for those customers that have purchased more than 5 times credit cards.

8) In Figures 15 and 16, we see the average deposit amount for buyers and non-buyers for the last campaign. The customers that bought the product have given approximately 2650 € to our company, while the customers that have not bought the last campaign's product, have given 2040. As a result, it is more likely for the company to gain more money, if we try to approach the customers that have bought the last campaign's product. In the graphs, we can also see the total amount deposited for buyers/non-buyers.

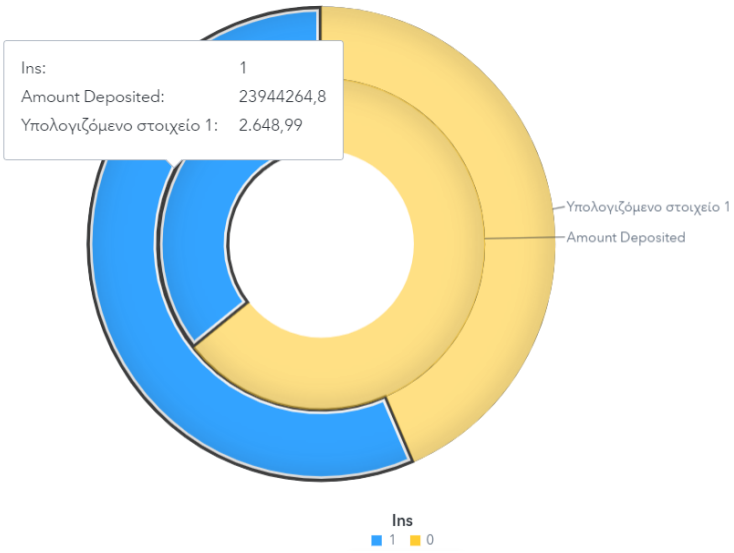


Figure 15: Average Deposit Amount for Buyers

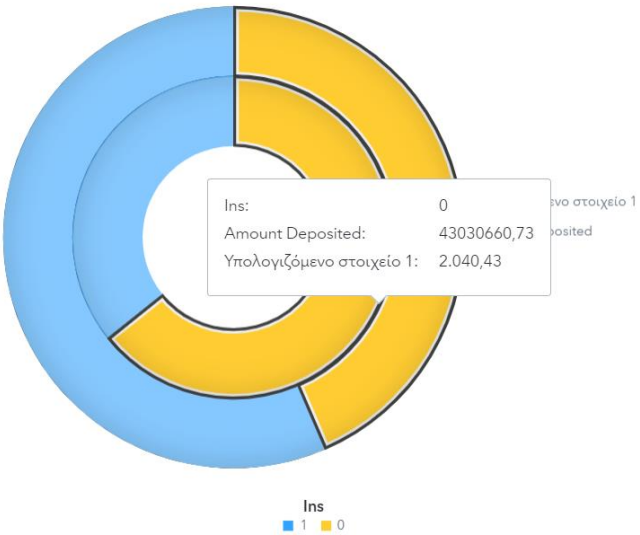


Figure 16: Average Deposit Amount for Non- Buyers

9) We continue our analysis by creating the models. The first model is a decision tree. The variable used for the first split is SavBal i.e. Saving Balance. This variable was chosen because it has the highest logworth. Logworth is a statistic metric, on which is based the decision about which variable will be selected in each split. The logworth statistic is a logarithmic transformation of the p-value and used to determine the contribution of each variable to the model. It measures the effectiveness of a particular split decision at differentiating values of the target variable. The bigger the logworth is, the more important this variable is for the model. Thus, when training the model, the variables that are selected to enter the tree are the ones with the highest logworth. Also, the variables that have higher logworth will be higher in the decision tree graph. As we can see in the figure 17, the customers that have a saving balance greater than or equal to 9102.86 are directed to the left node and the customer that their saving balance is less than 9102.86 or this value is missing are directed to the right node.

