Business Analytics Practicum I

Assignment

Kitsou Effranthia Anastasia

p2822118

Vioni Konstantina

p2822107

Supervisor: Andreas Zaras ,

Data Scientist & Artificial Intelligence Professional

ATHENS, MAY 2023

Case Study 1

Executive Summary

This report aims to identify the potential sales prospects of an online store, Buy-books-on-line.com, which specializes in selling books that cover topics related to science and information technology. The category of books that we were interested in was that related to “Business Analytics”. A very popular method that belongs to Pattern Discovery Techniques of Data Mining/Machine Learning, the Market Basket Analysis, also known as Association Rule Discovery or Affinity Analysis, was used to manage this problem. The Market Basket Analysis aims to determine the likelihood of purchasing a product or group of products, given that a customer has already bought or searched for a product or group of products. This method was utilized in the sales history of 19.805 customers who have previously purchased at least one book from the 'Business Analytics' category. Before conducting the analysis, we generated a bar chart using SAS Visual Analytics software to show the number of book sales by title, which contributed to identifying the titles with the highest and lowest sales. The Market Basket Analysis consisted of two parts. The first part focused on five specific books ('Managerial Analytics', 'Implementing Analytics', 'Customer Analytics for Dummies', 'Enterprise Analytics') and on identifying their relationships. For each book, two books were identified that should be advertised together. For “Managerial Analytics”, the two books were “Implementing Analytics” and “Web Analytics 2.0”. For “Implementing Analytics”, the two books were “Data Science and Big Data Analytics'' and “Managerial Analytics” while for “Customer Analytics for Dummies”, the two books were “Decision Analytics” and “Enterprise Analytics”. Last but not least, for “Enterprise Analytics”, the two books were “Customer Analytics for Dummies” and “Managerial Analytics”.The second part of the analysis identified the three books that were most frequently bought together. These books were “Data Science and Business Analytics”, “Business Analytics for Managers” and “Data Analytics Made Accessible”.

2)In the following bar chart is presented the number of units that have been sold for each book. As can be observed, the 'Data Science and Business Analytics' book has the highest sales (1,600 units) while the 'Managerial Analytics' has the lowest sales (152 units).

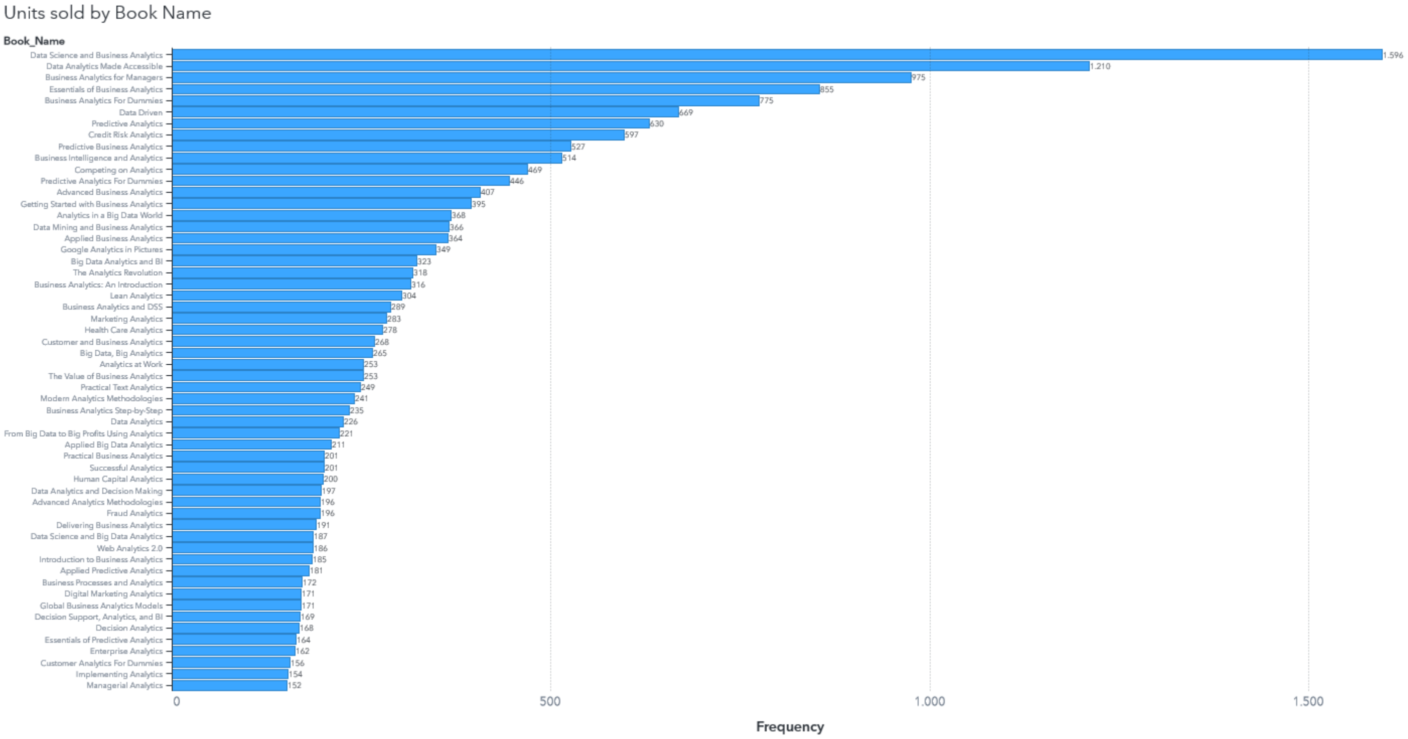


Figure 1: Sales for each book.

3)The store is looking for the best two books to advertise to customers who have purchased or are searching for four specific books. These books are Managerial Analytics, Implementing Analytics, Customer Analytics for Dummies, and Enterprise Analytics. To determine which books to advertise, we sorted the lift column in descending order and filtered the Market Basket Analysis table for association rules that included only the book of interest on the left side. This allowed us to identify the books that the store should promote for each case. Taking into consideration the book "Managerial Analytics", then the "Implementing Analytics" and "Web Analytics 2.0" books should be advertised due to the highest lift 11.4. Moreover, If a customer is interested in the book "Implementing Analytics" then with Lift 11.3 "Data Science and Big Data Analytics" and "Managerial Analytics" should be advertised. If a customer is interested in the book "Customer Analytics for Dummies", then with Lift 11.19, "Decision Analytics" and "Enterprise Analytics" should be advertised together. Also, If a customer is interested in the book "Enterprise Analytics" then with Lift 11.07 "Customer Analytics for Dummies" and "Managerial Analytics" should be advertised.

The association rule that had the highest lift among the three-item rules involving the books mentioned is the one where the 'Managerial Analytics' book is on the left, and the 'Implementing Analytics' and 'Web Analytics 2.0' books are on the right. The lift 11.4 implies that if a customer has purchased the 'Managerial Analytics' book, then it is 11.4 times more likely to buy the 'Implementing Analytics' and 'Web Analytics 2.0' books, compared to a random customer who has not purchased the 'Managerial Analytics' book.

4)If we specify the maximum number of items in a rule to be 3, then the three books that are most frequently bought together are “Data Science and Business Analytics”, “Business Analytics for Managers” and “Data Analytics Made Accessible”. This book collection was found together 794 times, indicating that the customers have bought these books together 794 times. The support metric for this rule is 41.877 and is calculated as the probability of intersections of those three books. This number depicts the number of times that these books have been bought together in all the transactions of our dataset.

Case Study 2

Executive Summary

Sports-OnLine.com is an online retailer that sells sports clothes and shoes. Following a discussion with the marketing team, it was determined that examining customer segmentation would be beneficial to better understand the market. After evaluating the accessible information, it was concluded that implementing a Recency Frequency Monetary (RFM) analysis would be the most appropriate approach to achieve the desired outcome. The process of RFM analysis involves grouping customers into sets that have comparable attributes, determined by three main elements. These elements are Recency, Frequency, and Monetary. Recency is related to how recently a customer has made a purchase, while Frequency refers to how frequently a customer has interacted with the retailer. Monetary relates to the amount of money that a customer has spent on the retailer's goods. The result of the RFM analysis was the creation of four groups of customers. The first group was called "Worst Customers" referring to the bad customers that, on average, compared to the average customer, they had neither purchased a product recently nor frequently and the total amount of money they had spent was lower than that of the average customer. The second group, named "Best Customers", indicated the customers that, on average, compared to the average customer, they had purchased products more recently and frequently while the total amount of money they had spent was higher. The third group is referred to as the "Churners". This group consisted of customers that, on average, compared to the average customer, had not purchased an item for a long time but the frequency and the total amount of money they had spent was higher. The fourth group, named "First Timers" represented customers that, on average, compared to the average customer, had purchased an item more recently, less frequently, and with a lower total amount of money spent. Finally, for the aforementioned groups of customers have been suggested some marketing actions, which encompass enhancing services based on customer feedback, providing customized deals, and introducing loyalty products.

