Business Analytics Practicum I

Assignment

Name

DIMITRIOS MATSANGANIS, F2822212

FOTEINI NEFELI NOUSKALI, F2822213

Contents

Table of Figures

Case Study 1

Buy-books-on-line.com is a highly reputable online store specializing in the sale of books pertaining to science and information technology. We have been tasked with assisting the analytics department of the store in conducting a market basket analysis of the "Business Analytics" category of books, which comprises 56 titles including Credit Risk Analytics, Marketing Analytics, and Analytics at Work. The aim of this analysis is to identify cross-selling opportunities that will enable the sales department to recommend relevant books to customers based on their previous purchases, thereby increasing overall sales. To this end, we have analyzed a dataset of 19,805 past sales transactions related to the "Business Analytics" book category. In this report, we will outline the methodology employed in conducting the analysis, detail the results obtained, and discuss their implications for the store.

Executive Summary

The objective of this report is to leverage cross-selling opportunities for Buy-books-on-line.com, an online store that specializes in science and information technology books, with a focus on the highly popular category of "Business Analytics". To this end, a Market Basket Analysis was performed using 19,805 past sales transactions. Prior to the analysis, a figure was constructed to visualize the number of book sales by title, enabling the identification of the best- and worst-performing titles.

The Market Basket Analysis aimed to determine the highest probability of a product or group of products being purchased, given that another product or group of products had already been bought or searched for by a customer. Two distinct parts were analyzed. In the first part, four specific books, "Managerial Analytics," "Implementing Analytics," "Customer Analytics for Dummies," and "Enterprise Analytics", were considered, and their strongest relationships were identified. Specifically, for "Managerial Analytics," the two books with the highest association were "Implementing Analytics" and "Web Analytics 2.0". For "Implementing Analytics," the two books with the highest association were "Data Science and Big Data Analytics" and "Managerial Analytics". For "Customer Analytics for Dummies," the two books with the highest association were "Decision Analytics" and "Enterprise Analytics". Finally, for "Enterprise Analytics," the two books with the highest association were "Customer Analytics for Dummies" and "Managerial Analytics". The strongest association rule depending on the Lift criterion among the aforementioned ones is the "Managerial Analytics" grouped by the "Implementing Analytics" and the "Web Analytics 2.0" with a Lift of 11.472.

In the second part, a set of three books that were frequently purchased together by customers was identified. The books with the highest frequency of co-purchase were "Business Analytics for Managers," "Data Analytics Made Accessible," and "Data Science and Business Analytics." These books were found to have a strong association, with a support metric of 41.877. As such, they could be recommended as a bundle to customers who purchase any one of these books, thereby further promoting cross-selling opportunities.

Buy-books-on-line.com is an e-commerce platform that specializes in selling books related to science and information technology, with a strong reputation among the academic community. The store's clientele primarily consists of university professors and librarians purchasing on behalf of their institutions. One of the most popular categories of books offered by the store is "Business Analytics," which includes a list of 56 titles such as "Credit Risk Analytics," "Marketing Analytics," and "Analytics at Work."

Over the past year, 1,896 customers have purchased at least one book from the "Business Analytics" category, indicating a significant demand for these titles. To further increase sales and capitalize on cross-selling opportunities, the store's sales department aims to provide customers with personalized next-best-offer propositions using association rules.

To achieve this goal, the store's analytics department has compiled a dataset of 19,805 sales transactions related to the "Business Analytics" book category and it will be used to generate insights and recommendations for the sales team that can lead to targeted book referential combinations that consist of book titles that present common or complementary interest for the community. In this way the community has benefits as the recommendation procedure facilitates it in the Data Analytics books' collection formation and offers the online bookstore enterprise the opportunity to augment effectively its sales' rate.

Upon initial examination of the dataset, it is possible to gain insight into the popularity of individual books within the "Data Analytics" category offered by the online store. Figure 1 displays a bar chart depicting the sales performance of each of the 56 book titles included in the analysis dataset. The chart provides a visual representation of the varying degrees of success of each book in terms of sales.