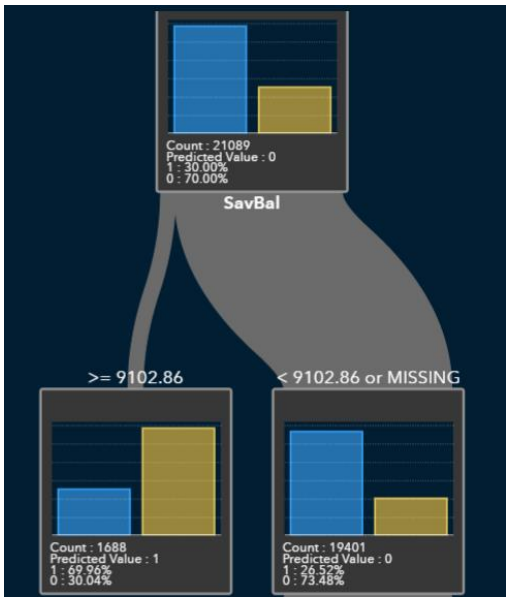


Figure 17: Decision tree's first split

10) The decision tree in Figure 18 is the maximal theoretical tree and it consists of 91 terminal leaves.

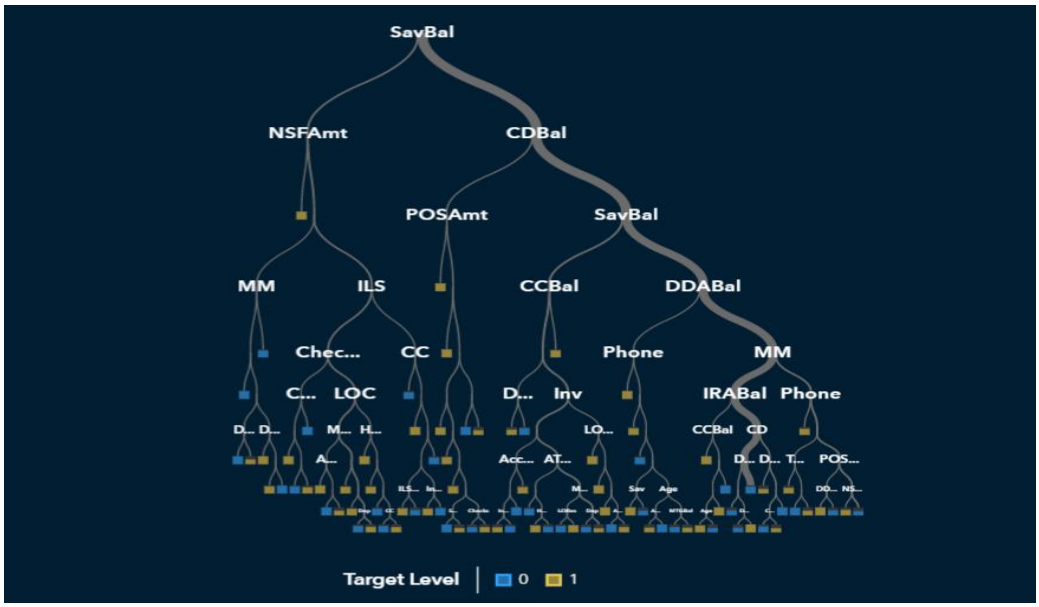


Figure 18: Maximal Decision Tree

The graph in Figure 19, is the subtree assessment plot for the maximal tree. The assessment criterion for a categorical target is the proportion of misclassified observations. The graph plots the Misclassification Rate on the y-axis and the Number of Leaves on the x-axis. It shows how the misclassification rate changes for subtrees, that are created by pruning the full decision tree to various numbers of leaves. The minimum misclassification rate indicates whether the model is a good classifier. The blue line depicts the training error for our maximal tree. Because this line is an assessment on the training dataset, there is bias. We can see that the training error decreases as the number of leaves increases, thus, the more leaves we add the less the error is. This phenomenon is called overfitting. To prevent that problem, we should use the validate partition to prune the full decision tree. The selected subtree for the decision tree model that we made, has a misclassification rate of 0.2601 for the validate partition (yellow line) and 91 leaves as it was mentioned before.