Exploratory Data Analysis

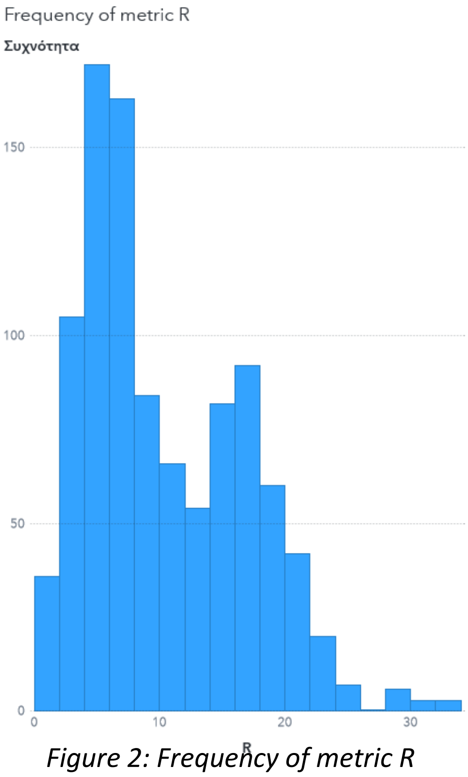
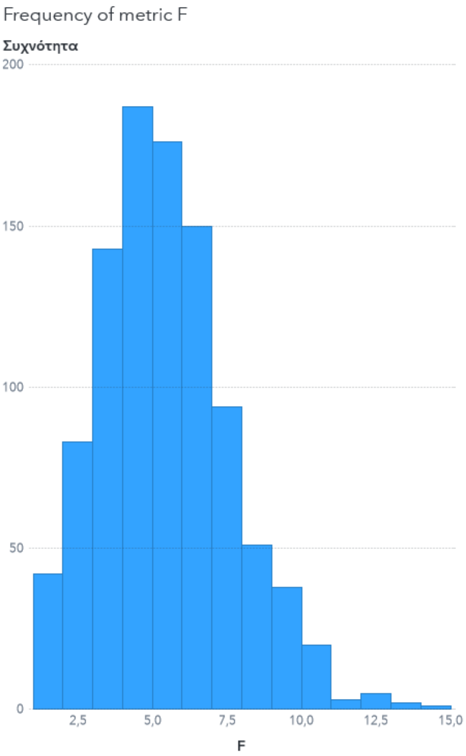


Figure 2: Frequency of metric R



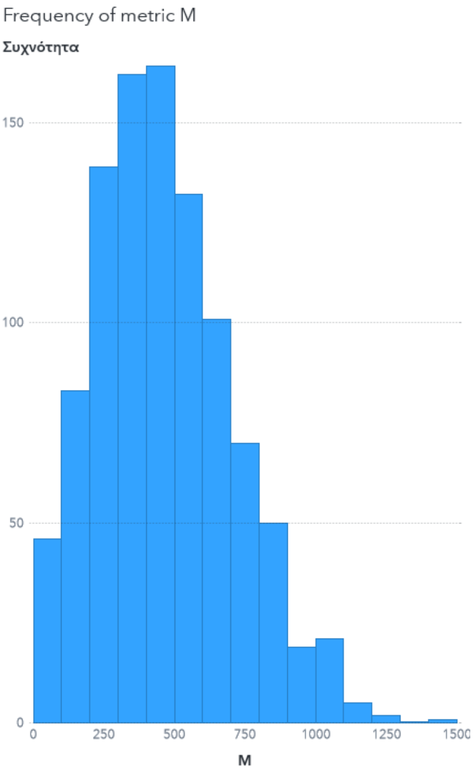


Figure 3: Frequency of metric F

Figure 4: Frequency of metric M

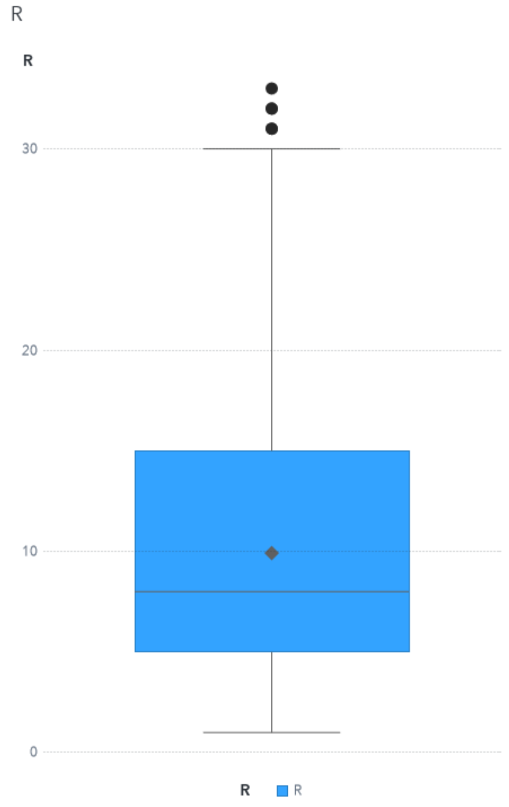
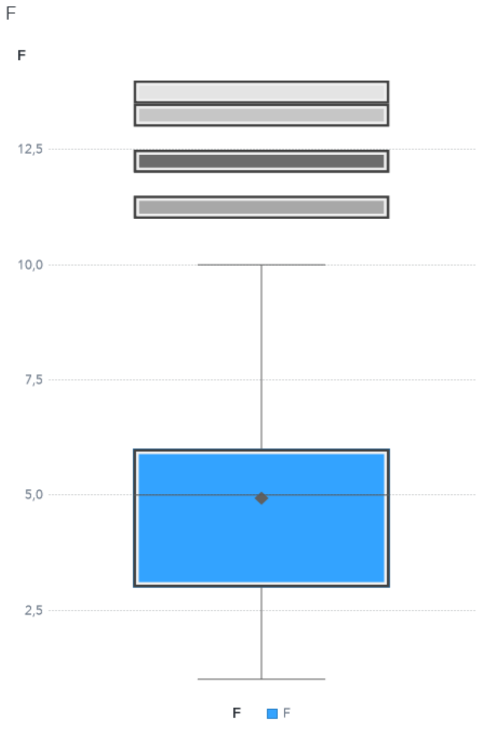


Figure 5: Boxplot for metric R



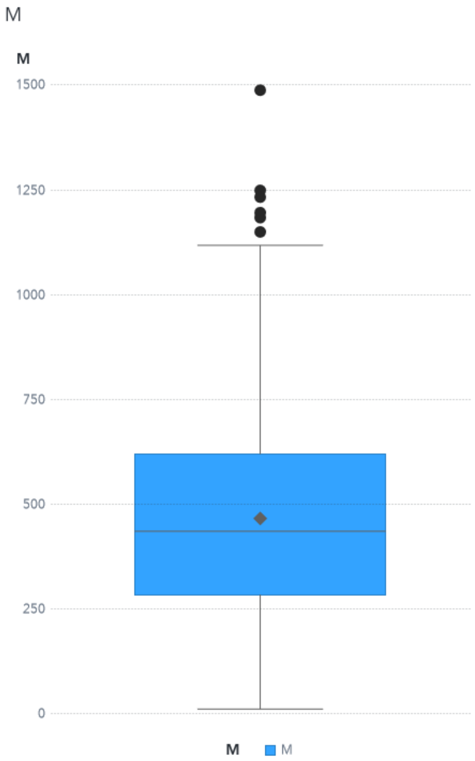
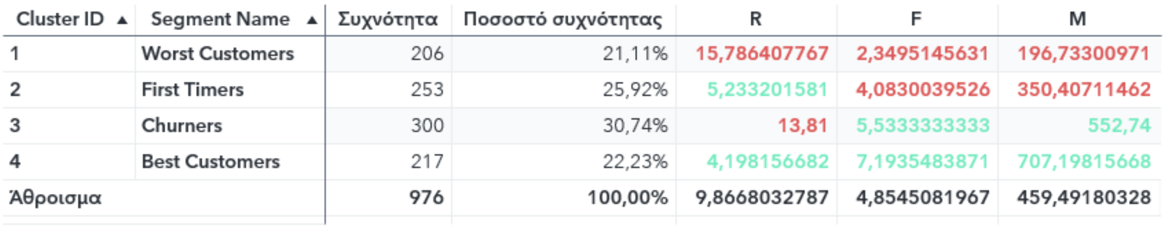


Figure 6: Boxplot for metric F

Figure 7: Boxplot for metric M



Cluster ID ▲	Segment Name ▲	Συχνότητα	Ποσοστό συχνότητας	R	F	M
1	Worst Customers	206	21,11%	15,786407767	2,3495145631	196,73300971
2	First Timers	253	25,92%	5,233201581	4,0830039526	350,40711462
3	Churners	300	30,74%	13,81	5,5333333333	552,74
4	Best Customers	217	22,23%	4,198156682	7,1935483871	707,19815668
Άθροισμα		976	100,00%	9,8668032787	4,8545081967	459,49180328

Figure 8: RFM Analysis

For the implementation of the RFM analysis, we used a dataset that was produced by the IT department and the Business Analytics department. This dataset was in an RFM format and contained information regarding customer id, recency, frequency, and monetary for each customer. Firstly, we conducted an exploratory data analysis of the dataset and we observed that there were some outliers. Also, we noticed that some variables had asymmetrical distribution. To solve this problem, we used the logarithm to transform these values. After we completed this procedure, we created the final dataset cleaned from outliers, and proceeded to the RFM analysis using k-means and Euclidean distance.

With the fulfillment of the RFM analysis, our customers had been clustered into 4 clusters. The process that we followed consisted of customer clustering and segmentation leading to grouping customers according to their behavior. We should mention that they were grouped with the logic of maximum similarity between the customers in the same group while also maximum dissimilarity with all the customers outside the group. These generated clusters were named according to customers' behavior. It can be easily understood that an average customer in general has almost 10 months to complete a purchase, has completed about 5 purchases, and on average spends almost 460€.

More specifically, for the segment of "Worst Customers" (group 1), it can be seen that these customers had almost 15 months to complete a purchase, had completed about 3 purchases and they also had spent on average almost 197€. Compared to the average customer, it can be concluded that they had neither purchased a product recently nor frequently and the total amount of money they had spent was lower than that of the average customer.