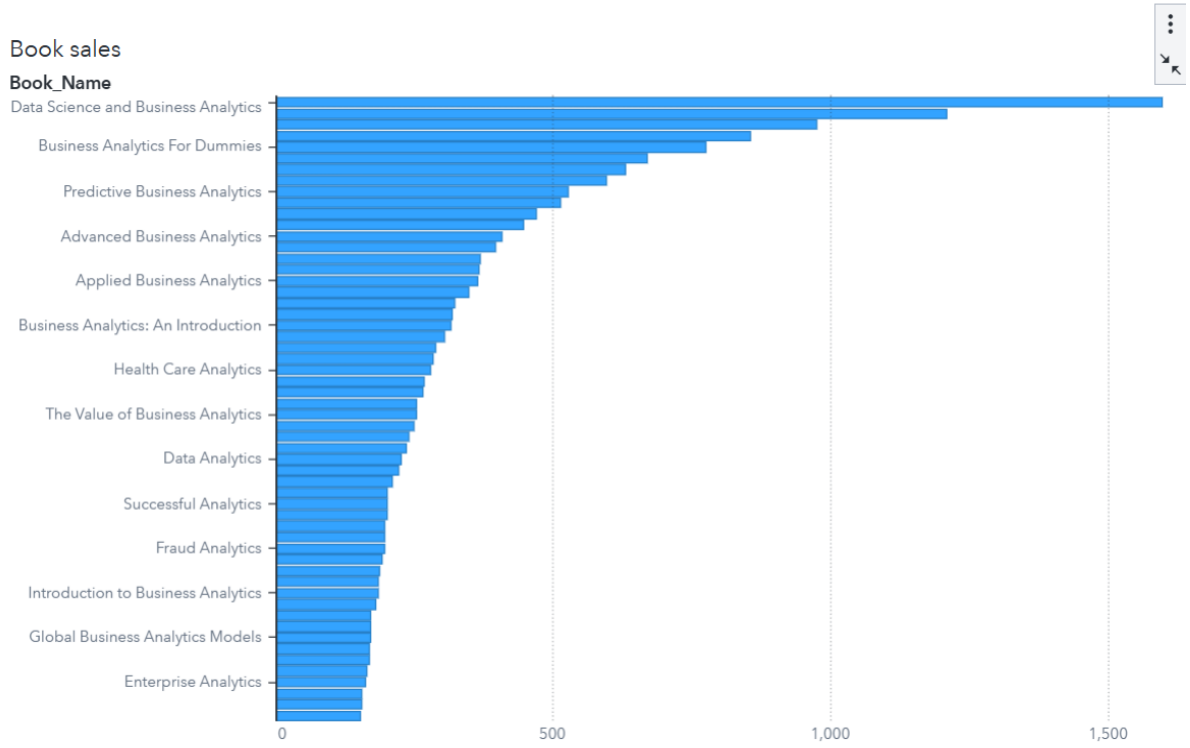


Figure 1: Data Analytics Book sales per book title.

Based on the bar chart in Figure 1, it is evident that the book "Data Science and Business Analytics" had the highest sales volume among all 56 titles in the "Data Analytics" category, with a total of 1,596 units sold. This represents a substantial lead over the second most popular title, "Business Analytics for Dummies," which sold 1,210 units. The book "Predictive Business Analytics" followed in third place with 975 units sold. It is worth noting that the difference in sales between the top-selling title and the third most popular title is significant, with the top-selling book nearly doubling the sales volume of the third-ranked book. Conversely, "Global Business Analytics Models" and "Enterprise Analytics" were the two titles with the lowest sales figures among the 56 titles analyzed.

In order to manage to discover useful associated patterns that combine book titles together that they are constantly bought by customers as a well value combination a Market Basket Analysis will be conducted. Market Basket Analysis is a very popular method that belongs to Pattern Discovery Techniques of Data Mining and Machine Learning.

The Market Basket Analysis technique is widely used by retailers to increase sales by better understanding customer purchasing patterns. In the simplest situation the data for Market Basket Analysis consists of two variables: a transaction and an item. For this case study, the transaction variable is the customer id, meaning that this client made a transaction for a certain group of items, while the item variable is the name of the book that was purchased. A core concept of Market Basket Analysis is the association rule. Association rules count the frequency of items subgroups that occur

together among groups of items coexisting in the same transaction, seeking to find associations that occur far more often than expected and consequently they cannot be considered random. The analysis aims to determine the strength of all the association rules among the set of items and help retailers to improve their business decisions, placement and packaging techniques, advertising strategy, pricing policy and discount patterns.

