

SAS group assignment consisting of three case studies, the first related to market basket analysis (association rules), the second to customer segmentation (clustering) and the third related to campaign management (customer response model)

# **Business Analytics Practicum I**

Assignment 2023

Georgios – Konstantinos Vlassis  
(f2822203)

Christos Vlassis (f2822204)

---

## Table of Contents

Case Study 1.....	2
<b>1) Executive Summary.....</b>	<b>2</b>
<b>2) Book Sales Bar Chart.....</b>	<b>3</b>
<b>3) Association Rules.....</b>	<b>3</b>
<b>4) Top triplet of books.....</b>	<b>4</b>
Case Study 2.....	5
<b>1) Executive Summary.....</b>	<b>5</b>
<b>2) Clustering and Customer Segmentation.....</b>	<b>6</b>
Clusters' graphs.....	8
<b>3) Marketing Actions.....</b>	<b>10</b>
<b>4) Technical Report.....</b>	<b>12</b>
Case Study 3.....	15
1) Executive Summary.....	15
2) Interpretation of the profit matrix.....	16
Question 3.....	16
Question 4.....	17
Question 5.....	18
Question 6.....	19
Question 7.....	20
Question 8.....	20
Question 9.....	21
Question 10.....	22
Question 11.....	24
Question 12.....	24
Question 13.....	26
Question 14.....	29
Question 15.....	30
Question 16.....	31
Question 17.....	32
Choosing the best model and making predictions.....	33
Question 18.....	34
Question 19.....	35

## Case Study 1

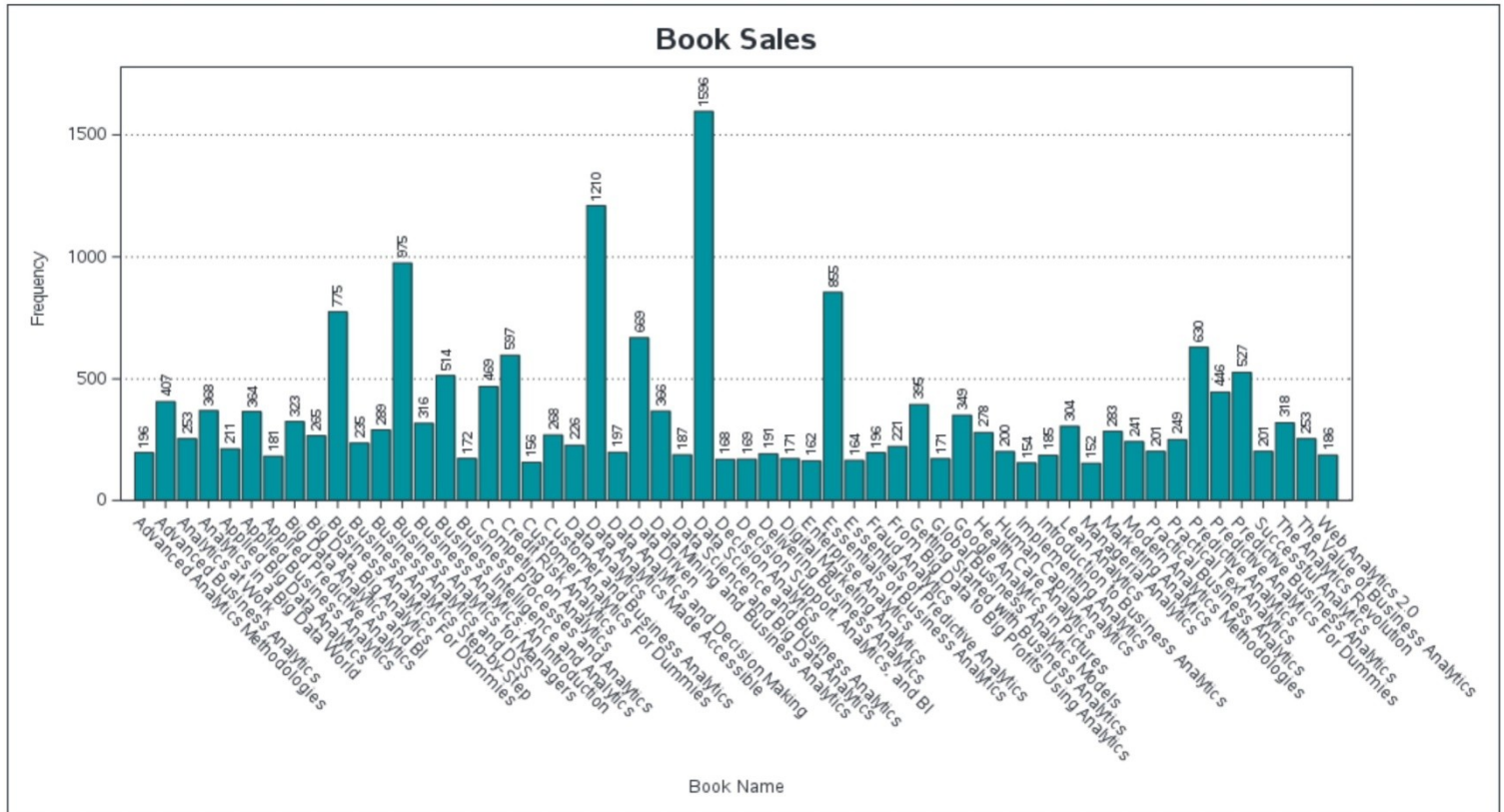
### 1) Executive Summary

The market basket analysis conducted for our business, 'Buy-books-on-line.com' revealed valuable insights into the purchasing behavior of our customers. The analysis involved identifying the frequent itemsets and association rules among the purchased items in our transactions. Based on the results, we can optimize our marketing strategies to increase sales and profitability. The analysis showed that certain items are commonly purchased together, and we can use this information to create product bundles or cross-sell complementary products to customers. We also identified the top-selling groups of items and can prioritize their availability in our store to meet customer demand.

To be precise, we identified a couple of books that sell frequently with 4 individual ones from our store. Such itemsets could be sold together as a bundle in order to increase sales. Additionally, we identified the books most commonly bought together and have presented all the sales of the store in a bar chart for your convenience.

## 2) Book Sales Bar Chart

Below follows a bar chart containing the book sales of Buy-books-on-line.com.



As we can see, Data Science and Business Analytics is the book with the most sales, which is 1,596.

## 3) Association Rules

Regarding the recommendations for book purchases, we have the following proposals. These results were achieved by sorting the results of the basket analysis in decreasing order of Lift, filtering the LHS to '1' and picking the occurrence of each book keeping the recommendation with the maximum Lift.

Target Book	Bought with			Lift
Managerial Analytics	Implementing Analytics	&	Web Analytics 2.0	11,47214752
Implementing Analytics	Data Science and Big Data Analytics	&	Managerial Analytics	11,33032185
Customer Analytics for Dummies	Decision Analytics	&	Enterprise Analytics	11,19203099
Enterprise Analytics	Customer Analytics For Dummies	&	Managerial Analytics	11,07350427

We are going to use 'Managerial Analytics' book as an example. Interpreting its Lift value, we can say that customers who buy Managerial Analytics are 11.47 times more likely to buy Data Science and Big Data Analytics & Managerial Analytics than customers who do not buy Managerial Analytics. In other words, the purchase of Managerial Analytics is strongly associated with the purchase of Data Science and Big Data Analytics & Managerial Analytics, and recommending the last two books to customers who have purchased the first can be a profitable strategy for the Buy-books-on-line.com store.

#### 4) Top triplet of books

The triplet of books mostly bought together is:

Data Science and Business Analytics    &    Business Analytics for Managers    &    Data Analytics Made Accessible
--

(Lift = 1.154526463)

- This means that customers who buy Data Science and Business Analytics are 1.15 times more likely to buy Business Analytics for Managers & Data Analytics Made Accessible.
- This triplet of books has been bought together 794 times (looking at the 'Count' column).
- The Support of this triplet is 41.877637131. This means that by counting all the purchases made in the online store, we find that this particular triplet of books appears in 41.88% of the total purchases.

## Case Study 2

### 1) Executive Summary

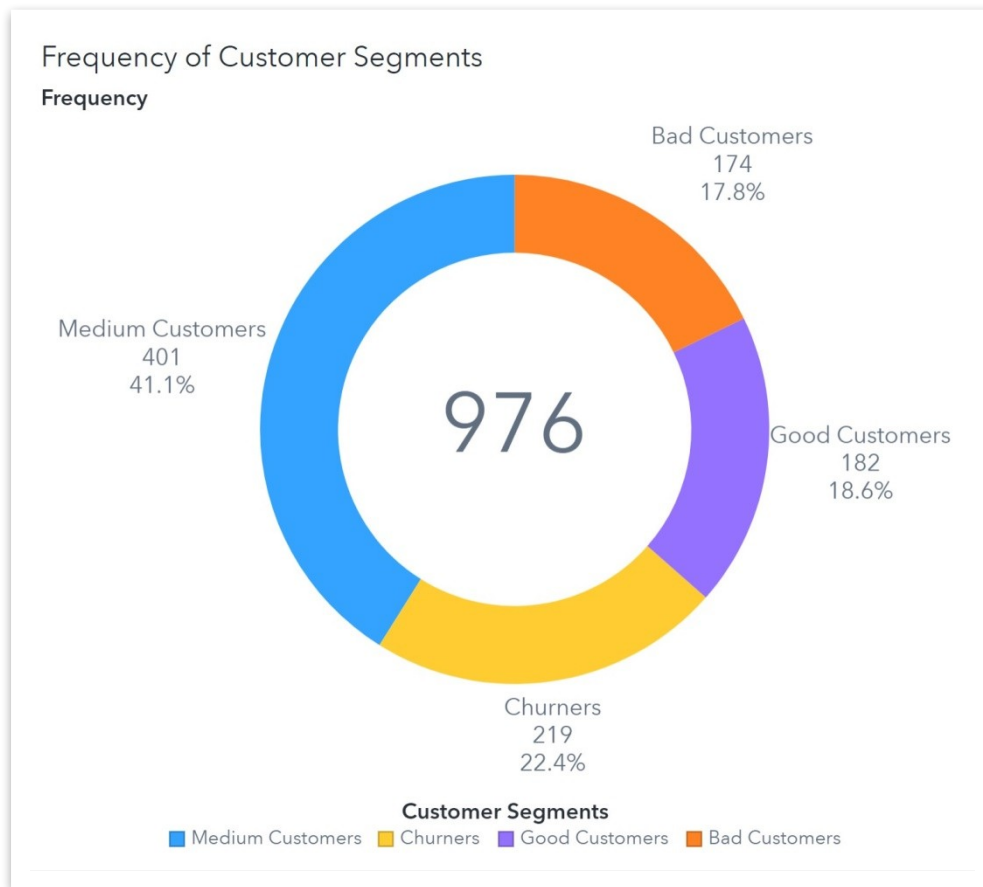
The clustering procedure conducted in this report utilized RFM data of customers as input to create customer segments for our online sports clothing shop. The resulting clusters represent four distinct customer segments: 'Churners', 'Good Customers', 'Bad Customers', and 'First Time Customers'. These segments were created based on the Recency, Frequency, and Monetary value of customer purchases, which allowed for a more targeted approach to marketing and customer retention strategies. In addition to identifying customer segments, this report also proposes specific marketing actions for each segment. By tailoring marketing actions to specific customer segments, our business can optimize their marketing efforts and improve customer retention.