The cluster of "First Timers" (group 2) had almost 5 months to complete a purchase, had completed about 4 purchases, and had spent on average 350€. Compared to the average customer, this group had purchased an item more recently, less frequently, and with a lower total amount of money spent.

For the segment of "Churners" (group 3), we observed that they had almost 13 months to complete a purchase, had completed about 5 purchases, and had also spent on average 552€. Compared to the average customer, this group had not purchased an item for a long time but the frequency and the total amount of money they had spent was higher than that of the average customer.

Finally, for the segment of "Best Customers" (group 4), we examined that they had almost 4 months to complete a purchase, had completed about 7 purchases and had spent on average 707€. Compared to the average customer, they had purchased products more recently and frequently while the total amount of money they had spent was higher than that of the average customer.

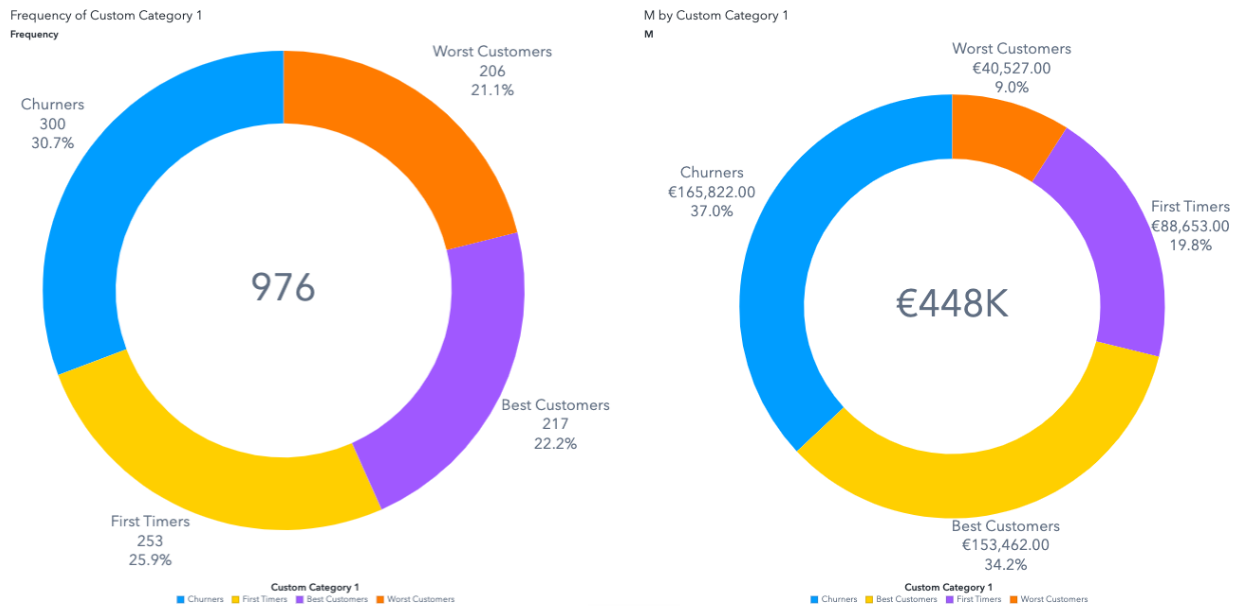


Figure 9: RFM Cluster Analysis Graph

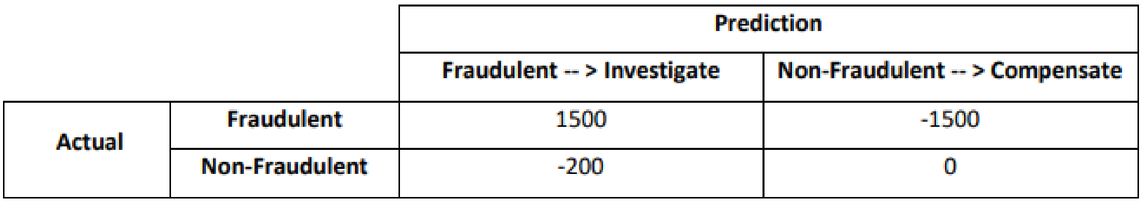
Based on the RFM analysis results, some actions could be implemented for each group of customers to accomplish the desired outcome. Firstly, considering the "Worst" group of customers which refers to the ones who stopped purchasing, we could send them a questionnaire requesting their feedback. By this, we could understand the reasons that led them to reduce their engagement with the company and offer them incentives to re-engage. For the "Best" group, we could create some loyalty programs that will reward them for their continued engagement with the company while maintaining their satisfaction. Regarding the "Churners", the customers that have discontinued their engagement with the company, we could propose a reactivation program that aims to regain them by providing new offers and personalized emails and advertisements with products that meet their needs. Finally, for the "First timers", we could also promote some attractive offers to maintain their engagement and keep them as satisfied as possible.

Case Study 3

Executive Summary

An insurance organization -XYZ- operating in the motor insurance industry for the past seven years is interested in achieving fewer losses and more profits while implementing better customer service. To accomplish this goal it was required to develop a mathematical model that will predict whether the claim will prove to be fraudulent or not. The fraud prevention department of the company, in cooperation with the IT department, has collected the dataset named “Historical_Claims_Final”, which contains data about claims from the period 1st of May 2017 to 30th of August 2017. This dataset consisted of claims’ characteristics (input variables) such as the age of the vehicle, whether a witness was present during the accident, the number of days elapsing between the accident day and the policy termination date, or even more the area where the accident took place. Firstly, we conducted an Exploratory data analysis that helped us understand the data. We identified that the proportion of fraudulent claims in our dataset was 30% while the proportion of non-fraudulent claims was 70%. Secondly, we generated four models. These were the Decision Tree, the Maximal Tree, the Logistic Regression, and the Neural Network. The best model was the Decision Tree due to some metrics that were considered. For validation, the best model was applied to the New claims dataset (issued after the 1st of October) to predict whether these claims are fraudulent or not. From the total 200 claims, the model classified 148 claims as non-fraudulent and the rest 52 as fraudulent.

2)



		Prediction	
		Fraudulent -- > Investigate	Non-Fraudulent -- > Compensate
Actual	Fraudulent	1500	-1500
	Non-Fraudulent	-200	0

Figure 10:Profit Matrix

This is the profit matrix generated by the management team of the fraud prevention department. This matrix takes into consideration all the possible outcomes based on the case that a claim could be fraudulent or not. The numbers represent monetary units in euros and they are calculated by the difference between the cost and the income of each potential scenario. More specifically, in the first case, if a claim is fraudulent and is also predicted as fraudulent, the fraud prevention department can proceed to further investigation to be sure that it is a legitimate claim and gain 1500 euros. In the second case, if a claim is fraudulent and is predicted as non-fraudulent, then the department proceeds to the compensation, losing 1500 euros. Moreover, in the case of non-fraudulent claims and fraudulent prediction, the department proceeds to a non-needed investigation and loses 200 euros. Finally, if a claim is non-fraudulent and is also predicted as non-fraudulent, then the company does not make either a loss or a profit.

3) The minimum probability (cut-off point) that a claim should have to be considered as fraudulent can be calculated after specifying the expected profit if the claim is fraudulent and the expected profit if the claim is non-fraudulent. By solving the system of the two , the probability of a claim to be fraudulent (p_1) can be estimated leading to the calculation of the cut-off point. The calculations that were implemented are presented below.

$$\textit{Expected profit}_{\textit{if the claim is fraudulent}} = p_1 \times 1500 + (1- p_1) \times (-200) = 1700\, p_1 – 200$$

$$\textit{Expected profit}_{\textit{if the claim is non-fraudulent}} = p_1 \times (-1500) + (1- p_1) \times 0 = -1500\, p_1 \quad ,$$

$$\textit{where} \quad \textit{Expected profit}_{\textit{if the claim is fraudulent}} = \textit{Expected profit}_{\textit{if the claim is non-fraudulent}}$$

$$=> \quad 1700\, p_1 – 200 \quad > \quad -1500\, p_1$$

$$=> \quad 1700\, p_1 + 1500\, p_1 > 200$$

$$\Rightarrow \quad 3200\, p_1 > 200$$

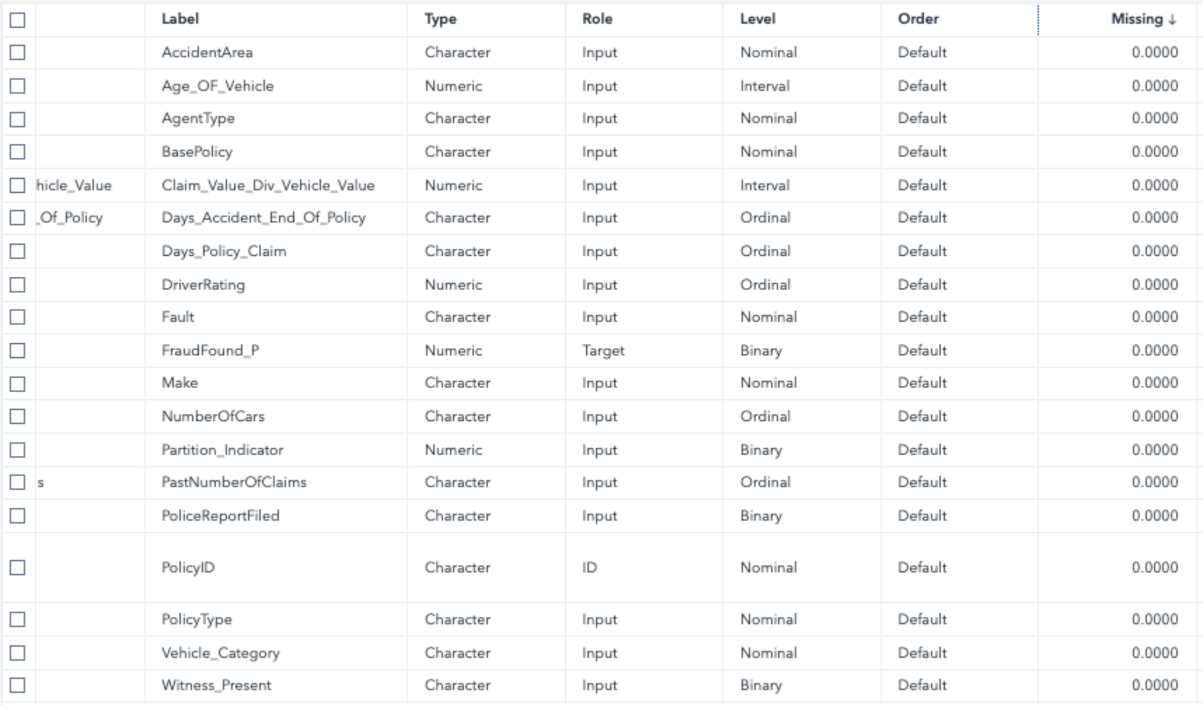
$$\Rightarrow \quad p_1 > \frac{200}{3200}$$