To implement this Market Basket Analysis method, firstly a relational table with two columns was created. The table had one column that contained the customer id and another column that contained the book that was purchased from the customer considered as a particular transaction-basket. In the simplest situation the data for Market Basket Analysis consists of two variables: a transaction and an item. For this case study, the transaction variable is the customer id, meaning that this client made a transaction for a certain item, while the item variable is the name of the book that was purchased. A core concept of Market Basket Analysis is the association rule. Association rules count the frequency of items that occur together, seeking to find associations that occur far more often than expected.

During the analysis steps, the metrics Support, Confidence, Expected Confidence and Lift are calculated in order to finally understand the associations between the books. The Support measure expresses the intersection of a potential association rule among all transaction-baskets that are occurred in the data set. More specifically, Support measures the times of occurrence of the subgroup of items that represent a rule over the total number of groups-basket-transactions recorded. The Confidence measure expresses the conditional probability that the items on the right side of the rule will be selected given that the items on the left side of the rule have already been selected in case these two probabilities of occurrence are dependent. Finally, the Expected Confidence measure expresses the conditional probability that the items on the right side of the rule will be selected given that the items on the left side of the rule have already been selected in case these two probabilities of occurrence are independent.

A metric called lift determines the strength of each association rule. This metric is calculated as the division of confidence to expected confidence. The lift expresses how probable is a customer that has made certain product selections to proceed in purchasing the products that are also suggested by an association rule in comparison with a total random customer. The final lift metric expresses a ratio of confidence to expected Confidence and consequently the lift measure does not take into consideration the occurrence frequency if an association rule. In this scenario an association rule could have a high lift but a very limited occurrence frequency that is expressed through a very small Support metric and this implies that it is not a very probable association rule in general.

Due to the aforementioned challenge which concerns the Lift metric, the high number of possible associations between products and the need to find the strongest relationships, some constraints were implemented. The association rules were calculated only for rules that had a minimum level of Support greater than or equal to 5%, minimum Confidence greater than or equal to 10% and a minimum Lift greater than or equal to 1. Finally, the maximum number of books that could be contained in each association rule was limited to 3 in total.

After the conduction of the Market Basket Analysis the association rules table is created by the software that contains the association rules along with their Lift measure and other important information as Table 1 shows below.

CASUSER.MBA_RESULTS | Columns: 11 of 11 | Total rows: 169400 | Rows 1 to 200

Enter expression

	LHS	RHS	COUNT	SUPP...	CONF	LIFT	ITEM1	ITEM2	ITEM3	RULE
1	2	1	147	7.753164557	84.482758621	9.3672111313	Business Analytics Step-by-Step	Delivering Business Analytics	Global Business Analytics Models	Business Analytics Step-by-Step & Delivering Business Analytics ==> Global Business Analytics Models
2	1	2	147	7.753164557	62.553191489	7.7516896112	Business Analytics Step-by-Step	Delivering Business Analytics	Global Business Analytics Models	Business Analytics Step-by-Step ==> Delivering Business Analytics & Global Business Analytics Models

Table 1: MBA Association Rules Results table.

The store seeks to find the two books that should be advertised to customers who bought or are searching to buy four particular books (Managerial Analytics, Implementing Analytics, Customer Analytics for Dummies and Enterprise Analytics). To identify the books that should be advertised, we sorted the lift column in descending order. Then, for each of the four books that are of interest, we filtered the Market Basket Analysis table for association rules that contained on the left side only the book of interest. Therefore, we identified the books that the store should advertise in each distinct case and present them in Table 2.

Book Titles of Interest	Highly Associated Book Titles		Lift
Managerial Analytics	Implementing Analytics	Web Analytics 2.0	**11.472**
Implementing Analytics	Data Science and Big Data Analytics	Managerial Analytics	11.330
Customer Analytics for Dummies	Decision Analytics	Enterprise Analytics	11.192
Enterprise Analytics	Customer Analytics for Dummies	Managerial Analytics	11.073

Table 2: Associations Rule.