### 2) Clustering and Customer Segmentation

In order to break the customers apart into segments for better comprehension and analysis we performed a clustering procedure using the Recency (R), Frequency (F) and Monetary value (M) of transactions made by the customers. After running the clustering procedure (details of which can be found in part 3 of this case study) and computing the average values of R, F and M (which were used in order to distinguish between the clusters) we created the customer segments as can be seen in the image below:

Cluster ID ▲	Customer Segments ▲	Frequency	F	M	..... R
1	Churners	219	2.7214611872	€247.48	17.698630137
2	First Time Customers	401	4.2793017456	€373.28	5.7506234414
3	Bad Customers	174	5.8218390805	€591.74	14.33908046
4	Good Customers	182	7.7637362637	€778.13	5.2362637363
Total		976	4.8545081967	€459.49	9.8668032787

We can also see the percentages of the clusters within our total customers in the following graph:



### Churners (C1)

This customer segment is characterized by the lowest F and M and worst R of all other clusters. They are people who have spent little money to buy goods, they have made a small number of purchases in total and, most importantly, they have not made any purchases for a considerable amount of time. This shows us that these people have either completely ceased to buy from our store or are in the process of ceasing. 219 customers fall into this category.

### First Time Customers (C2)

This customer segment is characterized by the second lowest F and M and second best R of all other clusters. They are people who have spent pretty low amounts of money to buy goods, they have made a mediocre number of purchases in total but, most importantly, they have made their purchases relevantly recently. This shows that these people have only recently begun purchasing from our store and could have a great potential. We have 401 customers in this category.

### **Bad Customers (C3)**

This customer segment is characterized by the second highest F and M and second lowest R of all other clusters. They are people who have spent little money to buy goods, they have made a small number a mediocre number of purchases in total and, most importantly, they have not made any purchases for a considerable amount of time. This makes them very low quality customers. This category is made up of 174 customers.

### **Good Customers (C4)**

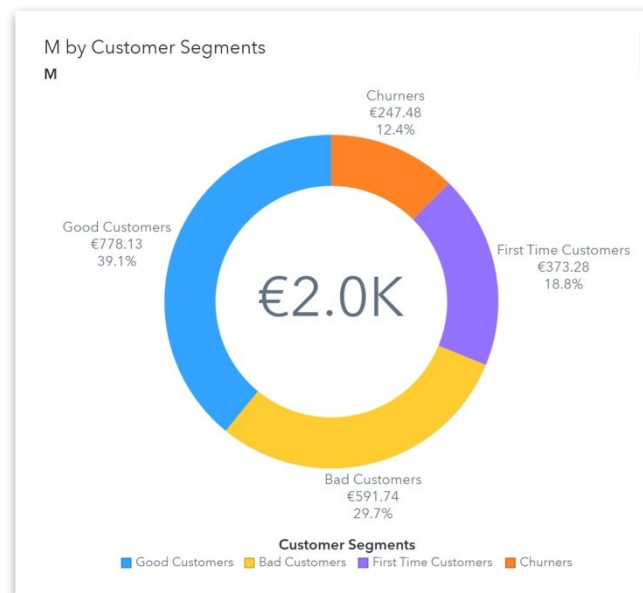
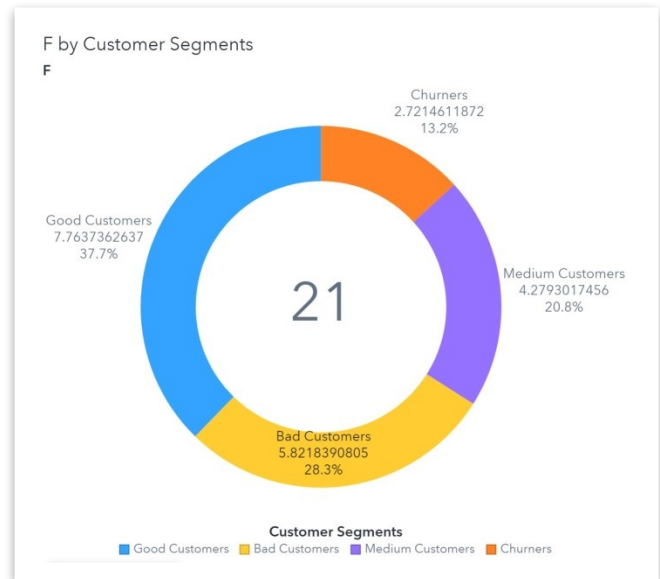
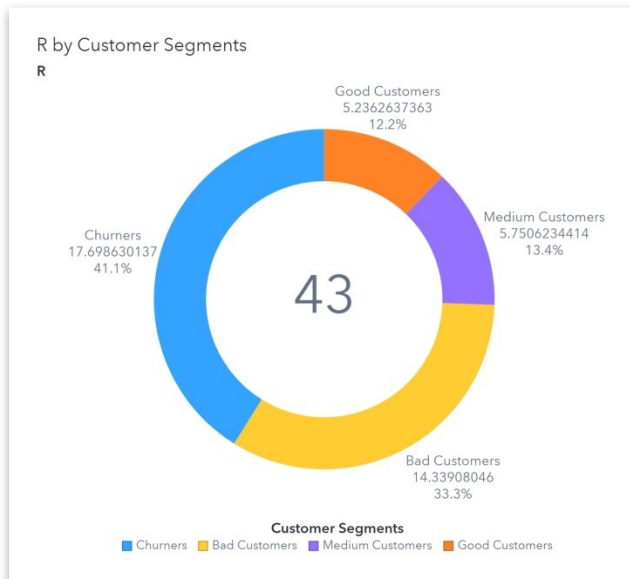
This customer segment is characterized by the highest F and M and lowest R of all other clusters. They are people who have spent a lot of money to buy goods, they have made the biggest number of purchases in total and they have made purchases very recently time-wise. These are the best type of customers available. They are loyal and who are worthy of being retained. We have 182 customers in this category.

## **Clusters' graphs**

A few additional graphs regarding our clusters will further give us an insight into their significance.

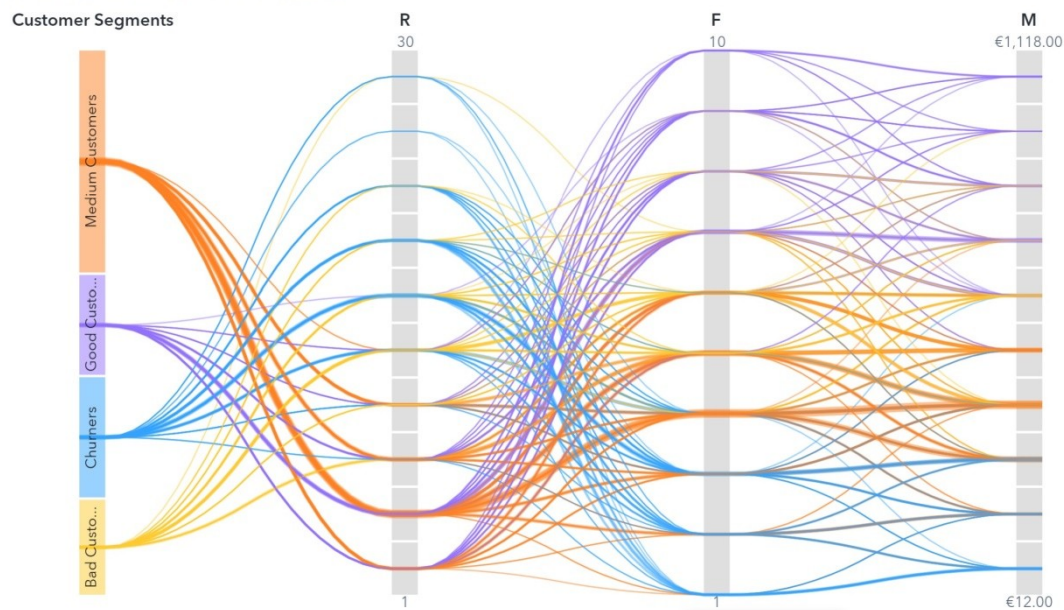
The bellow pie-charts by customer Segments showing the relationship between the R, F and M variables and customer segments can be helpful in providing a high-level overview of the distribution of customers across different recency, frequency and monetary value categories within each segment. It can visually represent the proportion of customers in each segment based on their variable (R,F,M) values.





The Parallel Coordinates graph can help us visualize the relevant fluctuation of the R, F and M variables within the same clusters and intuitively compare them to other clusters. It allows for simultaneous comparison of multiple variables across different clusters or groups.. The Parallel Coordinates graph enables the identification of clusters based on their distinct patterns. Clusters with similar behavior or characteristics will exhibit similar patterns in the graph, making it easier to visually distinguish them. Finally, it allows us to identify and analyze common patterns within clusters. We can look for consistent trends, such as increasing or decreasing values across certain variables, that are shared by a majority of data points within a cluster.

Parallel Coordinates of Selected Variables



### 3) Marketing Actions

Marketing to our customers can be a challenging task, but it's essential for any business that wants to succeed. One effective way to optimize our marketing efforts is to segment our customers based on their behavior and characteristics. By dividing our customer base into different groups, we can tailor our marketing campaigns to each segment's specific needs and preferences, increasing the likelihood of attracting and retaining customers. In this part, we'll discuss some effective marketing actions we can take for the four different customer segments we identified in the previous part of this analysis: Churners, First Time Customers, Bad Customers, and Good Customers.

#### Churners

This customer segment requires some special handling. They are people who either completely ceased to buy from our store or are in the process of ceasing. Thus we need to give them new incentive to purchase from us again. This can be achieved with the following actions:

- Send personalized win-back offers that incentivize them to return.
- Offer discounts on future purchases to encourage them to come back.
- Remind them of the benefits of being a customer and how they can continue to benefit from your products or services.

## First Time Customers

This customer segment can evolve into either good or bad customers, and it all comes down to the steps taken. Using the right motivation and marketing techniques we can retain these customers and add them to the list of loyal customers. A few such marketing actions are the following:

- Send a welcome message that thanks them for their business and offers a special discount on their next purchase.
- Follow up with them to ensure they're satisfied with their purchase and offer assistance if needed.
- Send personalized product recommendations based on their purchase history.
- Offer loyalty rewards or a referral program to encourage them to become repeat customers.