$$\Rightarrow \quad p_1 > 0.0625$$

As a result, the minimum probability that a claim should have to be considered as fraudulent is 0.0625 or 6.25%.

4) The historical data must be split to create two new datasets, the training, and the validation. The training dataset consists of 70% of the total observations of the initial dataset and it will be used for training the model. The validation dataset contains 30% of the total observations of the initial dataset and it will be useful for testing the model. This practice of splitting the dataset allows the models to be internally optimized using the validation testing after their initial training with the first dataset. It also prevents issues such as overfitting, where a model is trained too well on a particular dataset and begins to perform poorly on new unseen data. When the data are being stratified, each subgroup has equal representation in our dataset leading to more robust models as the model is less biased.

5) It can be easily observed that the dataset has no missing data as the respective field (column "Missing") is filled with zeros on every variable of the Dataset.



□		Label	Type	Role	Level	Order	Missing ↓
□		AccidentArea	Character	Input	Nominal	Default	0.0000
□		Age_OF_Vehicle	Numeric	Input	Interval	Default	0.0000
□		AgentType	Character	Input	Nominal	Default	0.0000
□		BasePolicy	Character	Input	Nominal	Default	0.0000
□	hicle_Value	Claim_Value_Div_Vehicle_Value	Numeric	Input	Interval	Default	0.0000
□	_Of_Policy	Days_Accident_End_Of_Policy	Character	Input	Ordinal	Default	0.0000
□		Days_Policy_Claim	Character	Input	Ordinal	Default	0.0000
□		DriverRating	Numeric	Input	Ordinal	Default	0.0000
□		Fault	Character	Input	Nominal	Default	0.0000
□		FraudFound_P	Numeric	Target	Binary	Default	0.0000
□		Make	Character	Input	Nominal	Default	0.0000
□		NumberOfCars	Character	Input	Ordinal	Default	0.0000
□		Partition_Indicator	Numeric	Input	Binary	Default	0.0000
□	s	PastNumberOfClaims	Character	Input	Ordinal	Default	0.0000
□		PoliceReportFiled	Character	Input	Binary	Default	0.0000
□		PolicyID	Character	ID	Nominal	Default	0.0000
□		PolicyType	Character	Input	Nominal	Default	0.0000
□		Vehicle_Category	Character	Input	Nominal	Default	0.0000
□		Witness_Present	Character	Input	Binary	Default	0.0000

Figure 11:Missing values

Also, we noticed that the proportion of fraudulent claims in our dataset was 30% while the proportion of non- fraudulent claims was 70%.

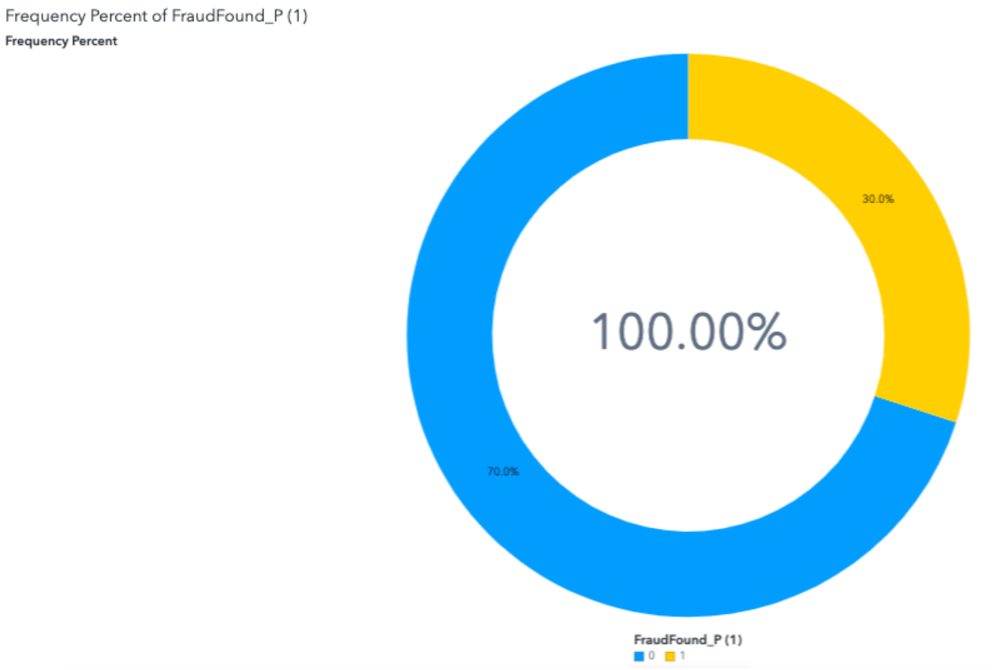


Figure 12: Proportion of fraudulent and non-fraudulent claims.

6) The proportion of fraudulent and non-fraudulent claims in the historical dataset is 30% - 70%. If the proportions of the dataset were 10% fraudulent and 90% non-fraudulent, the dataset would be imbalanced. This type of dataset can be difficult to handle in predictive modeling, since many classification algorithms were designed with the assumption of an equal distribution of examples across each class. As a result, the model will perform poorly, especially when it comes to predicting outcomes for the minority class. To address this issue, it may be necessary to rebalance the dataset by adjusting the proportions of each category of claim. Since the value of interest (fraudulent) is 10%, we could use under-sampling. This means that we will select all the observations from the minority class (fraudulent) and a subset of the observations from the majority class (non-fraudulent) to create a new dataset that contains 30% fraudulent claims and 70% of non-fraudulent claims.

7) Filtering the claims to keep only them that have Claim Value Divided by the Vehicle Value greater than 120% we observed that the proportion of fraudulent and non-fraudulent claims changed to 47.9% fraudulent and 52.1% non-fraudulent.

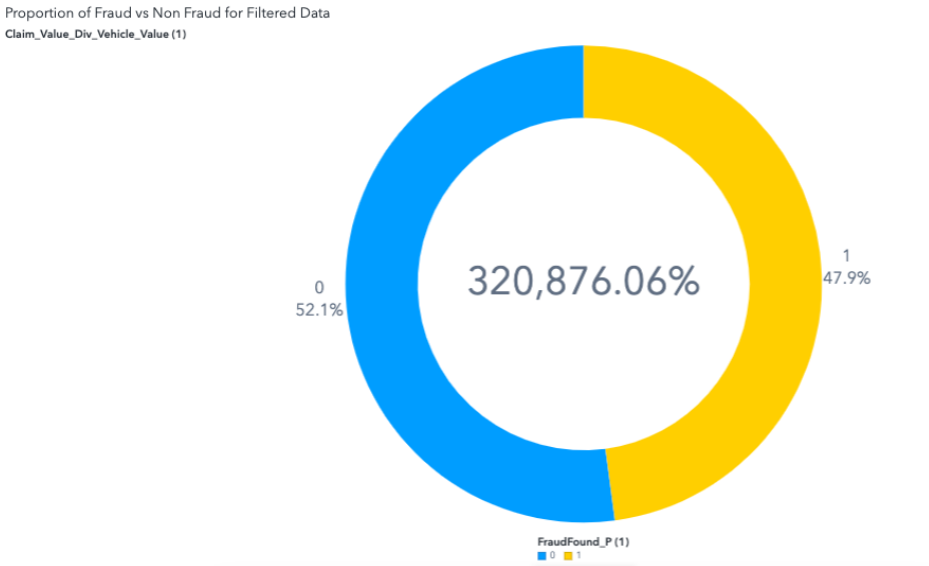


Figure 13:The proportion of fraudulent and non-fraudulent claims with Claim Value Divided by the Vehicle Value greater than 120%.

8) The average AgeOfVehicle differs significantly between fraudulent and non-fraudulent claims. We concluded that the average age of the vehicle for the non - fraud claims is smaller, about 7.5 years while the average age of the vehicle for fraud cases is more than 10 years. This also means that the age of the vehicle in our claims could be a crucial factor for a good fraud detection model.