The association rule with the highest lift with three (3) items where each of the above-mentioned book is on the left side, is the rule where on the left side is the 'Managerial Analytics' book and on the right side are the 'Implementing Analytics' and 'Web Analytics 2.0' books. This means that if a customer has bought the 'Managerial Analytics' book, then it is 11.472 times more probable to buy the 'Implementing Analytics' and 'Web Analytics 2.0' books, compared to a random customer that has not bought the 'Managerial Analytics' book.

When considering rules with a maximum length of three books, the top three books that are most frequently purchased together are "Business Analytics for Managers," "Data Analytics Made Accessible," and "Data Science and Business Analytics." This set of books was found to be purchased together by customers a total of 794 times, indicating a strong association between them. The support metric for this set of books was calculated to be 41.877.

The support metric is a measure of the probability of the intersection of the three books, expressed as the number of times the three books were purchased together by customers ($n(A \cap B \cap C)$) divided by the total number of purchases made on the online store (N) during the study period.

$$support = \frac{n(A \cap B \cap C)}{N}$$

Therefore, a high support metric of 41.877 indicates that these books are frequently purchased together, and thus, could be recommended as a bundle to customers who purchase any one of these.

Case Study 2

The case study is about Sports-OnLine.com, an online retailer that sells sports clothes and shoes. The company has been operating in the market since October 2001 and has recorded 4,906 sales transactions from 995 customers between October 2001 and December 2006. The management team of the store wants to exploit the electronic data captured during these years to better understand the market. After a meeting with the marketing department, it was decided that a customer segmentation analysis should be performed. Based on the available data, a Recency Frequency Monetary (RFM) analysis would be the most suitable technique for the desired objective.

The IT department in cooperation with the Business Analytics department transformed the sales data into RFM format and produced a SAS data set named RFM_Final_Practice.sas7bdat. The data set has 995 rows, each one corresponding to a single customer, and three columns - R, F, and M - representing the Recency, Frequency, and Monetary value of each customer.

The Marketing Analytics consultant is hired to perform the RFM segmentation using machine learning software SAS Visual Data Mining and Machine Learning in SAS Viya. The consultant is required to cluster the customers and profile the segments created. The consultant is also required to name the segments and describe briefly what marketing actions are appropriate for each segment and why.

Executive Summary

As Marketing Analytics consultants for Sports-OnLine.com, we conducted an in-depth analysis of the company's historical data to better understand their market and customer base. Using the machine learning software SAS Visual Data Mining and Machine Learning in SAS Viya, we employed the Recency Frequency Monetary (RFM) analysis technique to segment customers based on their shopping activities. Our analysis was based on data collected between October 2001 and December 2006, and we identified four distinct customer groups: Lost Customers, First Timers, Churners, and Good Customers.

Lost Customers were defined as customers who had not made a purchase for a longer period of time, purchased products less frequently, and spent less money compared to the average customer from the entire customer base. First Timers were defined as customers who had made a purchase more recently, purchased products less frequently, and spent less money compared to the average customer. Churners were defined as customers who had not made a purchase for a longer period of time, purchased products more frequently, and spent more money compared to the average customer. Finally, Good Customers were defined as customers who had made a purchase more

recently, purchased products more frequently, and spent more money compared to the average customer.

We proposed several strategies based on our analysis to improve customer engagement and increase revenue. For First Timers, we recommended special promotions to attract them, as they represent 26.3% of total customers and 20% of total customer monetary. For Churners, we suggested implementing a reactivation program or offering special promotions to regain their business, as they represent 30.6% of total customers and 36.7% of monetary. To improve the engagement of Lost Customers, we recommended trying to make them better buyers, as they represent a high percentage of 20.9% of total customers. Finally, we suggested providing loyalty credits and cross-selling or upselling to Good Customers, as they represent 34.3% of total monetary. Overall, our analysis revealed valuable insights into Sports-OnLine.com's customer base and proposed effective strategies to improve customer engagement and increase revenue. These strategies include reactivation programs, special promotions, feedback, improvement of services, personalized offers, and loyalty products.