## Bad Customers

This customer segment is probably the most difficult to handle. Since these people are not loyal customers we need to find the proper marketing action in order to encourage them to purchase more. But such a task can be very challenging if they are not happy with their purchasing experience. Some marketing actions could be the following:

- Identify the reasons why they are unhappy and make changes to address their concerns.
- Offer personalized solutions to resolve any issues they have had with your products or services.
- Offer incentives or discounts to encourage them to give your business another chance.

## Good Customers

This customer segment is the most important one. They are loyal customers how spend a considerable amount of money to our business. Such customers need not only be retained and encouraged to buy more, but also to be made feel appreciated for their loyalty toward us. Thus the required marketing actions toward them must make the feel special. A few such examples are the following:

- Offer exclusive promotions or discounts to show your appreciation for their loyalty.
- Ask for feedback on how you can improve their experience and implement changes based on their responses.
- Send personalized product recommendations based on their purchase history.
- Provide early access to new products or services to reward their loyalty.
- Invite them to participate in customer surveys or focus groups to gather insights and feedback.

## 4) Technical Report

Looking into the given dataset containing the results of the RFM analysis procedure (RFM\_Final\_Practice.sas7bdat) we see that there are no values missing. Thus, no imputing of the data was needed.

### ▼ More information

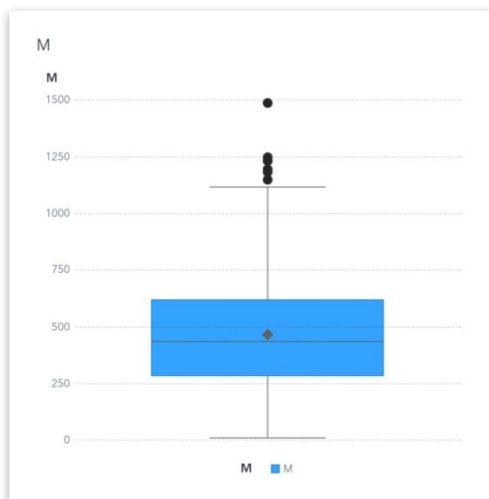
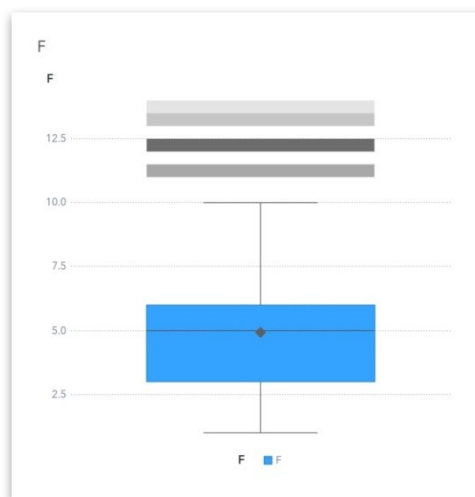
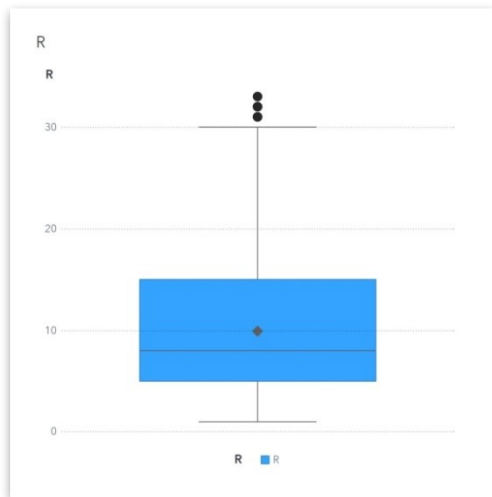
Standard Deviation:	2.20
Standard Error:	0.07
Variance:	4.86
Distinct Count:	14
Number Missing:	0
Total Observations:	995
Skewness:	0.5717
Kurtosis:	0.4293
Coefficient of Variation:	44.6988
Uncorrected Sum of Squares:	29,018.00
Corrected Sum of Squares:	4,828.22

For the execution of the clustering procedure, the numbers were used as they were, since their values were not very dispersed. No log was applied.

Their level was set as intervals with 'T' being the target variable and 'Cust\_ID' the ID.

Variable Name	↑	Label	Type	Role	Level	Order
Cust_ID			Character	ID	Nominal	Default
F			Numeric	Input	Interval	Default
M			Numeric	Input	Interval	Default
R			Numeric	Input	Interval	Default
T			Numeric	Target	Unary	Default

After creating box-plots for all R, F and M we noticed that there were some outliers in the data.



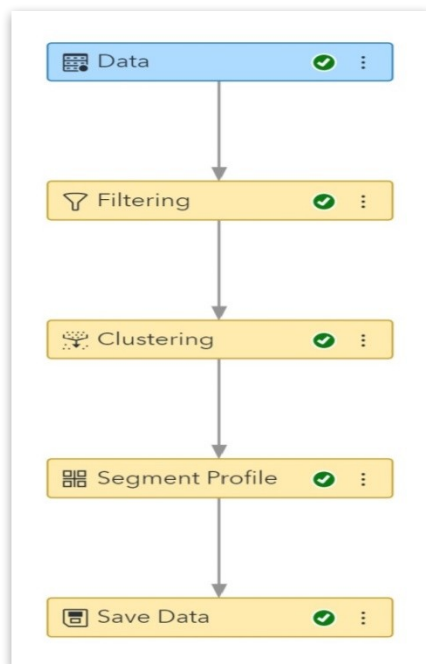
In order to optimize the clustering procedure and avoid the creation of nonsensical clusters, we decided to remove the outliers from the input data.

Thus, we capped the maximum values of R, F and M according to the maximum whiskers of their box-plots, like it is shown in following table

Variable	R	F	M
Max value	30	10	1118

Additionally, a filtering was applied to the data, using the Metadata Limits as the interval filtering limits method.

Below follows the complete pipeline used:



The results of the clustering procedure were saved in a table named 'RFM CLUSTERING TABLE' and then used to create the charts and visualizations used in the precious parts of the Case Study.

The clusters produced can be seen below:

Cluster ID ▲	Customer Segments ▲	Frequency	F	M	.... R
1	Churners	219	2.7214611872	€247.48	17.698630137
2	Medium Customers	401	4.2793017456	€373.28	5.7506234414
3	Bad Customers	174	5.8218390805	€591.74	14.33908046
4	Good Customers	182	7.7637362637	€778.13	5.2362637363
Total		976	4.8545081967	€459.49	9.8668032787

## Case Study 3

### 1) Executive Summary

Our objective was to develop multiple mathematical models to aid the fraud prevention department in predicting the likelihood of a claim being fraudulent. We utilized historical data from May to September 2017 and built four different models: Decision Tree, Maximal Decision Tree, Neural Network, and Logistic Regression. These models were trained using various claim characteristics such as vehicle age, presence of a witness during the accident, time elapsed between the accident and policy termination date and accident location. In the end, after comparing the performance of each model, the Decision Tree seems to be the best one according to the metrics we used and especially the Misclassification Rate.

The developed models can now be applied to new claims issued after October 1st, 2017, allowing us to predict their probability of being fraudulent. Each model provides its own unique insights and strengths in identifying fraudulent claims. Claims with higher probabilities of fraud, as determined by the models, will be directed to the investigation department for further scrutiny. By

leveraging these models, we can optimize our resources and enhance the efficiency of fraud prevention measures.

## 2) Interpretation of the profit matrix

The given profit matrix can be interpreted like so:

		Prediction	
		Fraudulent --> Investigate	Non-Fraudulent --> Compensate
Actual	Fraudulent	The model correctly predicted that a claim is fraudulent	The model falsely predicted that a claim is non-fraudulent, when it was fraudulent
	Non-Fraudulent	The model falsely predicted that a claim is fraudulent, when it was non-fraudulent	The model correctly predicted that a claim is non-fraudulent

### Question 3

In order to calculate the cut-off point using the profit matrix provided, we do the following calculations:

$$\text{Cut-off point: } 1500p_1 - 200(1-p_1) = -1500 \cdot p_1 \Leftrightarrow$$

$$3000p_1 - 200 + 200p_1 = 0 \Leftrightarrow$$

$$3200p_1 = 200 \Leftrightarrow$$

$$p_1 = 200/3200 \Leftrightarrow$$

$$p_1 = 0.0625$$

So, the cut-off point is equal to 0.0625.

### Question 4

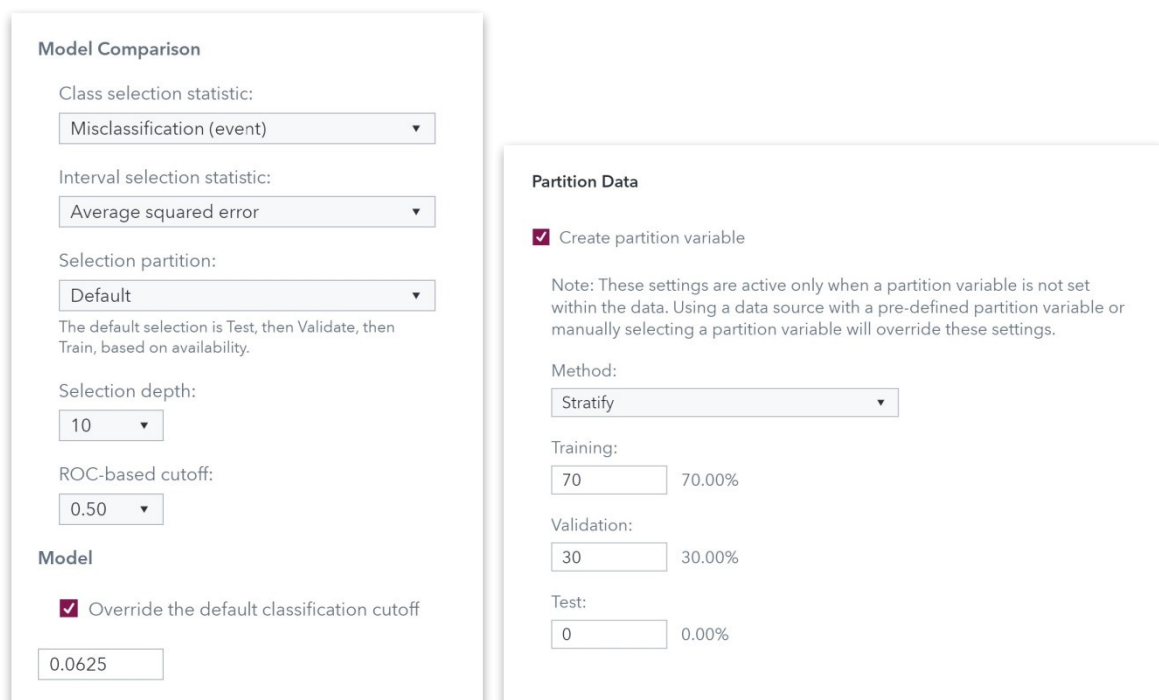
Partitioning our data into training and validation sets is an essential step in building machine learning models. This allows us to train our model on a subset of the data and evaluate its



performance on a separate subset that the model has not seen before. The main reason for partitioning the data is to prevent overfitting, which occurs when a model learns to fit the training data too closely and fails to generalize well to new data. Using a validation set also allows us to estimate the performance of the model on new data. If we evaluate the model only on the training data, it may perform well on that specific dataset, but its performance on new data may be much worse. By evaluating the model on a separate validation set, we can get a more accurate estimate of its performance on new data.