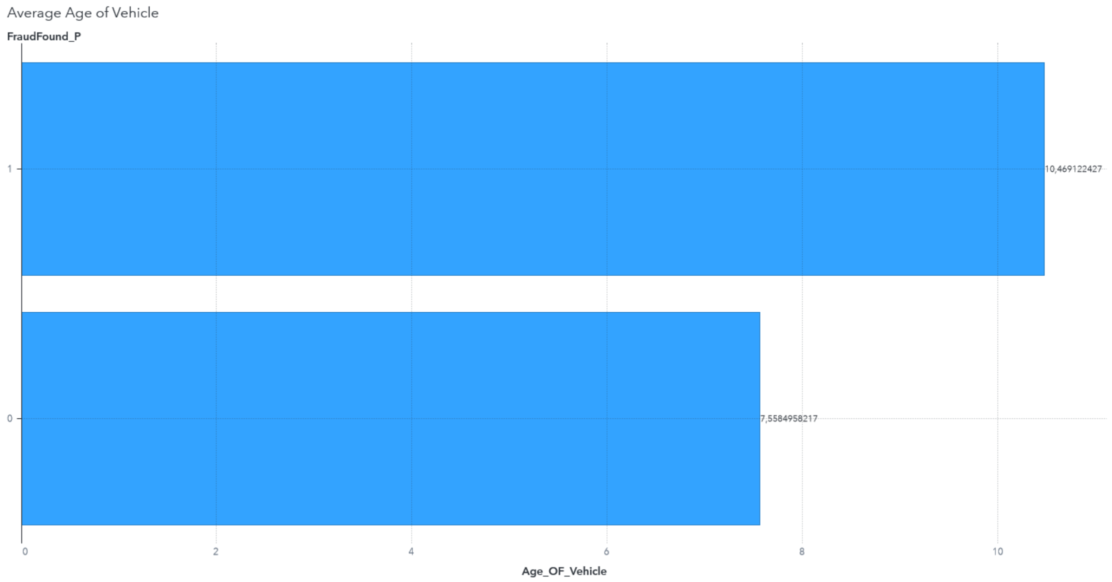


Figure 14: The average Age of Vehicle for fraudulent and non-fraudulent claims.

9) We created the Decision Tree node and connected it with the data source node. In our generated tree, the first variable used was the Age of the Vehicle as we had already guessed. This variable was chosen because it had the highest quality split. The quality of the split was calculated by the logworth. So, for each split, the variable that has been chosen is the one with the highest observed logworth. In general, customers' claims that the Age of the vehicle is greater than or equal to 8 years are directed to the left node while in other cases are directed to the right node. The left part of the tree is associated with fraudulent cases while the right part of the tree is generally associated mostly with non-fraudulent cases. Missing data in the tree are being associated with the right part of the tree, the non-fraudulent claims.

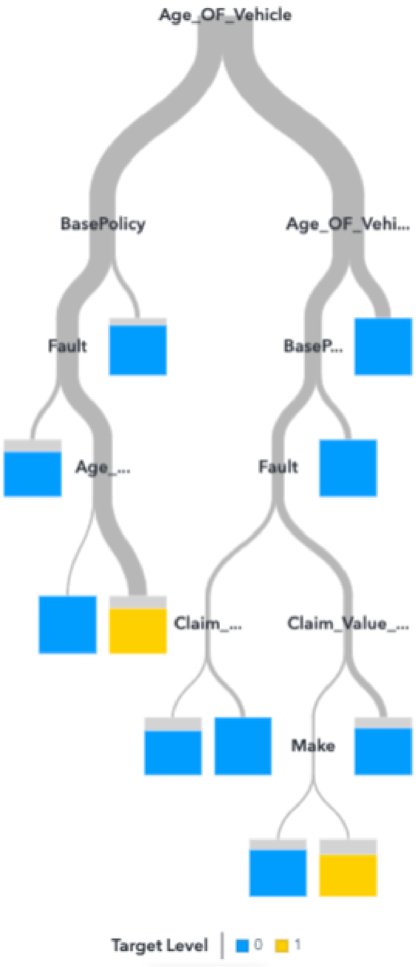
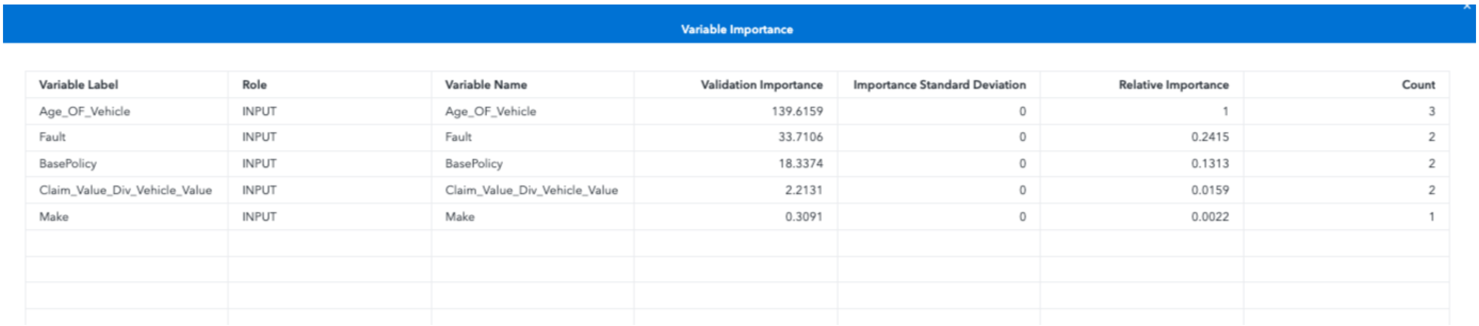


Figure 15:The Decision Tree

This variable has been selected as it has the most validation importance in our model (136,6).



Variable Importance

Variable Label	Role	Variable Name	Validation Importance	Importance Standard Deviation	Relative Importance	Count
Age_OF_Vehicle	INPUT	Age_OF_Vehicle	139.6159	0	1	3
Fault	INPUT	Fault	33.7106	0	0.2415	2
BasePolicy	INPUT	BasePolicy	18.3374	0	0.1313	2
Claim_Value_Div_Vehicle_Value	INPUT	Claim_Value_Div_Vehicle_Value	2.2131	0	0.0159	2
Make	INPUT	Make	0.3091	0	0.0022	1

Figure 16:Variable Importance

10) The maximal tree, which is the tree with the highest number of trees,in our case has 31 terminal leaves.

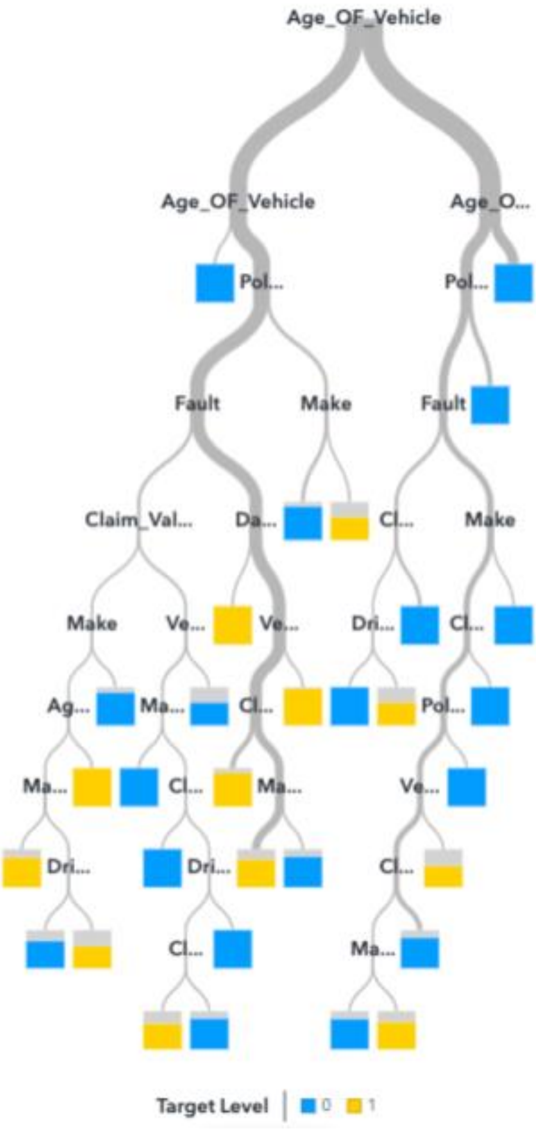


Figure 17: Maximal Tree

The plot in Figure 13 presents how the performance of the model changes as the tree size increases. The x-axis represents the tree size while the y-axis represents the error rate of the model. This plot describes how the training and validation error change as the tree size increases. The line representing the training dataset decreases as the tree size increases. This is because a larger tree can capture more complex relationships in the data and fit the training set more closely leading to overfitting.

Overfitting is noticed when the built model is fitted so close to the actual data becoming so complex that reflecting each validation in the initial dataset, there is almost no misclassification error. This makes the model perform poorly on new data as it is not robust. A simple and easy solution to this is Pruning. During Pruning nonessential and non useful branches or nodes from the tree are erased, leading to more parsimonious models.