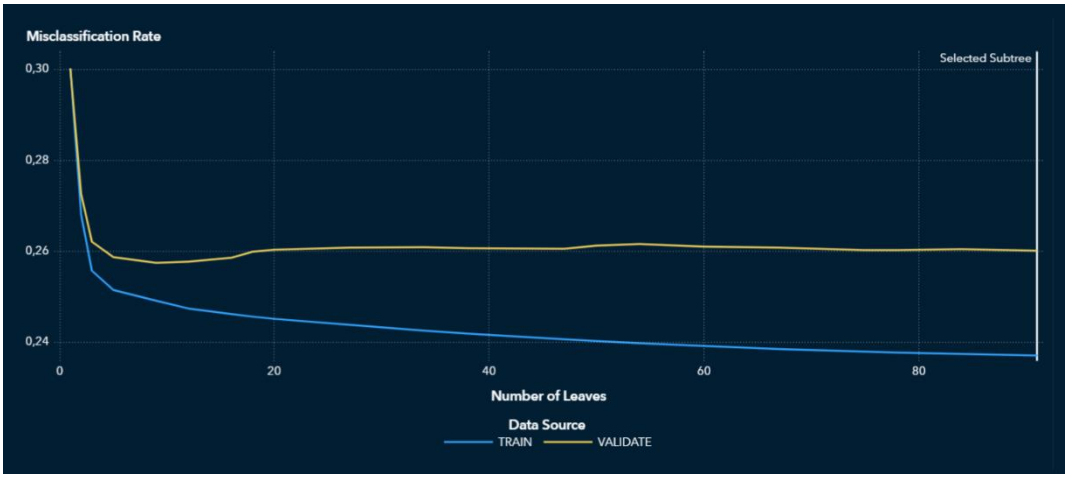


Figure 19: Maximal tree's subtree assessment plot

11) In Figure 20, we can see the optimal decision tree. This tree consists of 9 terminal leaves. The number of leaves is much lower than it was for the maximal tree. In Figure 21, we see the subtree assessment plot for the optimal tree. As it was mentioned before, the blue line depicts the training error for our tree and the yellow one the validation error. To prevent the overfitting phenomenon and perform an honest assessment, we evaluate the misclassification error of our model on validation data. The misclassification rate for the validate partition is 0.2574 and the selected subtree has 9 terminal leaves. That number gives us the best honest assessment on validation data.