In the context of data partitioning, stratified sampling is commonly used to ensure that the training and validation sets have a similar distribution of target variables. For example, if we are building a binary classification model and the target variable is imbalanced, meaning that one class is much more common than the other, we can use stratified sampling to ensure that both the training and validation sets have a similar proportion of the two classes. By using stratified sampling, we can improve the quality of our model by ensuring that the training and validation sets are representative of the population, and that the model is trained on a balanced sample of the target variable. This can help prevent issues such as overfitting and poor generalization to new data.

Bellow follows a screenshot of the use of the Misclassification Rate (Event) as the performance criterion and the previously calculated cut-off point (0.0625). Also we see the partition of the historical data into training and validation in a 70/30 ratio.



**Model Comparison**

Class selection statistic:  
Misclassification (event) ▼

Interval selection statistic:  
Average squared error ▼

Selection partition:  
Default ▼  
The default selection is Test, then Validate, then Train, based on availability.

Selection depth:  
10 ▼

ROC-based cutoff:  
0.50 ▼

**Model**

☒ Override the default classification cutoff

0.0625

**Partition Data**

☒ Create partition variable

Note: These settings are active only when a partition variable is not set within the data. Using a data source with a pre-defined partition variable or manually selecting a partition variable will override these settings.

Method:  
Stratify ▼

Training:  
70 70.00%

Validation:  
30 30.00%

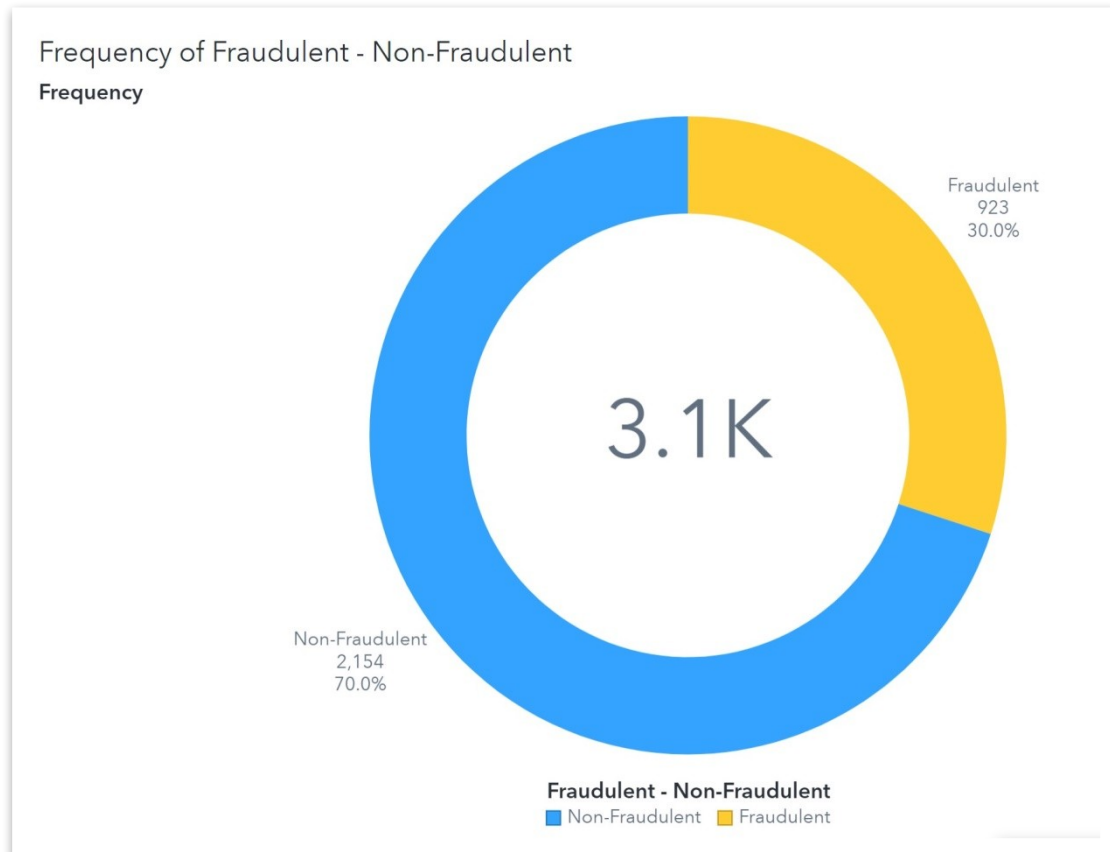
Test:  
0 0.00%

## Question 5

Looking into the data, we see that there are no missing values.

▼ More information	
Standard Deviation:	3.07
Standard Error:	0.06
Variance:	9.44
Distinct Count:	15
Number Missing:	0
Total Observations:	3,077
Skewness:	0.3313
Kurtosis:	-0.1282
Coefficient of Variation:	36.4483
Uncorrected Sum of Squares:	247,800.00
Corrected Sum of Squares:	29,050.85

In order to calculate the proportion of fraudulent and non-fraudulent claims we created a custom category called “Fraudulent – Non-Fraudulent”. The results can be seen in the graph below:



We can see that the proportion of Fraudulent to Non-Fraudulent claims is 30/70.

### Question 6

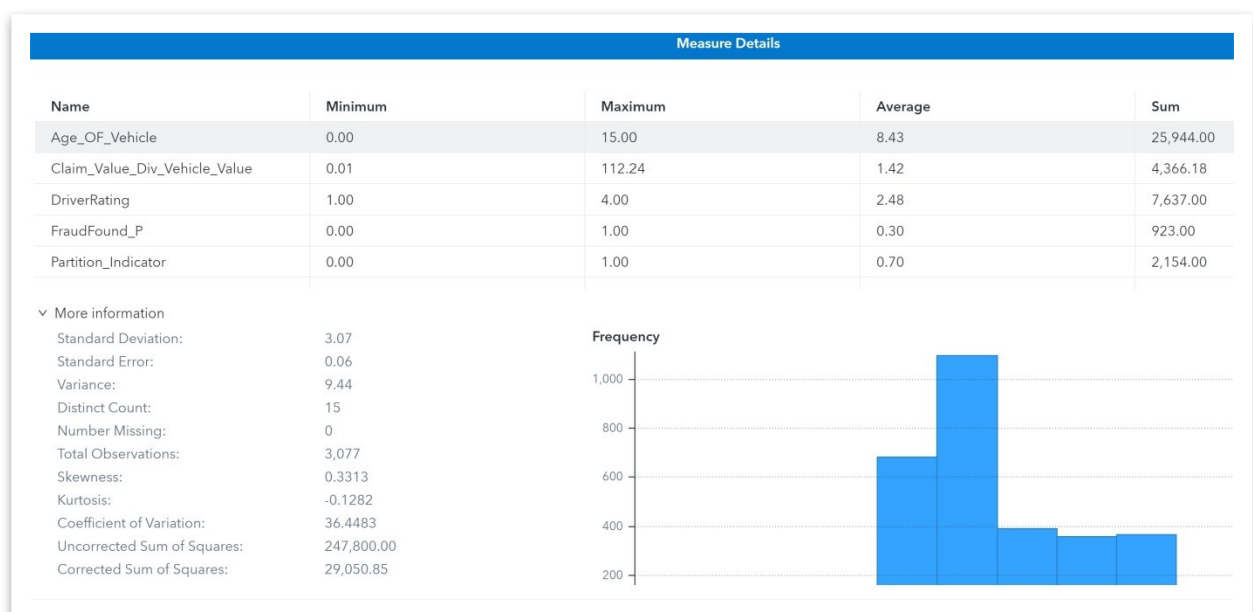
In the case that the proportion of fraudulent and non-fraudulent claims in the historical data set was 10/90 we could utilize oversampling and undersampling to handle the imbalanced data in our machine learning algorithm. We could use them in the following ways:

- **Oversampling the minority class:** We could duplicate or create new synthetic instances of the minority class (fraudulent claims) to balance out the data. Oversampling could help increase the representation of the minority class and provide more examples for the machine learning model to learn from.
- **Undersampling the majority class:** We could randomly remove instances of the majority class (non-fraudulent claims) to balance out the data. Undersampling can help reduce the number of examples from the majority class and prevent the machine learning model from being biased towards this class.

Although, we should keep in mind that even though doth oversampling and undersampling can be useful techniques for handling imbalanced data, they should be used with caution and evaluated carefully to avoid overfitting or underfitting the machine learning model.

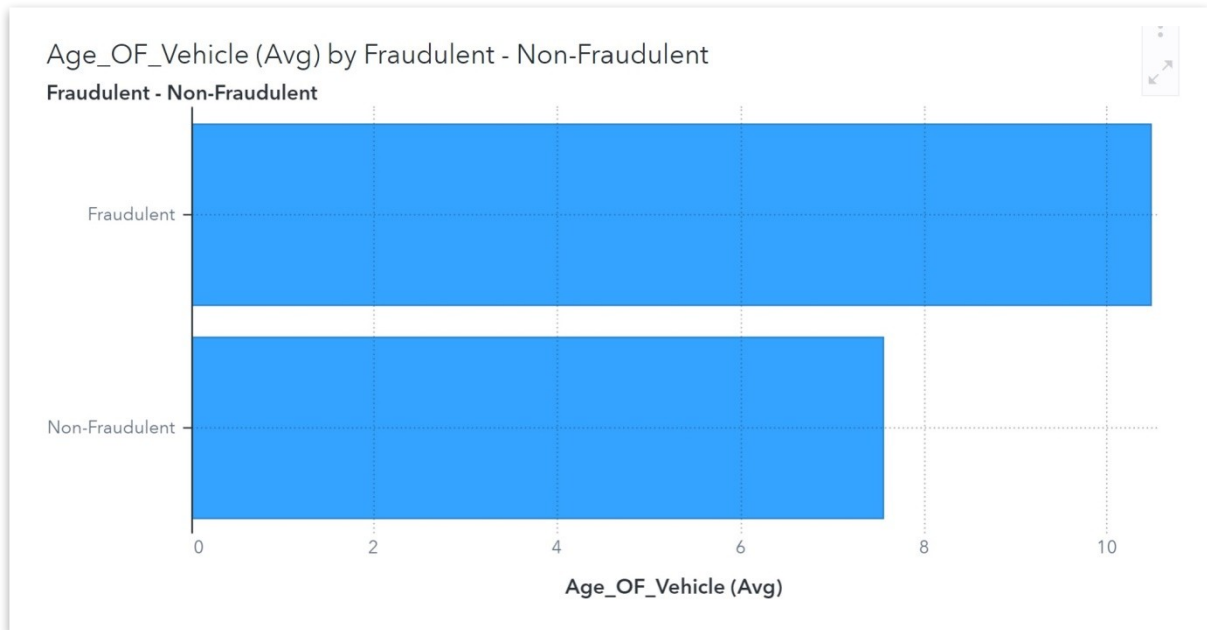
### Question 7

Looking into the details of the “HISTORICAL\_CLAIMS\_PARTITION” dataset, see can see the values of ‘Claim\_Value\_div\_Vehicle\_Value’ go up to 112.24. There are no values equal to or above 120. Thus, we can’t create such a graph that shows the proportion of fraudulent and non-fraudulent claims for those claims that have Claim Value Divided by the Vehicle Value greater than 120%.