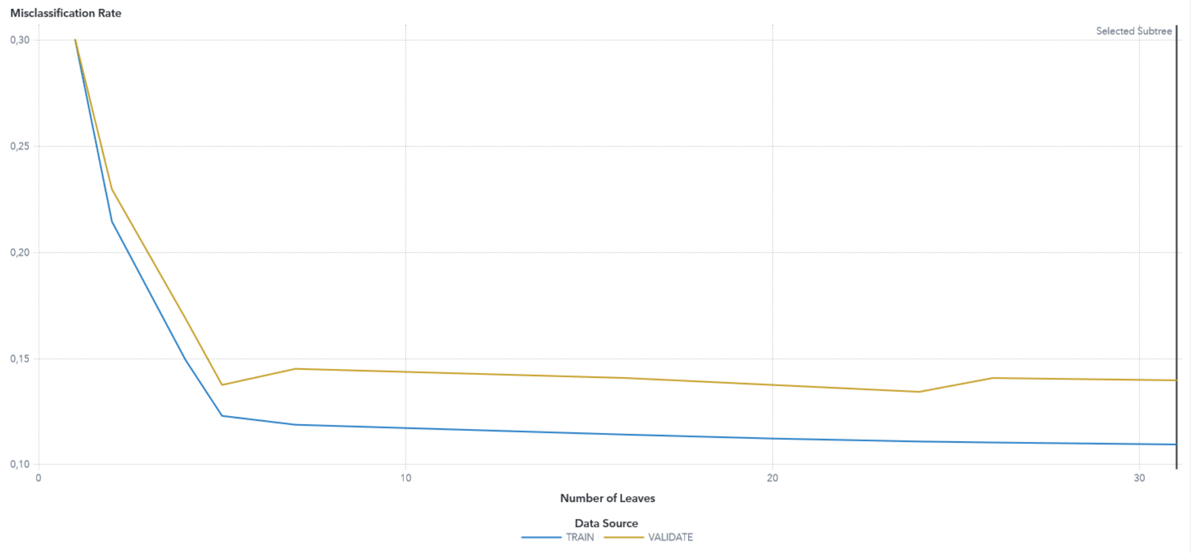


Figure 18: Misclassification Rate of Maximal Tree

11) The optimal tree has 11 terminal leaves. It is also observed that the model behind this tree is very robust leading to a Misclassification error around 0.10 for the train data and less than 0.15 for the validation data. That means that the model is being trained well and can perform great on unseen data. Moreover, the optimal tree in the graph presented in Figure 15 is selected when the Misclassification Rate takes the lowest value. After this point, the Misclassification Rate on the validation set starts increasing again while the training is still relatively low. This leads to a larger tree with higher errors and consequently overfitting.

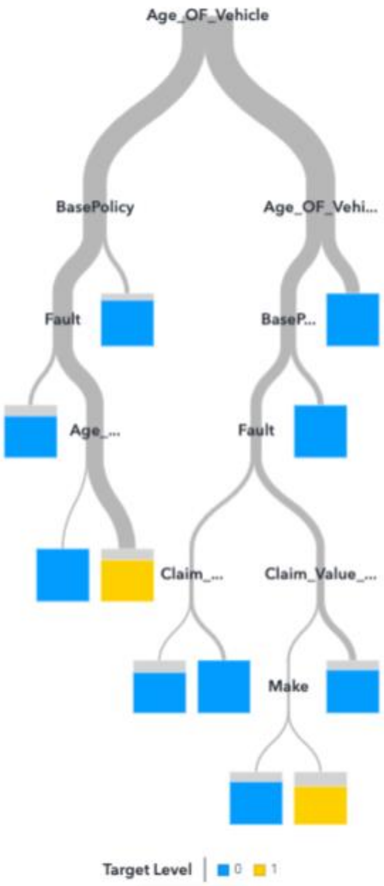


Figure 19: Optimal Tree

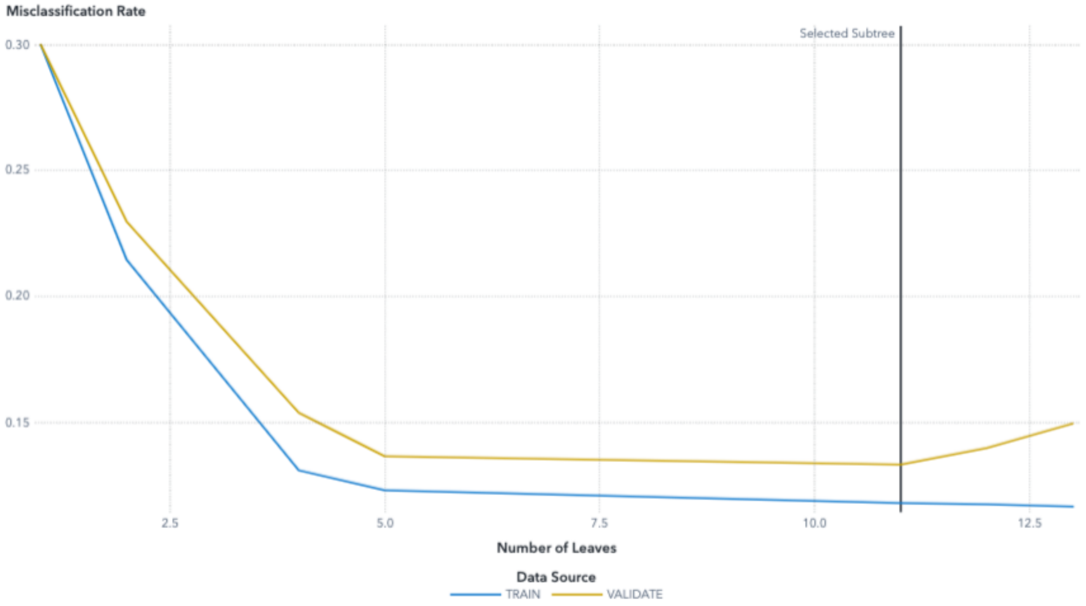


Figure 20: Misclassification Rate of Optimal Tree

12) A decision tree model is characterized by three important characteristics. These are the posterior probability (the probability of the observations in a bracket belonging to each category), the final decision (the category to which the observations are classified), and the rules (the thresholds that each node uses to split the data until it reaches the terminal leaf).In this part, we will interpret the decision tree model by using only 5 of the terminal leaves where p_1 > 6.25%.

Taking into consideration the terminal leaf with node id 7, the posterior probabilities are po:78.91 % and p1:21.09%. We should note that po is the probability of a claim to be non-fraudulent and p1 is the probability of a claim to be fraudulent. The decision at this node is 0 which means that the observations are classified in the non-fraudulent category. The rules are Age of vehicle >=8, Base policy= Collision, All Perlis, and Fault = Third party.

For the terminal leaf node id 12, the posterior probabilities are po: 21.69% and p1:78.31%. Moreover, the decision at this node is 1 which means that the observations are classified to the fraudulent category. Moreover, the rules are Age of vehicle >=8, Base policy = Collision, All Perlis, Fault = Policy Holder, Age of vehicle 0<= and <15.

Regarding the terminal node id 15, the posterior probabilities are p_o:76.92 % and p_1:23.08%. The decision at this node is 0 and the observations are classified to the non-fraudulent category. The rules are Age of vehicle <8 or Missing, Age of vehicle >=7, Base policy= Collision, All Perlis, Fault = Third party, and Claim value divided by the vehicle value >= 1.826507.

In the terminal node id 18, the posterior probabilities are p_o:81.48% and p_1:18.52%. The decision at this node is 0 and the observations are classified to the non-fraudulent category. The rules are Age of vehicle <8 or Missing, Age of vehicle >=7, Base policy= Collision, All Perlis, Fault = Policy Holder, and Claim value divided by the vehicle value < 1.483497 or Missing.

Furthermore, in the terminal node id 20, the posterior probabilities are po:28% and p1:72%. The decision at this node is 1 and the observations are classified to the fraudulent category. The rules are Age of vehicle <8 or Missing, Age of vehicle >=7, Base policy= Collision, All Perlis, Fault = Policy Holder, Claim value divided by the vehicle value >=1.483497, and Make = Toyota, Accura , Honda.

13) This part will interpret the decision tree in a non-technical manner. More specifically, it can be noted that if a claim has an Age of Vehicle greater than or equal to 8 years, and if the Base policy is equal to Collision - All Perils, while the Fault is a third party, then the probability that a claim will be fraudulent is 21,09% and the probability of that claim to be non-fraudulent is 78.91 %. Therefore following the defined threshold, the customer’s claim will be non-fraudulent (node id 7).

If the vehicle's age is greater than or equal to 8 years old, and if the Base policy of collusion is Collision-All Perils, while the Fault lies with the policy holder, and the age of Vehicle is less than 15 years, then the probability that a claim will be fraudulent is 78,31% and 21.69% for non-fraudulent. Therefore, in accordance with the defined threshold, the customer's claim will be fraudulent (node id 12).

If a claim has an Age of Vehicle less than 8 years or the age is unknown but the age is greater than 7 years, and if the Base policy is Collision-All Perils, while the Fault is third party and the claim value divided by the Vehicle value is greater or equal to 1.8, then the probability that a claim will be fraudulent is 23,8% and 76.92% for non-fraudulent. Therefore, following the defined threshold, the customer's claim will be non-fraudulent (node id 15).

If a claim has an Age of Vehicle less than 8 years or the age is unknown, and if the age is greater than 7 years, and the Base policy is Collision-All Perils, while the Fault lies by the policyholder and the claim value divided by the Vehicle value is less than 1.4 or this value is unknown, then the probability that a claim will be fraudulent is 18,52% and 81.48% for non-fraudulent. Therefore, following the defined threshold, the customer's claim will be non-fraudulent (node id 18).

If a claim has an Age of Vehicle less than 8 years or the age is unknown, and the age is greater than 7 years, and if the Base policy is Collision-All Perils, while the fault lies by the policyholder and the claim value divided by the Vehicle value is greater or equal to 1.4 or this value is unknown and if the manufacturer is Toyota, Accura or Honda, then the probability that a claim will be fraudulent is 72% and 28% for non-fraudulent. Therefore, following the defined threshold, the customer's claim will be fraudulent (node id 20).

14) In the following graph, it is presented the cumulative % response for the five created models. As it can be observed, if we check the 20% of the most highly ranked customers according to the probability that the optimal Tree model gives them to be fraudulent, 75.59% of this 20% will be fraudulent. Moreover, if we check the 100% of the most highly ranked customers according to the probability that the optimal Tree model gives them to be fraudulent, then 30% of this 100% will be fraudulent.