Technical Analysis

After importing the data to our software, we took the first step towards analyzing it by creating visualizations to better understand the distribution of our data. We decided to plot frequency histograms for each of the three RFM variables: Recency, Frequency, and Monetary. The following figure contains two histograms for each variable, one is the frequency and the other is the frequency percentage (*See Figure 2*).

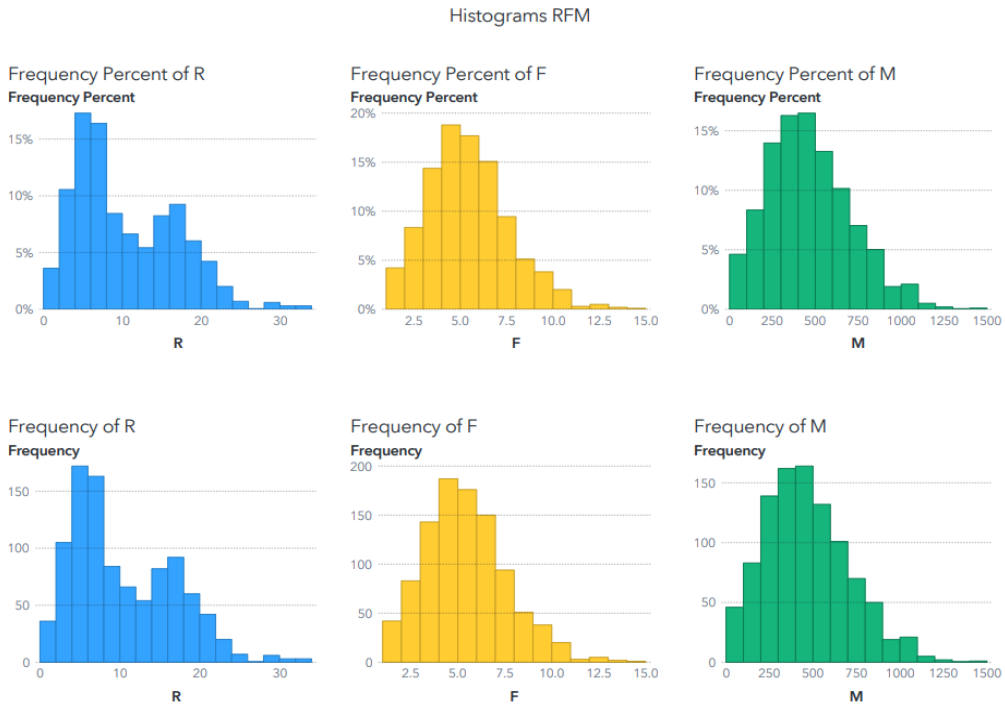


Figure 2: Frequency Histograms for RFM.

The histograms for Recency (R) reveal a bimodal-like distribution with two distinct peaks. The first and higher peak indicates that the majority of our customers made a purchase within 2-8 months prior to the analysis, accounting for a total of 44.22% of the total share. Within this range, the intervals 2-4 months, 4-6 months, and 6-8 months account for 10.55%, 17.29%, and 16.38% of the total share, respectively. The second peak reveals that a significant number of our customers made a purchase over 14 months prior to the analysis. Specifically, the range of 14-16 months and 16-18 months account for 8.24% and 9.25% of the total share, respectively, and together, they make up 17.49% of the total share. Understanding the distribution of Recency is important for businesses as it indicates the amount of time that has elapsed since a customer's last purchase. In this case, the bimodal distribution suggests that there may be two distinct customer segments, one of which purchases regularly while the other makes purchases less frequently. In addition to Recency, we also examined the histograms for Frequency and Monetary.

R (lower)	R (upper)	Frequency Percent
0	2	3.62%
2	4	10.55%
4	6	17.29%
6	8	16.38%
8	10	8.44%
10	12	6.63%
12	14	5.43%
14	16	8.24%
16	18	9.25%
18	20	6.03%
20	22	4.22%
22	24	2.01%
24	26	0.70%
26	28	0.00%
28	30	0.60%
30	32	0.30%
32	34	0.30%

Table 3: Table of the frequencies of R.