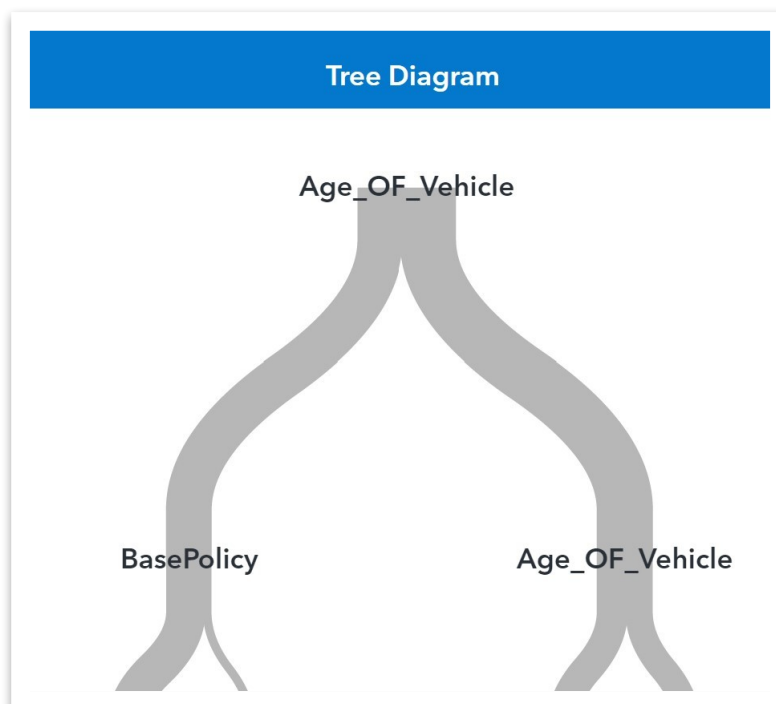
### Question 8

As we can see from the following graph, most fraudulent claims concern cars of greater age. This could be a sign that people with older cars try to trick the insurance company into giving them money in order to fix said cars by filling false claims.



### Question 9

The variable which is used for the first split is the 'Age\_OF\_Vehicle'.

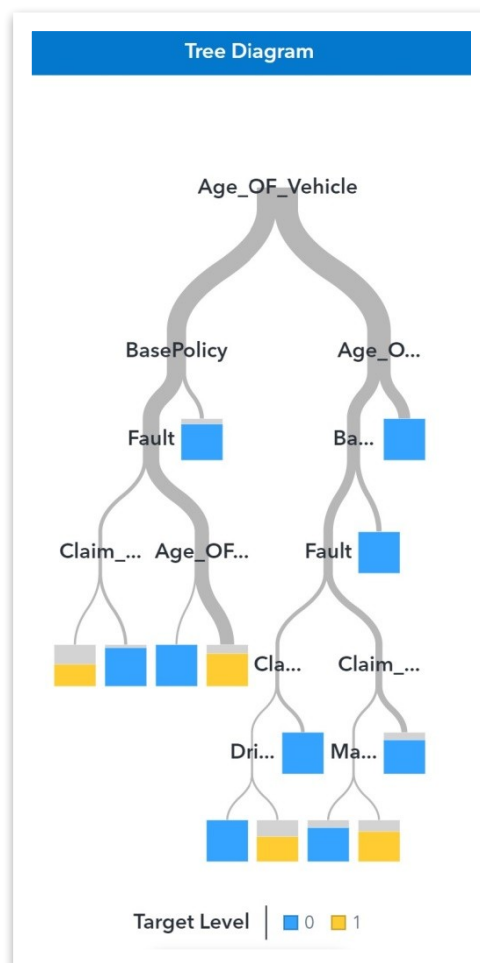


In decision tree algorithms, the first variable to split on is selected based on its ability to best separate the target variable into distinct groups. The logworth is a statistical measure that calculates the significance of the relationship between a categorical variable and the target variable. The

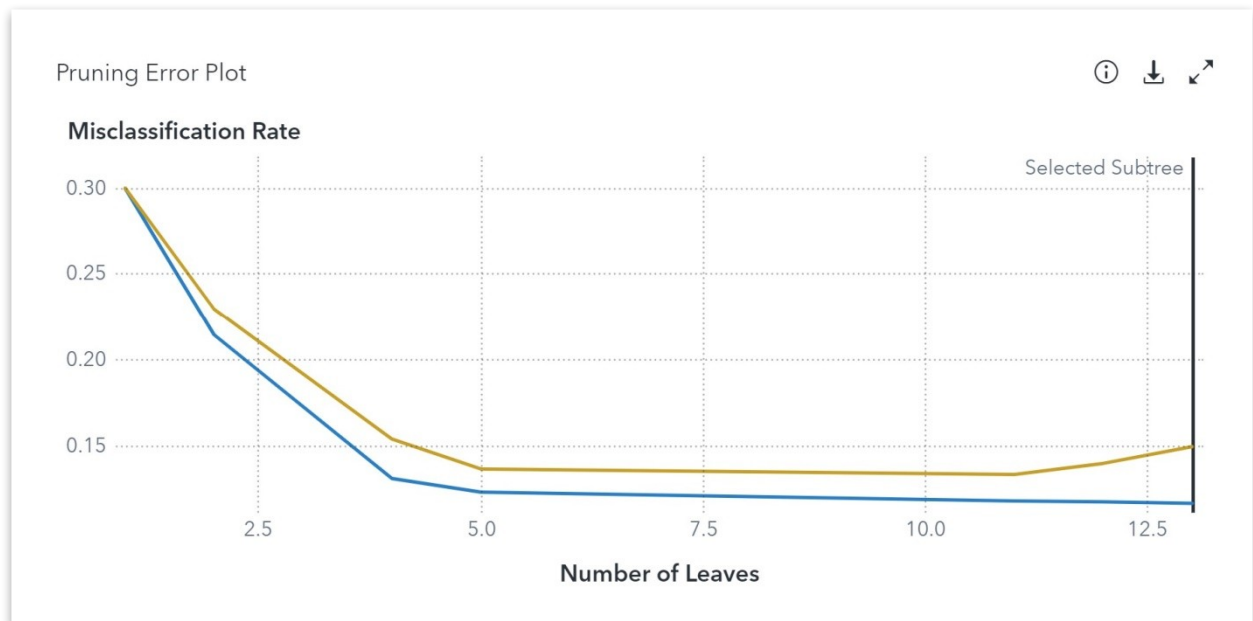
logworth value is one method that can be used to measure the predictive power of each variable in the dataset. When constructing a decision tree, the logworth is used to rank the variables in terms of their predictive power. The variable with the highest logworth value is selected as the first variable to split on because it has the strongest association with the target variable.

### Question 10

This decision tree (Maximal tree) has 12 leaves. It is also the biggest of the two trees that we created.



This type of decision tree is called Maximal tree. The Maximal tree is a type of decision tree that is built by growing the tree until all the nodes are pure (all data points in each leaf node belong to the same class) or until no further splits can be made.



The phenomenon presented in line for the training dataset (blue line) is called overfitting. Overfitting occurs when the decision tree model becomes too complex and starts to fit the training data too closely, resulting in poor generalization to new data. This is indicated by a decrease in the Misclassification Rate on the training dataset as the size of the tree increases.

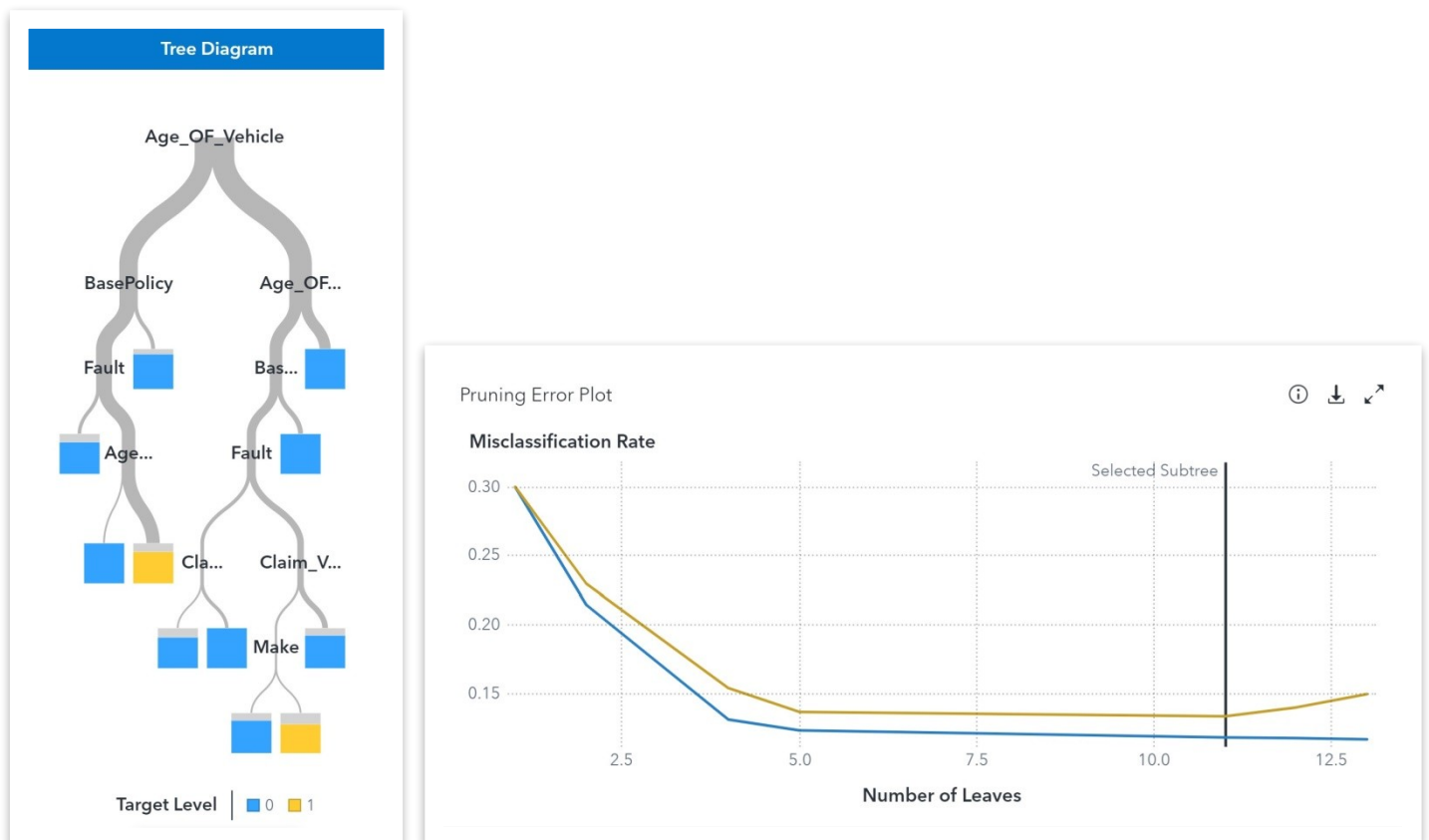
In the subtree assessment plot, we can observe that the Misclassification Rate on the training dataset decreases as the size of the tree increases, while the Misclassification Rate on the validation dataset first decreases and then starts to increase again. This indicates that the optimal subtree is obtained when the tree is pruned to a certain size, beyond which overfitting occurs.