It can also be understood that if we check the 20% of the most highly ranked customers according to the probability that the Logistic Regression model gives them to be fraudulent, 66% of this 20% will be fraudulent. Additionally, if we check the 100% of the most highly ranked customers according to the probability that the Logistic Regression model gives them to be fraudulent, then 30% of this 100% will be fraudulent.

As it can be seen, if we check the 20% of the most highly ranked customers according to the probability that the Maximal Tree model gives them to be fraudulent, 78.39 % of this 20% will be fraudulent. If we check the 100% of the most highly ranked customers according to the probability that the Maximal Tree model gives them to be fraudulent, then the 30% of this 100% will be fraudulent.

If we check the 20% of the most highly ranked customers according to the probability that the Neural Network model gives them to be fraudulent, the 30% of this 20% will be fraudulent. Furthermore, if we check the 100% of the most highly ranked customers according to the probability that the Neural Network model gives them to be fraudulent, then the 30% of this 100% will be fraudulent.

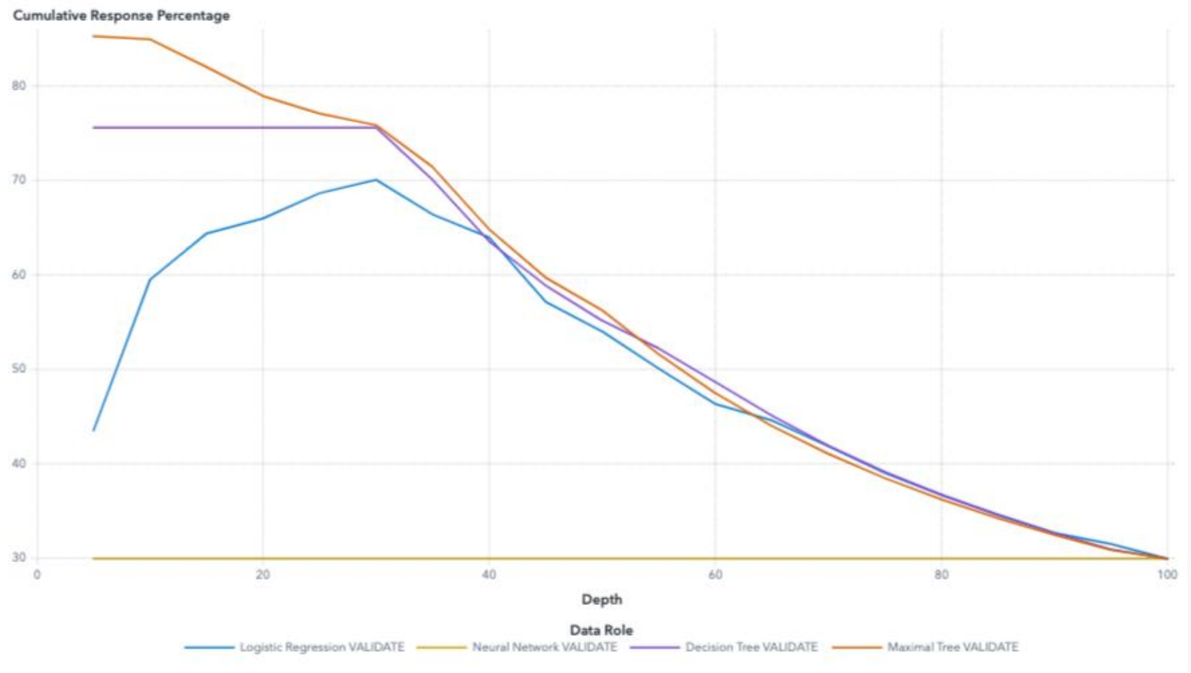


Figure 21: Cumulative Response Percentage

15) In the following graph, the response % of the five created models is presented. The graph is generated by sorting the data based on the likelihood of a customer being fraudulent, and then dividing them into equal-sized buckets. The x-axis shows these buckets. In our analysis, we focused on the fifth bucket which is between 20%-25% of the most highly ranked customers based on the created models.

As it can be observed,if we check the 5th bucket (20%-25%) of the most highly ranked customers according to the probability that the optimal Tree model gives them to be fraudulent, 75,59% of this bucket will be fraudulent.

It can also be understood that if we check the 5th bucket (20%-25%) of the most highly ranked customers according to the probability that the Logistic Regression model gives them to be fraudulent, 70,76% of this bucket will be fraudulent.

Moreover, if we check the 5th bucket (20%-25%) of the most highly ranked customers according to the probability that the Maximal Tree model gives them to be fraudulent, 69,71% of this bucket will be fraudulent.

If we check the 5th bucket (20%-25%) of the most highly ranked customers according to the probability that the Neural Network model gives them to be fraudulent, the 30,01% of this bucket will be fraudulent.

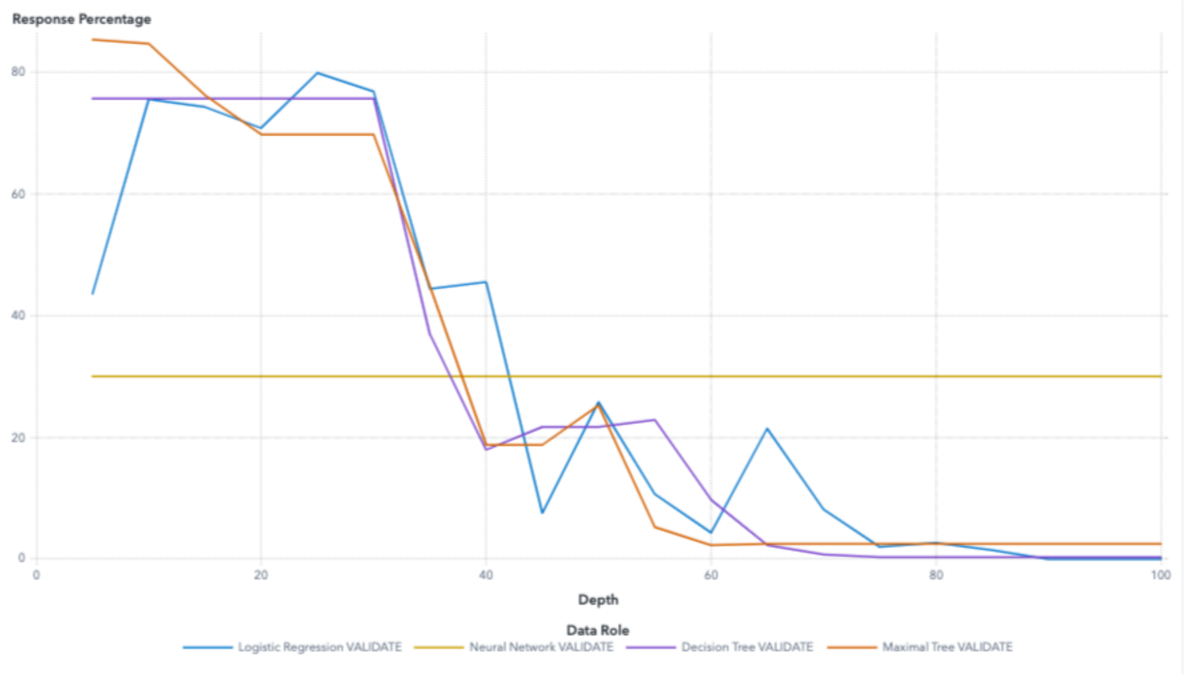


Figure 22: Response percentage

16) In figure 18 it is presented the cumulative lift of the five created models. As it can be understood, if we check the 20% of the most highly ranked customers according to the probability that the optimal Tree model gives them to be fraudulent, we will capture 2.5653 times more fraudulent claims than if we did the same job without a model i.e. at random.

if we check the 20% of the most highly ranked customers according to the probability that the Logistic Regression model gives them to be fraudulent, we will capture 2.2398 times more fraudulent customers than if we did the same job without a model i.e. at random.

Additionally, if we check the 20% of the most highly ranked customers according to the probability that the Maximal Tree model gives them to be fraudulent, we will capture 2.6788 times more fraudulent customers than if we did the same job without a model i.e. at random.

If we check the 20% of the most highly ranked customers according to the probability that the Neural Network model gives them to be fraudulent, we will capture 1.0184 times more fraudulent customers than if we did the same job without a model i.e. at random.