Upon analyzing the Frequency (F) variable, the histograms indicate that a significant proportion of our customer base made 3-7 purchases from our company. More specifically, the majority of customers fall in the range of 4-5 purchases, accounting for 18.79% of the total share, followed by 5-6 purchases (17.69%), 6-7 purchases (15.08%), and 3-4 purchases (14.37%). In total, these ranges account for approximately 66% (65.93%) of the total share, indicating that a significant portion of our customer base is repeat purchasers. Therefore, the company may want to focus on increasing the frequency of

purchases from this group of customers. By developing loyalty programs, personalized offers, and targeted campaigns, it may be possible to increase customer loyalty and encourage repeat purchases.

F (lower)	F (upper)	Frequency Percent
1	2	4.22%
2	3	8.34%
3	4	14.37%
4	5	18.79%
5	6	17.69%
6	7	15.08%
7	8	9.45%
8	9	5.13%
9	10	3.82%
10	11	2.01%
11	12	0.30%
12	13	0.50%
13	14	0.20%
14	15	0.10%

Table 4: Table of the frequencies of F.

Finally, the histograms for Monetary reveal that the majority of our customers made purchases between 200-700$, with the highest proportion of customers falling in the 400-500$ range (16.48%). This is followed by the 300-400$ range (16.28%), 200-300$ range (13.97%), 500-600$ range (13.27%), and 600-700$ range (10.15%). Together, these ranges comprise 70.15% of the total share, indicating that a significant proportion of our customers make purchases within this price range. Thus, the company can use this information to develop pricing strategies, promotions, and discounts to encourage customers to purchase more items, or to increase the average order value.

M (lower)	M (upper)	Frequency Percent
0	100	4.62%
100	200	8.34%
200	300	13.97%
300	400	16.28%
400	500	16.48%
500	600	13.27%
600	700	10.15%
700	800	7.04%
800	900	5.03%
900	1000	1.91%
1000	1100	2.11%
1100	1200	0.50%
1200	1300	0.20%
1300	1400	0.00%
1400	1500	0.10%

Table 5: Table of the frequencies of M.

Identifying outliers in our data is an important step in any analysis, as they can have a significant impact on the results. Outliers are data points that fall outside of the typical range of values, and may represent unusual or extreme cases that are not representative of the majority of the data.