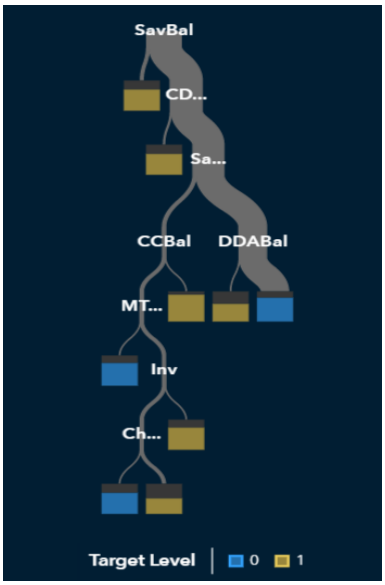


Figure 20: Optimal Decision Tree

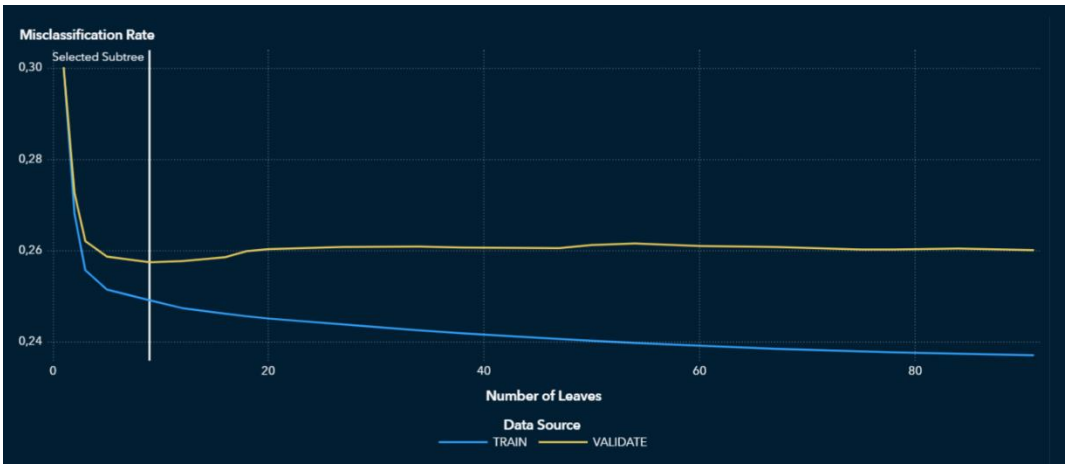


Figure 21: Optimal tree's subtree assessment plot

12) Now we are going to provide a description of the decision tree model. We will do this interpretation by using the 5 terminal leaves that are circled in the Figure 22. The target variable is coded as 1/ 0 and it indicates whether a customer from the solicited ones bought the insurance products or not (1=bought, 0=not bought). The decision, if the target variable will be classified as 1 or 0, is defined from the cut-off point. This point is equal to 0.2727. If the posterior probability for 1 is greater than the cut-off point, then the predicted value is 1 and if it is lower it is 0. In Table 3, we can see the posterior probabilities for a customer being classified as 1 or 0 and the decision that was made based on the comparison with the cut-off point.

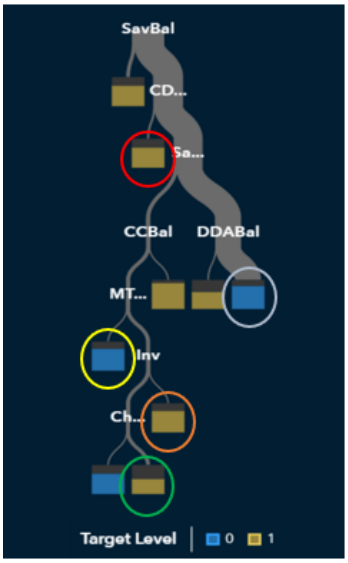


Figure 22: Optimal tree's selected leaves for interpretation

	Posterior Probability for 1	Posterior Probability for 0	Decision
Red Circled Leaf	69%	31%	1
Grey Circled Leaf	21.22%	78.78%	0
Yellow Circled Leaf	25%	75%	0
Orange Circled Leaf	73.47%	26.53%	1
Green Circled Leaf	51.58%	48.42%	1

Table 3: Posterior probabilities and Classification for the selected terminal leaves

Here we provide a technical interpretation for the leaves mentioned above:

- If SavBal < 9102.86 or missing and CDBal >= 13300 then the predicted value is 1, with probability 69%.
- If SavBal < 9102.86 or missing and CDBal < 13300 or missing and SavBal < 2709.642 or missing and DDABal < 14950.71 or missing then the predicted value is 0, with probability 78.78%.
- If SavBal < 9102.86 or missing and CDBal < 13300 or missing and SavBal >= 2709.642 and CCBal >= 0 and MTGBal >= 135061.4 then the predicted value is 0, with probability 75%.
- If SavBal < 9102.86 or missing and CDBal < 13300 or missing and SavBal >= 2709.642 and CCBal >= 0 and MTGBal < 135061.4 or missing and Inv = 1 then the predicted value is 1, with probability 73.47%.
- If SavBal < 9102.86 or missing and CDBal < 13300 or missing and SavBal >= 2709.642 and CCBal >= 0 and MTGBal < 135061.4 or missing and Inv = 0 and Checks < 16 or missing then the predicted value is 1, with probability 51.58%.

13) Now we want to describe the results of the decision tree model to the management team of the insurance organization. Thus, we will provide a business interpretation of the tree and not a technical one. We will interpret the same 5 terminal leaves that we did before.