The solution to the overfitting phenomenon in decision trees is to apply pruning to the tree. Pruning is a process of removing branches or nodes from the decision tree that do not contribute to its accuracy. By reducing the complexity of the tree, pruning can improve the generalization performance of the model on new data. The goal is to find the optimal subtree that balances model complexity and accuracy. The optimal subtree is usually determined by evaluating the accuracy of the model on a validation dataset, which is a separate dataset from the training dataset used to build the decision tree. By applying pruning techniques, we can reduce the overfitting phenomenon and improve the generalization performance of the decision tree model on new data.

## Question 11

The optimal tree has 11 terminal leaves.

Bellow follow the optimal Tree Diagram and the subtree assessment plot when Misclassification Rate is selected as the performance criterion.



In the subtree assessment plot, we can observe that the Misclassification Rate on the training dataset decreases as the size of the tree increases, while the Misclassification Rate on the validation dataset first decreases and then starts to increase again. This indicates that the optimal subtree is obtained when the tree is pruned to a certain size, beyond which overfitting occurs. In our case, 11 leaves is the optimal point for the pruning operation to stop.

## Question 12

The decision tree provides valuable insights into the factors that influence the outcome of insurance claims and can be a great assistance when trying to identify fraudulent claims. The key



variables that are most important in determining if a claim is fraudulent or non-fraudulent are in decreasing order of importance:

1. **Age of Vehicle:** The age of the vehicle appears to be a significant factor in detecting fraudulent claims. Older vehicles may be more susceptible to fraudulent claims, as they might be easier to manipulate or have higher chances of pre-existing damage.
2. **Fault:** The fault variable plays a crucial role in determining the likelihood of a claim being fraudulent. Claims where the claimant is not at fault are more likely to be legitimate, while those where the claimant is at fault might raise suspicions of potential fraud.
3. **BasePolicy:** The type of insurance policy, represented by the BasePolicy variable, also influences the probability of a claim being fraudulent. Certain policy types may be more attractive to fraudsters due to higher coverage limits or loopholes that can be exploited.
4. **Claim Value Divided by Vehicle Value:** This variable measures the ratio of the claimed amount to the value of the insured vehicle. Unusually high values in this ratio may indicate potential fraud attempts, as the claimed amount significantly exceeds the vehicle's value.
5. **Make:** The make of the vehicle is another important factor in identifying fraudulent claims. Certain vehicle brands may be associated with higher occurrences of fraud due to factors such as their desirability, market value, or involvement in previous fraudulent claims.

By understanding these key variables, we can better identify potential fraudulent claims and take appropriate actions. Thus, we can use these metrics in order to separate fraudulent from non-fraudulent claims during the first screening phase.

Bellow follows a table containing the aforementioned key variables:

Variable Importance						
Variable Label	Role	Variable Name	Validation Importance	Importance Standard D...	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

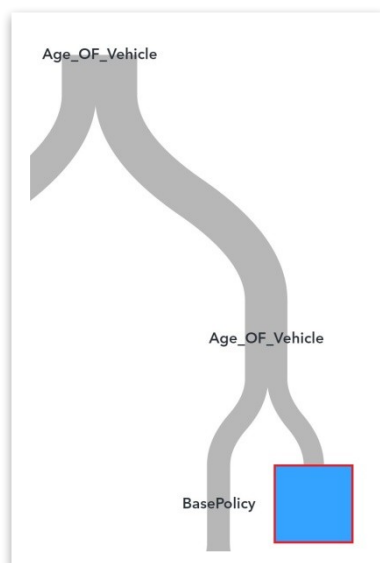
Now, let's interpret five leaves of the decision tree using the above key variables:

- If a car has 'Age\_OF\_Vehicle':<8 or MISSING + 'Age\_OF\_Vehicle':<7 or MISSING,  $p1=0.00\% < 6.25\%$ , thus the claim is Non-Fraudulent
- If a car has 'Age\_OF\_Vehicle':>=8 + 'BasePolicy': Liability,  $p1=13.19\% > 6.25\%$ , thus the claim is Fraudulent
- If a car has 'Age\_OF\_Vehicle':<8 or MISSING + 'Age\_OF\_Vehicle':>=7 + 'BasePolicy': Liability,  $p1=0.49\% < 6.25\%$ , thus the claim is Non-Fraudulent
- If a car has 'Age\_OF\_Vehicle':>=8 + 'BasePolicy': Collision or All Perils + 'Fault': Third Party,  $p1=21.09\%$ , thus the claim is Fraudulent
- If a car has 'Age\_OF\_Vehicle':>=8 + 'BasePolicy': Collision or All Perils + 'Fault': Policy Holder + 'Age\_OF\_Vehicle': not MISSING,  $p1=0.00\% < 6.25\%$ , thus the claim is Non-Fraudulent

### Question 13

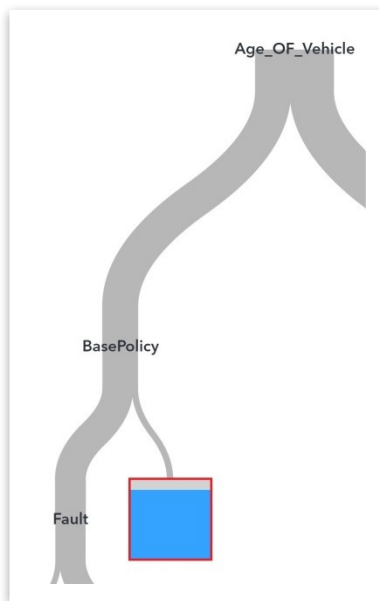
We are going to interpret the leaves in the red boxes in the below screenshots. The way the interpretation is carried out is by following the path leading from the base of the tree all the way to the leaf in question, picking the appropriate value of the variables on each intersection we come across.

a)



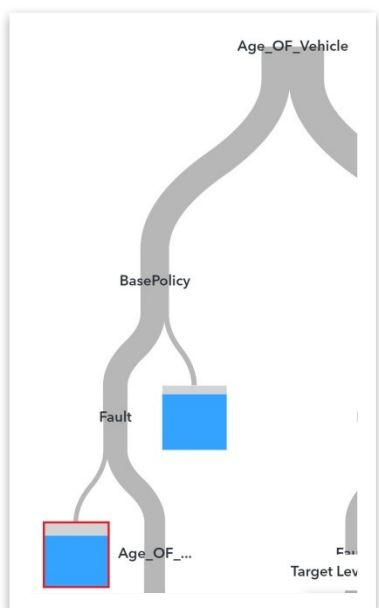
If a car's age is <8 or missing and again if its age is <7 or missing, then that car's claim is non-fraudulent.

b)



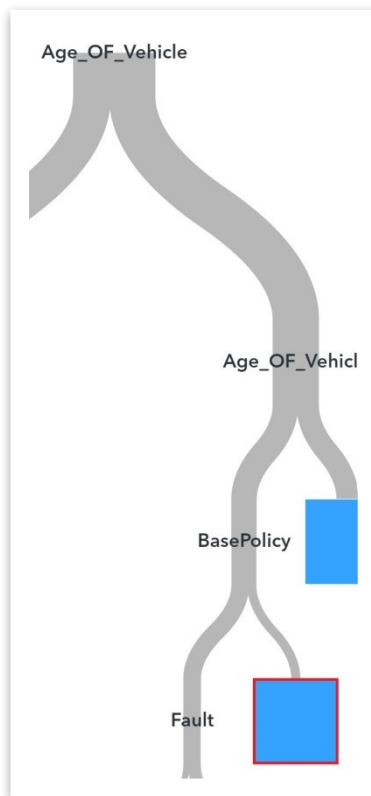
If a car's age is  $\geq 8$  and it has a base policy of Liability, then that car's claim is non-fraudulent.

c)



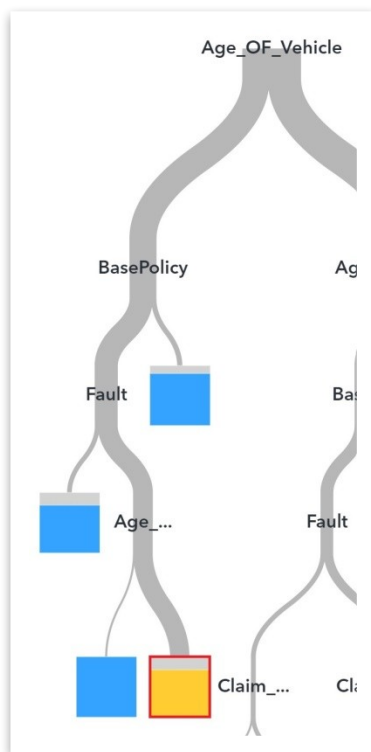
If a car's age is  $\geq 8$  and it has a base policy of Collisions or All Perils and the fault was of a Third Party, then that car's claim is non-fraudulent.

e)



If a car's age is  $<8$  or missing and again if its age is  $\geq 7$  and it has a base policy of Liability, then that car's claim is non-fraudulent.

f)



If a car's age is  $\geq 8$  and it has a base policy of Collisions or All Perils and the fault was of Policy Holder and the car's age is between 0 and 15, then that car's claim is fraudulent.