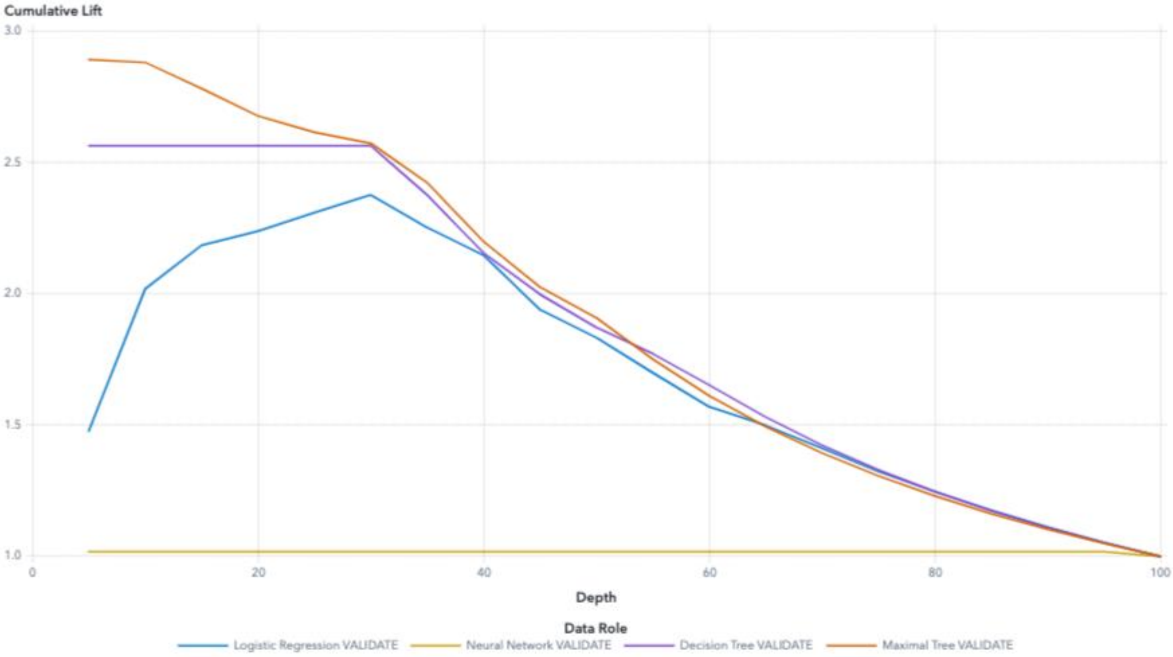


Figure 23:Cumulative lift

17) In the following figure, the cumulative % captured response graph for the five generated models is illustrated. If we take the 40% of the most highly ranked customers according to the probability that the optimal Tree model gives them to be fraudulent, we will capture the 86.2% of all the fraudulent customers of the whole validation dataset.

If we take the 40% of the most highly ranked customers according to the probability that the Logistic Regression model gives them to be fraudulent, we will capture the 85.92% of all the fraudulent customers of the whole validation dataset.

If we take the 40% of the most highly ranked customers according to the probability that the Maximal Tree model gives them to be fraudulent, we will capture the 88% of all the fraudulent customers of the whole validation dataset.

If we take the 40% of the most highly ranked customers according to the probability that the Neural Network model gives them to be fraudulent, we will capture the 40.73% of all the fraudulent customers of the whole validation dataset.

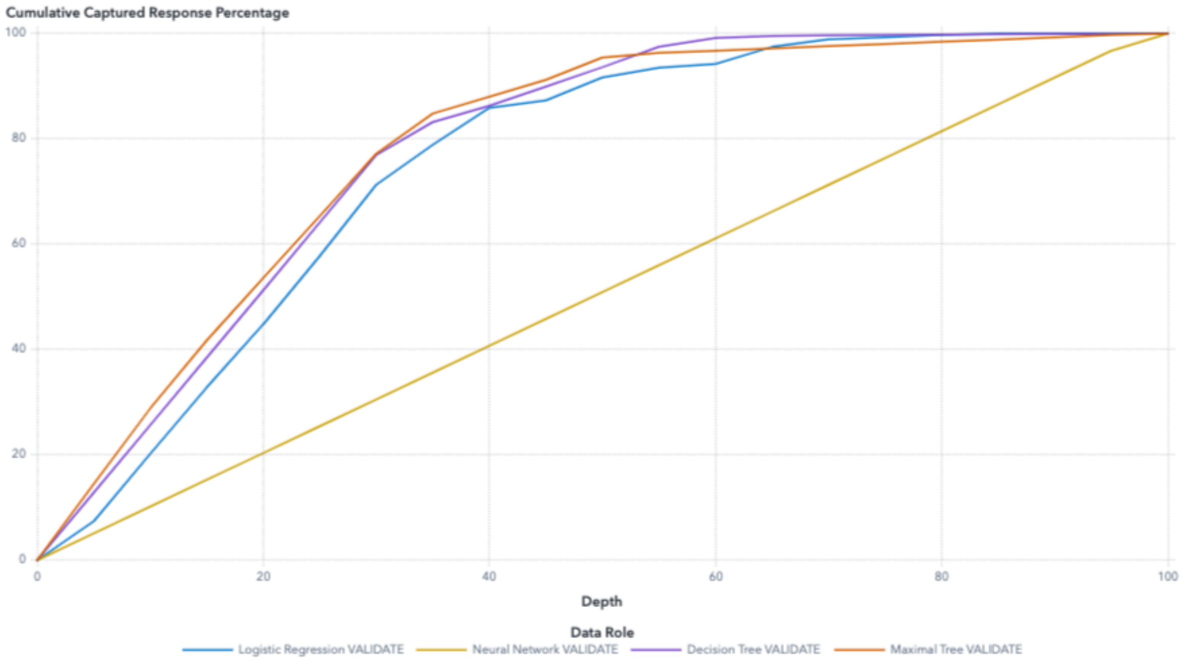


Figure 24:Cumulative Captured Response Percentage

Following in our analysis comes the Scoring of New Data. In the below figure it is presented the whole process.

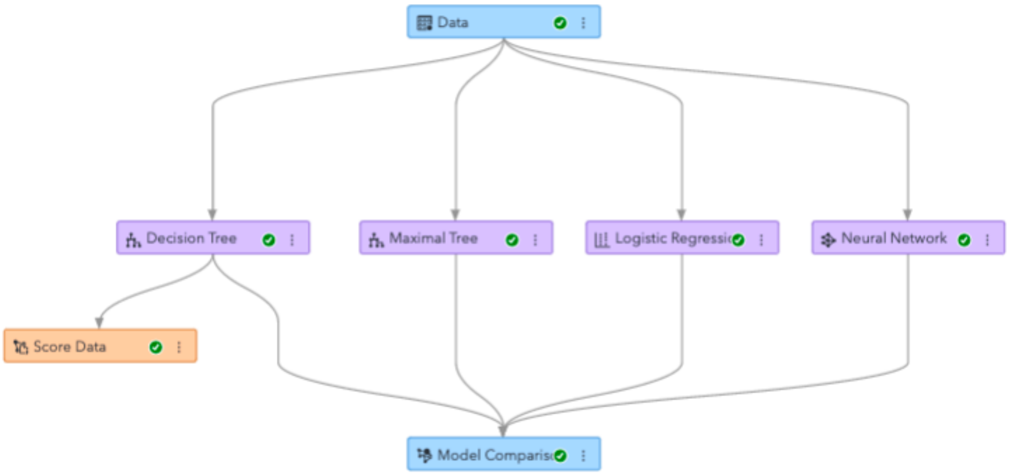


Figure 25: Process

The best model identified by our analysis in accordance to the lowest Average Scored error is the Decision Tree with 0.1.

18) There are 200 claims in the “New_Claims_Final” data. The 148 of them are being classified as non-fraudulent and the rest are being identified as fraudulent.

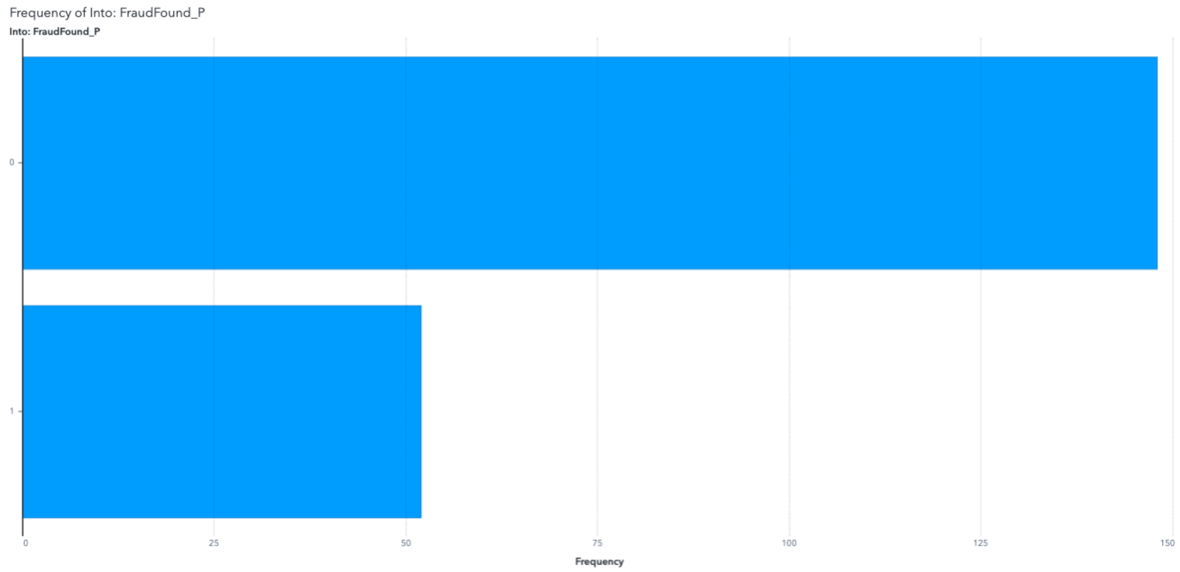


Figure 26: Decision Tree Prediction of fraudulent and non-fraudulent claims.

19) The biggest probability of a claim to be fraudulent is 0.7830609212, this means that it is almost certain that the claim is fraud. The smallest probability is zero meaning that there is no possibility that the specific claim is fraudulent.

20) The claim with PolicyID= 15 is scored as non-fraudulent while the claim with PolicyID=107 is scored as fraudulent. As previously seen, the variable with the biggest importance is the Age of Vehicle. It must not be forgotten also that this variable is used for the first split in our Decision tree based on whether the age of the vehicle is less than 8 years or more than that. In fact the split is proved again in those 2 cases, the claim with PolicyID= 15 regards a vehicle 7 years old and it is scored as non-fraudulent, while the claim with PolicyID= 107 is scored as fraud with a vehicle 10 years old.