- If a customer has a saving balance that is less than 9102.86 or this value is missing and a certificate of deposit balance greater than or equal to 13300, then he is classified as a buyer, with probability 69%.
- If a customer has a saving balance that is less than 9102.86 or this value is missing and a certificate of deposit balance less than 13300 or this value is missing, and a saving balance less than 2709.642 or this value is missing, and a checking account balance less than 14950.71 or this value is missing, then he is classified as a non-buyer, with probability 78.78%.
- If a customer has a saving balance that is less than 9102.86 or this value is missing and a certificate of deposit balance less than 13300 or this value is missing, and a saving balance greater than or equal to 2709.642, and a credit card balance greater than or equal to 0, and a mortgage balance greater than or equal to 135061.4, then he is classified as a non-buyer, with probability 75%.
- If a customer has a saving balance that is less than 9102.86 or this value is missing and a certificate of deposit balance less than 13300 or this value is missing, and a saving balance greater than or equal to 2709.642, and a credit card balance greater than or equal to 0, and a mortgage balance less than 135061.4 or this value is missing, and he has an investment account, then he is classified as a buyer, with probability 73.47%.
- If a customer has a saving balance that is less than 9102.86 or this value is missing and a certificate of deposit balance less than 13300 or this value is missing, and a saving balance greater than or equal to 2709.642, and a credit card balance greater than or equal to 0, and a mortgage balance less than 135061.4 or this value is missing, and he has not an investment account, and the number of his checks is lower than 16 or this value is missing, then he is classified as a buyer, with probability 51.58%.

As we can see on the Figure 23, the most important variables used are SavBal, CDBal, CCBal, DDABal, MTGBal, Inv and Checks. Our model is based in these features to class a new observation as buyer or non-buyer.

Variable Importance

Variable Label	Role	Variable Name	Validation Importance	Importance Standard Deviation	Relative Importance	Count
Saving Balance	INPUT	SavBal	329,3886	0	1	2
CD Balance	INPUT	CDBal	92,0109	0	0,2793	1
Checking Balance	INPUT	DDABal	29,7310	0	0,0903	1
Investment	INPUT	Inv	3,4579	0	0,0105	1
Number of Checks	INPUT	Checks	-0,7500	0	-0,0023	1
Mortgage Balance	INPUT	MTGBal	-0,7799	0	-0,0024	1
Credit Card Balance	INPUT	CCBal	-3,1983	0	-0,0097	1

Figure 23:Variable Importance

In the figures below, we see the score rankings overlay plots for the best model that we got after the model comparison. This model is the optimal decision tree. We will interpret these plots by focusing on the validation data set.

14)

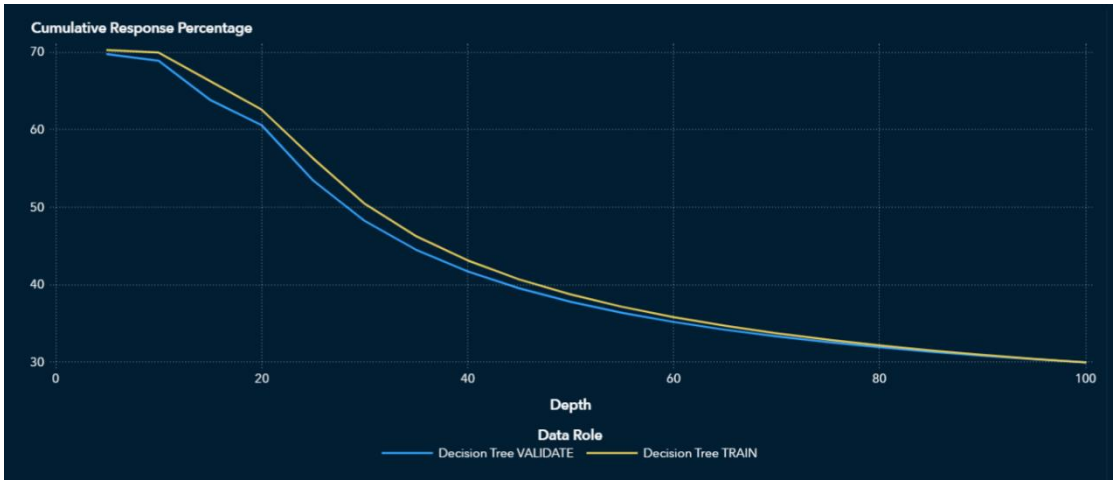


Figure 24: Cumulative Response Percentage Plot

If we solicit the 20% of the most highly ranked customers according to the probability that the best model gives them to be buyers, the 60.609% of this 20% will be buyers.