## Question 14

At the 20% point on the x-axis, the graph shows the cumulative percentage of customers from the validation dataset that have been ranked by the model. This means that the x-axis represents the proportion of customers included in the analysis, sorted from the highest predicted score to the lowest. The y-axis represents the cumulative percentage of actual positive responses. At the 20% point on the x-axis, the graph shows the cumulative percentage of customers from the validation dataset that have been ranked by the model. This means that the x-axis represents the proportion of customers included in the analysis, sorted from the highest predicted score to the lowest. The y-axis represents the cumulative percentage of actual positive responses

At the 100% point on the x-axis, this represents the cumulative percentage of customers considered when the entire dataset is analyzed. The y-axis at this point represents the cumulative percentage of actual positive responses for all customers in the validation dataset. Analyzing the cumulative percentage response at the 100% point provides an overall measure of the model's performance in identifying positive responses across the entire customer population. A higher value on the y-axis at the 100% point indicates a more accurate prediction of positive responses for the entire dataset. It implies that the model has a good understanding of customer behavior and can effectively identify potential positive outcomes.

For example, in our case, the maximal tree using the train dataset, when taking into account the top ranked 20% of the customers can capture 85% of actual positive responses. For the 100% mark all of our models converge to 100% success, which is expected, as they all have taken into account the entirety of the dataset.

Bellow follows the Cumulative Response Percentage graph for our case:

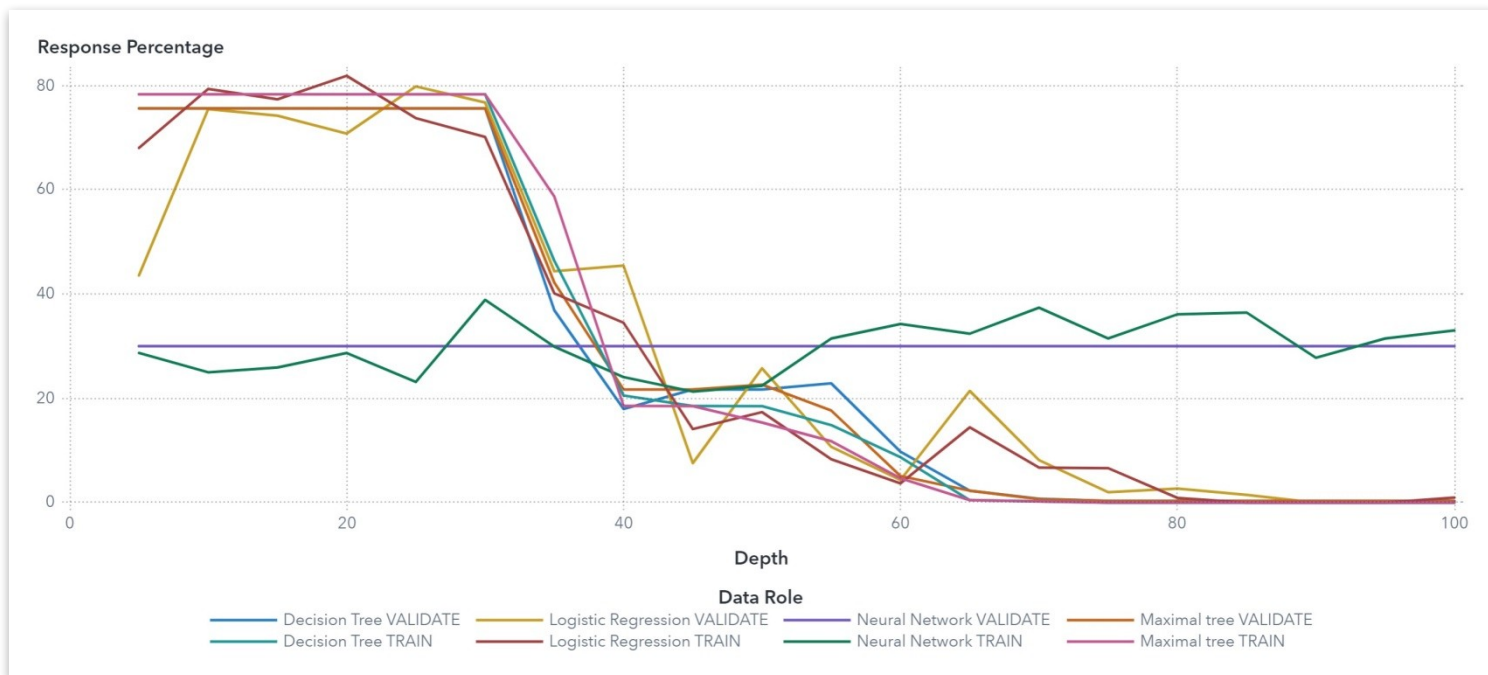


### Question 15

In the percentage response chart for the validation dataset, the x-axis represents the percentage of the customer population that is included in the analysis, rather than specific values. The graph shows the cumulative percentage of actual positive responses on the y-axis for the given percentage of the customer population on the x-axis. At the 25% point, the graph illustrates the cumulative percentage of positive responses observed within the first 25% of the customer population ranked by the model's predictions. In other words, it shows the proportion of positive outcomes achieved among the top 25% of customers based on the model's ranking. For example, in our case, the maximal tree using the train dataset, when taking into account the cumulative top ranked 20% of the customers can capture approximately 87% positive responses.

A higher value on the y-axis indicates a higher percentage of positive outcomes among the top 25% ranked customers, suggesting that the model is effectively identifying and targeting potential positive responders. The percentage response chart helps evaluate the model's ability to identify positive outcomes at different proportions of the customer population. It provides insights into the model's effectiveness in ranking customers based on their likelihood of responding positively and helps understand the potential gains achievable by targeting specific segments of customers.

Bellow follows the Response Percentage graph for our case:



## Question 16

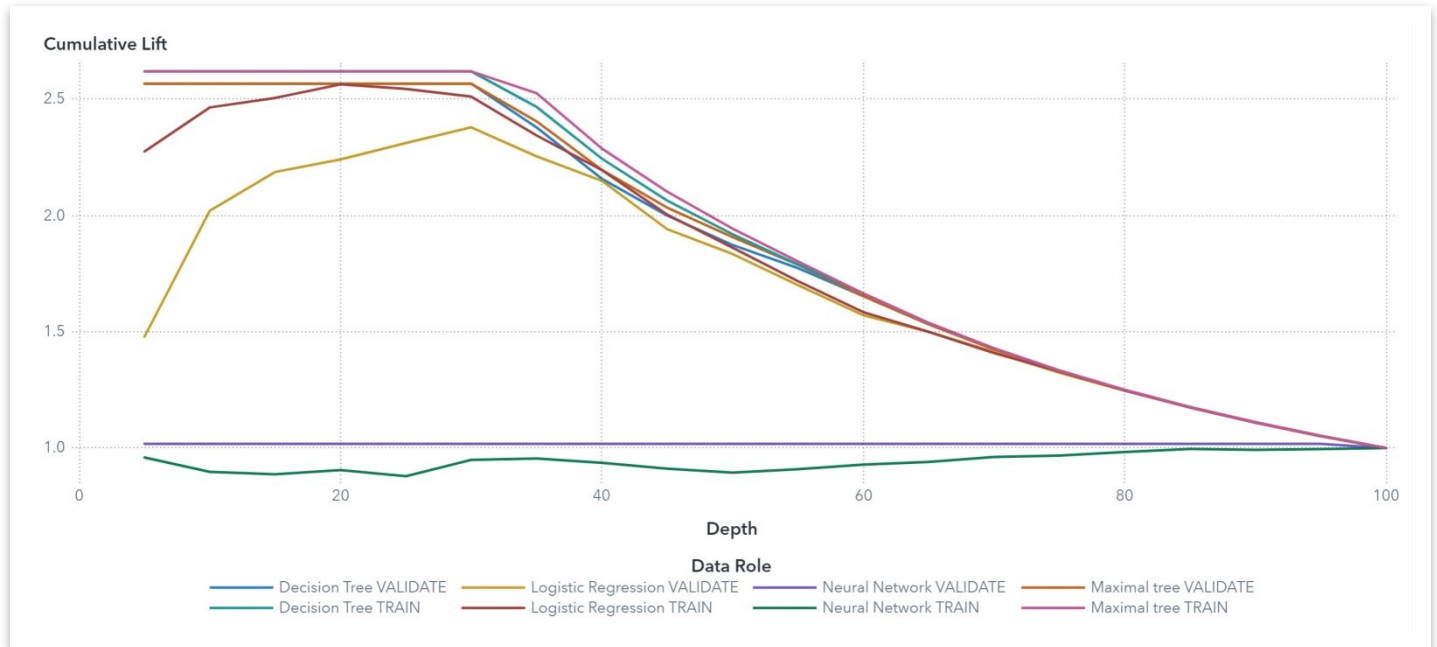
The cumulative lift chart for the validation dataset provides insights into the effectiveness of a model in comparison to a random selection approach. The x-axis represents the percentage of the customer population that is included in the analysis, while the y-axis represents the cumulative lift value. At the 20% point, the cumulative lift chart demonstrates the lift value achieved within the first 20% of the customer population ranked by the model's predictions. Lift measures the performance improvement of a model compared to a random selection approach.

The lift value at the 20% point on the y-axis indicates how much better the model performs in terms of targeting positive outcomes compared to selecting customers randomly. A lift value greater than 1 indicates that the model is outperforming random selection. For example, if the lift value at the 20% point is 2.5 (like the linear regression with the validation dataset in our case), it means that the model is capturing 2.5 times more positive outcomes within the top 20% ranked customers compared to random selection.

The cumulative lift chart helps assess the model's ability to identify and prioritize potential positive responders. Higher lift values at a given percentage of the customer population indicate a more effective targeting strategy. The chart allows us to understand the relative improvement

gained by using the model's predictions compared to random selection as we move through different segments of the customer population.