Model name:	Decision Tree
Data Role:	VALIDATE
Depth:	20
Cumulative Response Percentage:	60,609

If we solicit the 100% of the most highly ranked customers according to the probability that the best model gives them to be buyers, the 30.003% of this 100% will be buyers.

Model name:	Decision Tree
Data Role:	VALIDATE
Depth:	100
Cumulative Response Percentage:	30,003

15.

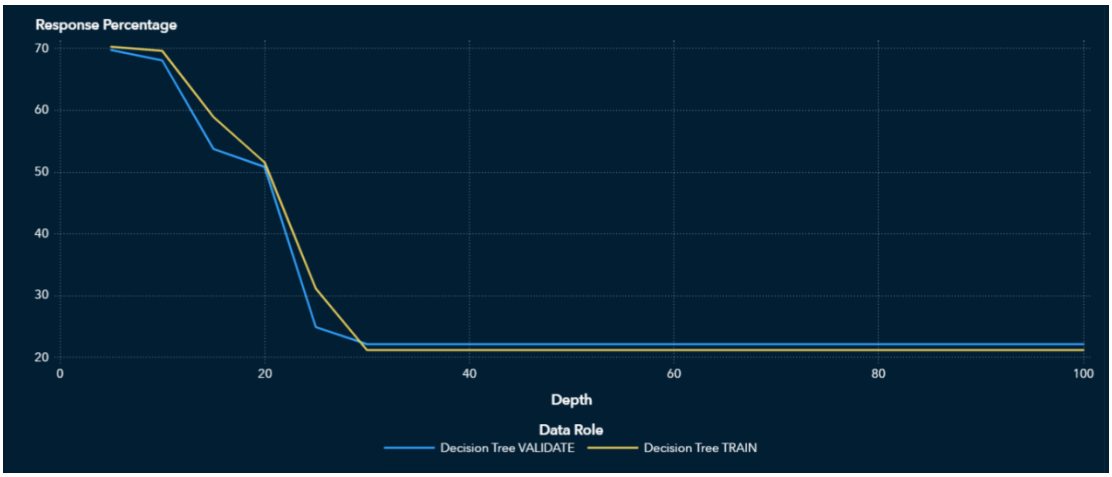


Figure 25: Response Percentage Plot

To create the % response plot, we separate data in partitions of 5% , thus each value on x axis represents a bucket of 5% of the data. If we solicit the fifth (5th) bucket (20% - 25%) of the most highly ranked customers according to the probability that the best model gives them to be buyers, the 24.931% of this bucket will be buyers.

Model name:	Decision Tree
Data Role:	VALIDATE
Depth:	25
Response Percentage:	24,931

16.

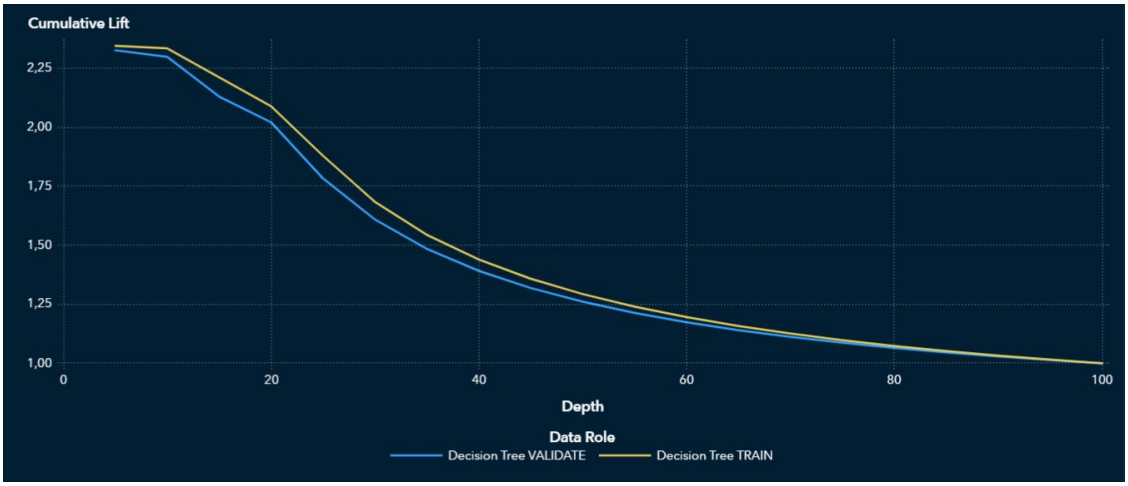


Figure 26: Cumulative Lift Plot

If we solicit the 20% of the most highly ranked customers according to the probability that the best model gives them to be buyers, we will capture 2.0203 times more buyers than if we did the same job without a model i.e. at random.

Model name:	Decision Tree
Data Role:	VALIDATE
Depth:	20
Cumulative Lift:	2,0203

17.

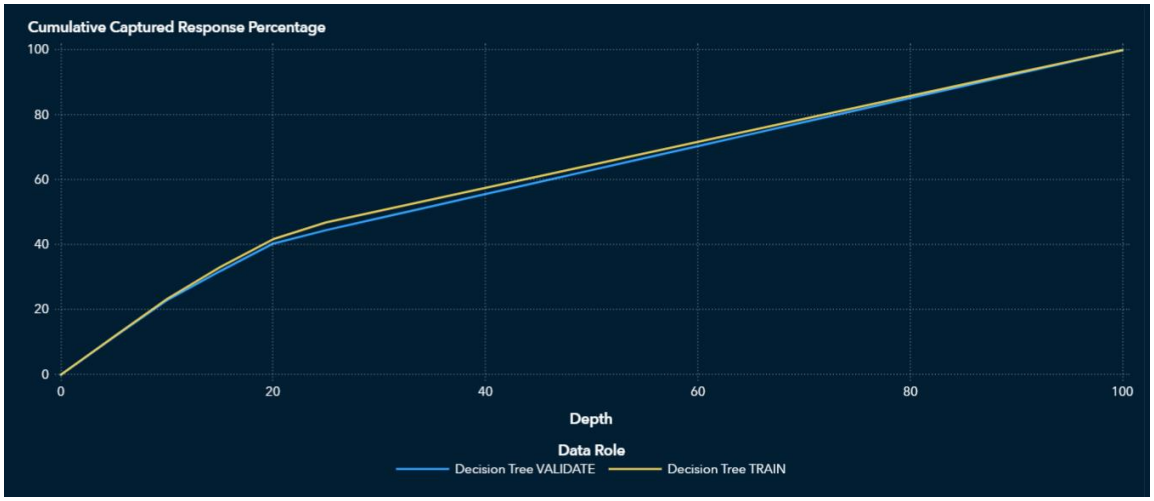


Figure 27: Cumulative Captured Response Percentage Plot

If we solicit the 40% of the most highly ranked customers according to the probability that the best model gives them to be buyers, we will capture the 55.651% of all the buyers of the whole validation data set.

Model name:	Decision Tree
Data Role:	VALIDATE
Depth:	40
Cumulative Captured Response Percentage:	55,651

In Figure 28, we can see the completed process flow.