Bellow follows the Cumulative Lift graph for our case:



### Question 17

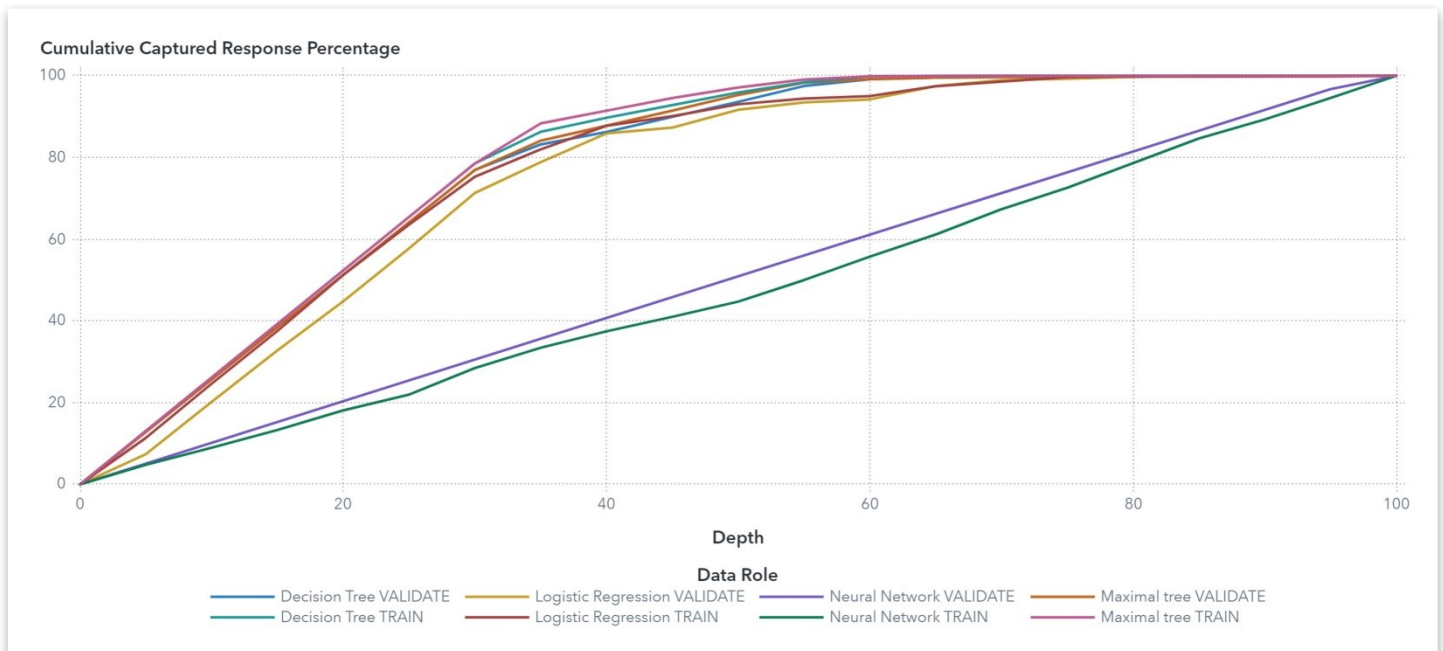
In a cumulative percentage captured response graph, the x-axis represents the percentage of the customer population analyzed, while the y-axis represents the cumulative percentage of positive responses captured. This graph helps us understand the model's ability to identify and capture positive outcomes as we progress through different segments of the customer population.

The 40% point on the x-axis, it represents the analysis of the first 40% of the customer population. The y-axis value at this point indicates the cumulative percentage of positive responses that have been captured within this segment. In other words, it shows the proportion of positive outcomes obtained among the top 40% ranked customers based on the model's predictions. By examining the y-axis at the 40% point, we can assess how well the model is capturing positive responses within this segment of the customer population. A higher value on the y-axis suggests that the model is effective in identifying and targeting potential positive responders, as a larger proportion of positive outcomes have been captured compared to random or baseline selection. For example, in our case, the neural network using the train dataset, when taking into account the top ranked 40% of the customers can capture 85% positive responses.



The cumulative percentage captured response graph helps evaluate the model's ability to prioritize and capture positive outcomes across different portions of the customer population. It provides insights into the model's performance and can guide decision-making in terms of resource allocation, campaign targeting, or customer segmentation strategies.

Bellow follows the Cumulative Captured Response Percentage graph for our case:

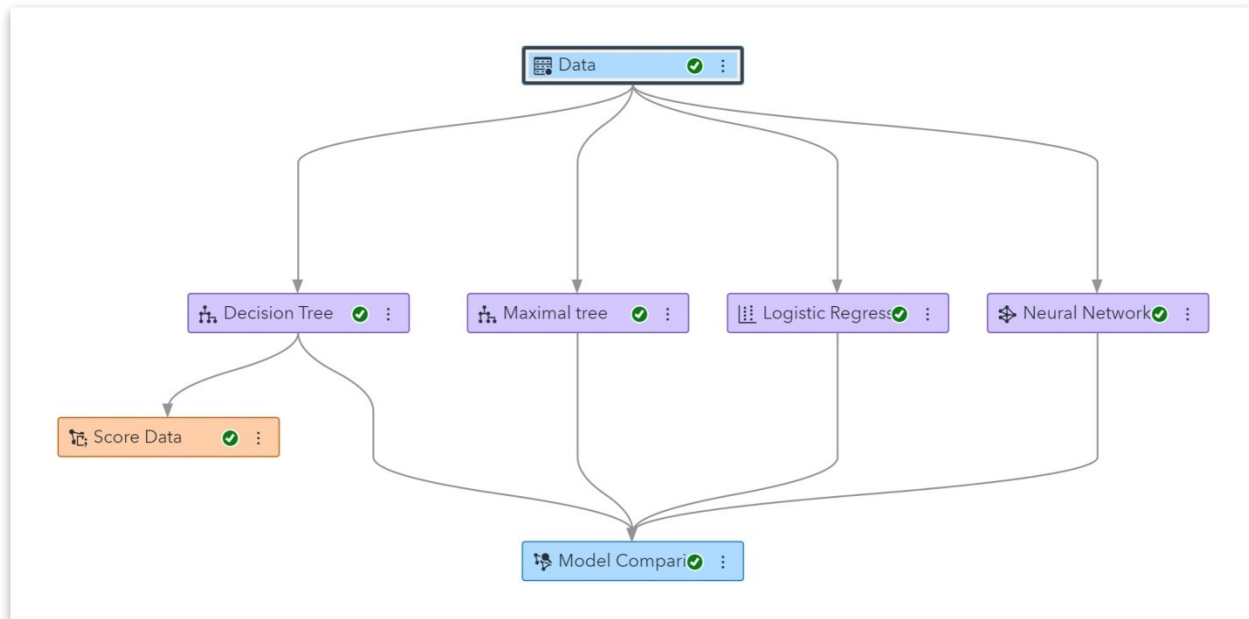


## Choosing the best model and making predictions

Assessing the models that we created, we can see that the best one is the decision tree. This can be seen in the following table from the model comparison:

Champion	Name	Algorithm Name	Misclassification Rate (Event)	Misclassification Rate	Misclassification at Cutoff	Root Average Squared Error	Average Squared Error
	Decision Tree	Decision Tree	0.1333	0.1333	0.2849	0.3238	0.1049
	Maximal tree	Decision Tree	0.1495	0.1495	0.2828	0.3294	0.1085
	Logistic Regression	Logistic Regression	0.1863	0.1863	0.3684	0.3762	0.1415
	Neural Network	Neural Network	0.3001	0.3001	0.6999	0.4583	0.2100

Additionally, we proceed with using said decision tree to predict if the claims in the new dataset (New\_Claims\_Final) are fraudulent or non-fraudulent:



We use CASUSER as the output library and save the results in a table named 'Scored\_Claims'.

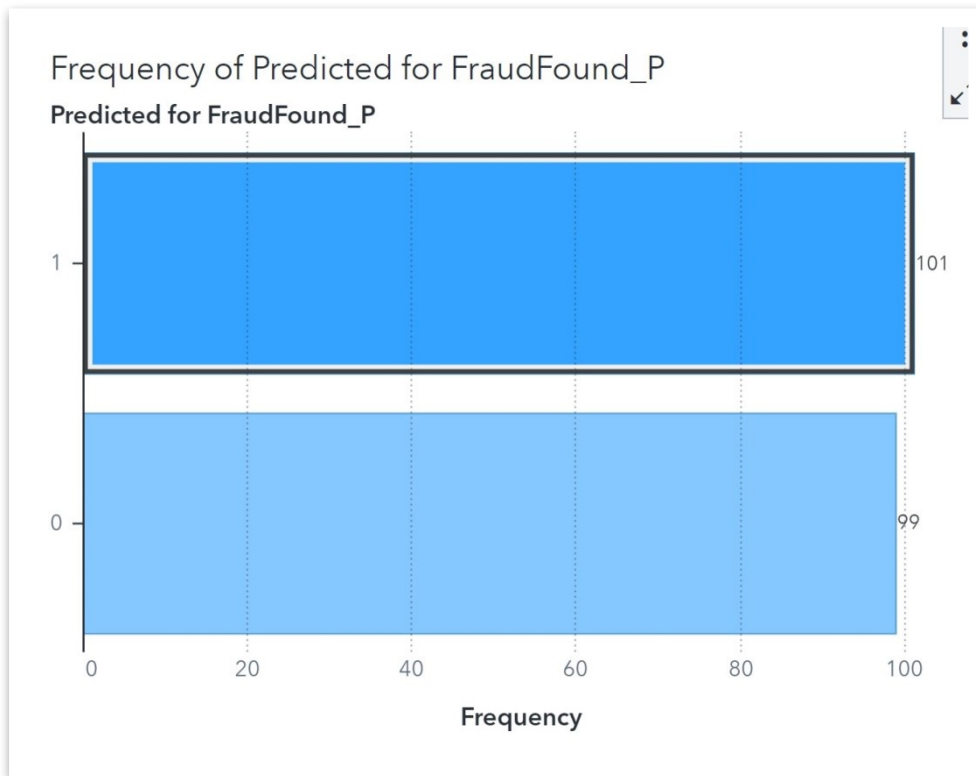
### Question 18

As we can see from the dataset details below, there are 200 claims available.

#### More information

Standard Deviation:	4.41
Standard Error:	0.31
Variance:	19.45
Distinct Count:	8
Number Missing:	0
Total Observations:	200
Skewness:	0.3552
Kurtosis:	-1.1031
Coefficient of Variation:	42.2670
Uncorrected Sum of Squares:	25,649.00
Corrected Sum of Squares:	3,871.16

Of the 200 total claims, 101 have been predicted as fraudulent and 99 as non-fraudulent.  
This can also be seen on the bar-chart below:



### Question 19

Looking into the dataset details below, we can see that the biggest probability of being fraudulent assigned to a claim is 0.78 and the smallest is 0.0.

Name	Minimum	Maximum	Average	Sum
Age_Sex_Premium	0.00	1000.00	200.00	170000.00
DriverRating	1.00	4.00	2.45	490.00
FraudFound_P	0.00	1.00	0.20	40.00
PolicyID	1.00	200.00	100.50	20,100.00
Predicted: FraudFound_P=0	0.22	1.00	0.75	150.29
Predicted: FraudFound_P=1	0.00	0.78	0.25	49.71

## Question 20

The software uses the decision tree model we trained in order to assign 0/1 value to the predicted fraudulent or non-fraudulent claim. The software pushes the line of PolicyID= 15 and PolicyID=107 to the top of the tree and then allows it to reach a terminal leaf by moving left and right on the branches according to the values in the designated variables on each intersection. The terminal leaf it will reach will estimate the characterization of the PolicyID as fraudulent or non-fraudulent.

Additionally, the software assigns a value of  $p1$  to the policy under study. If that value is below the cut-off point then that policy is categorized as Non-Fraudulent and if it is above, then it is categorized as Fraudulent.

In our case:

- Policy15:  $p1=0.0048 < 0.0625$ , thus the claim is categorized as Non-Fraudulent
- Policy107:  $p1=0.7830 > 0.0625$ , thus the claim is categorized as Fraudulent.