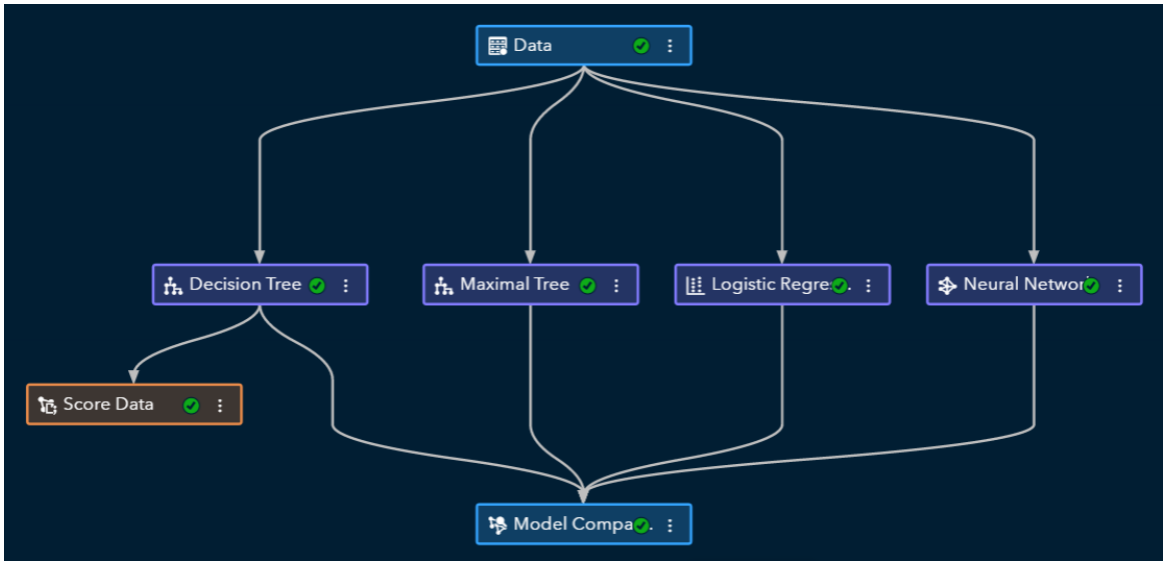


Figure 28: Completed Process Flow

18) The total number of customers in the "insurance_campaign_score" data set is 3013. These customers are classified as buyers and non-buyers. The number of buyers is 629 and the number of non-buyers is 2384.

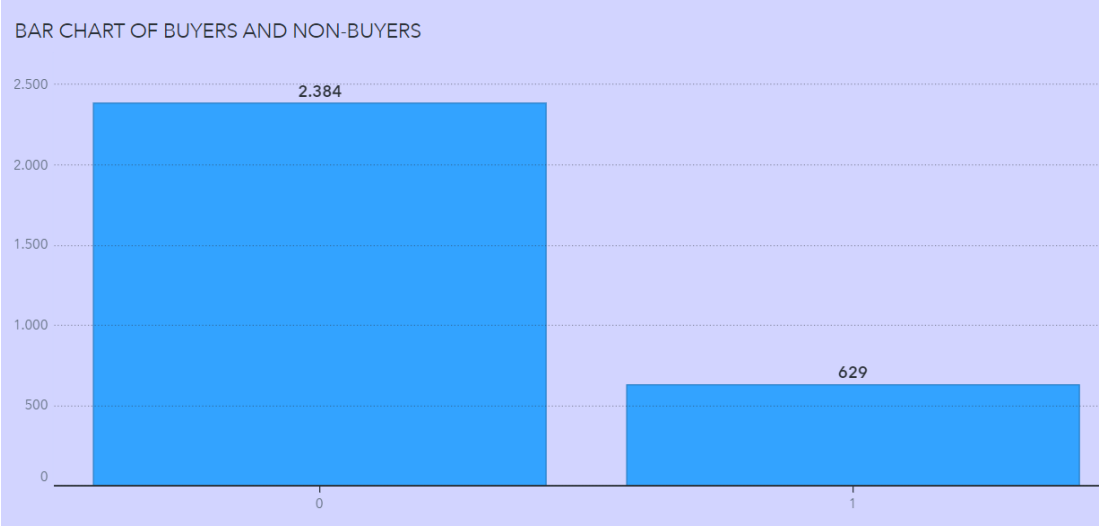


Figure 29: Bar Chart for Buyers-Non-Buyers

19) According to the column Probability for Ins=1 the biggest probability of being a buyer assigned to a customer is 0.7346. On the contrary the smallest one is 0.2122.

20) The software assigns the value 1 or 0 to the customers based on the column named "Probability for Ins=1". It compares the values of this column with a certain threshold which is equal to 0.2727. The customers that have a value greater than this threshold are predicted to be buyers and the ones that have lower as non-buyers. The customer with Cust_ID= 07636 has a probability of 0.2122 which is lower than 0.2727. So the software assigns the value 0 to this customer i.e. he is classified as a non-buyer. Similarly, the customer with Cust_ID =29773 has the same probability of 0.2121 and as a result he is classified as a non-buyer too.