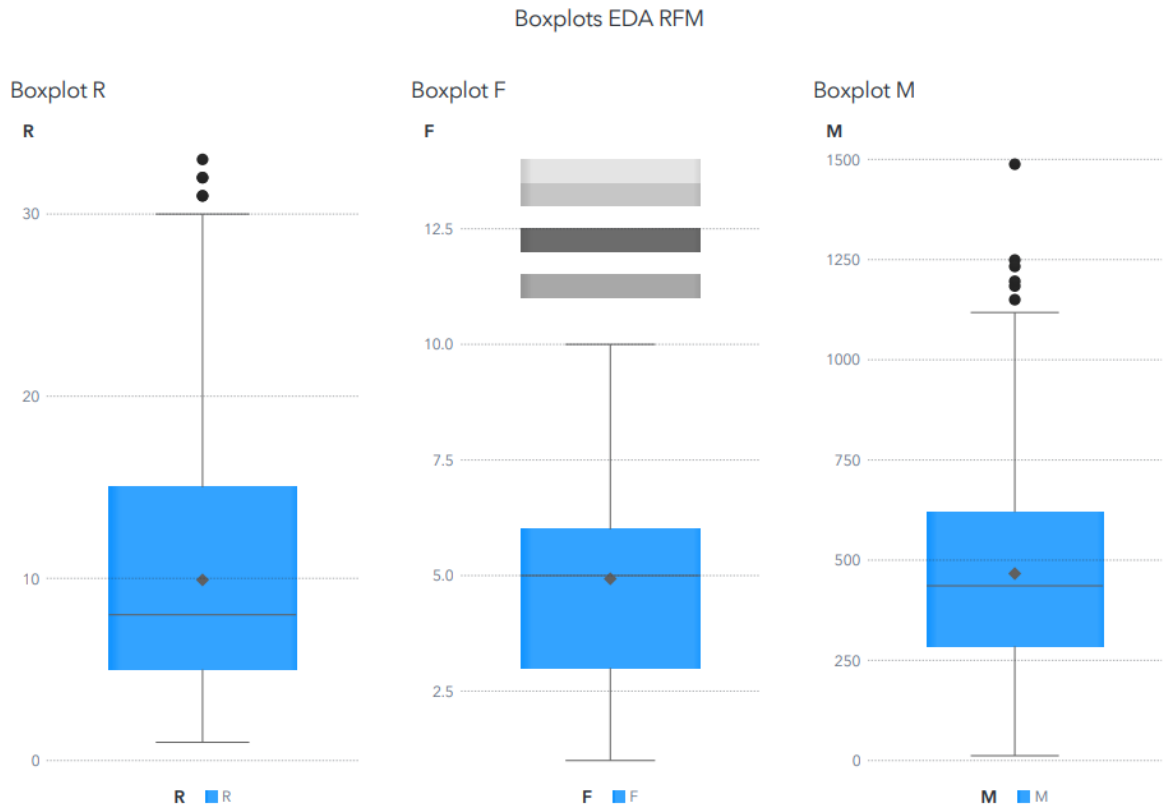


Figure 3: Boxplots for RFM - Outliers Detection Process.

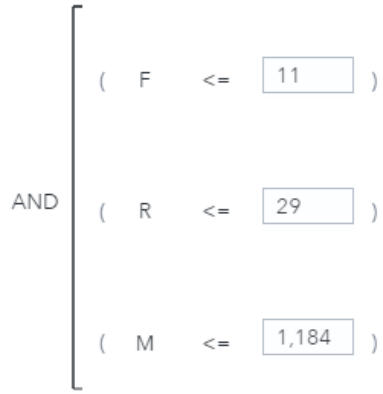
In our analysis of the RFM variables, we identified outliers in each of the three categories. For the Recency variable, the outliers represent customers who made a purchase much less recently than the typical range of values. These customers may represent old or very low-value customers who require special attention or targeted marketing efforts, if the marketing department wants to attempt to make them return to the company. To be more precise, while the median value for the variable R is 8 months the outliers with values of 31, 32, and 33 months are considerably higher than the typical range of values (above the upper whisker). These outliers represent customers who have not made a purchase in an extended period, indicating that they may be lost customers. However, it is worth noting that these customers could have been valuable customers in the past and may require targeted marketing efforts to win back their loyalty.

For the Frequency variable, the outliers represent customers who have made a much larger number of purchases than the majority of our customer base. These customers may be high-value or loyal customers who are worth investing in to maintain their business and increase their lifetime value. More specifically, for the Frequency variable the median value is 5 times of purchases while the

outliers have values of 11-14 times. Some ideas to maintain them could be to implement a loyalty program or other incentives to reward these customers and encourage them to continue doing business with us. This can not only increase their lifetime value but also help to strengthen their loyalty to the brand. Additionally, these high-frequency customers may be good candidates for upselling or cross-selling opportunities, as they have shown a willingness to make multiple purchases.

Finally, from the boxplot of the Monetary variable, the outliers represent customers who have spent a much larger amount of money than the majority of our customer base. These customers may be high-value or high-potential customers who require targeted marketing efforts or additional incentives to continue doing business with our company. The median of this category is 435$, while the outliers have values in the price range of 1150 - 1488$.

However, to ensure the accuracy and reliability of our results, we therefore chose to exclude these outliers from our dataset. In total, only 16 observations were removed out of an initial sample of 995 customers, resulting in a remaining sample size of 979 customers for our analysis. Overall, while outliers can provide valuable insights into customer behavior, excluding them can help to ensure that our analysis is based on a more representative sample of our customer base, which in turn can support more informed and optimal business decisions. Therefore, with the following rule, we excluded the Frequency observations with values over 11, the Recency ones with values over 29, and the Monetary ones with values over 1184$. By making these filtering of the data we ensure that the data will depict the customers behavior, will not be influenced by extreme-value outliers. After the Figure 4 that contains the data filtering rules based on the upper whisker values of the boxplots, we will plot again the boxplots to validate our changes which have been made also to the model (*Figure 5*).




```
( 'F'n <= 11 ) AND ( 'R'n <= 29 ) AND ( 'M'n <= 1184 )
```

Figure 4: Data Filtering Rule to remove outliers.

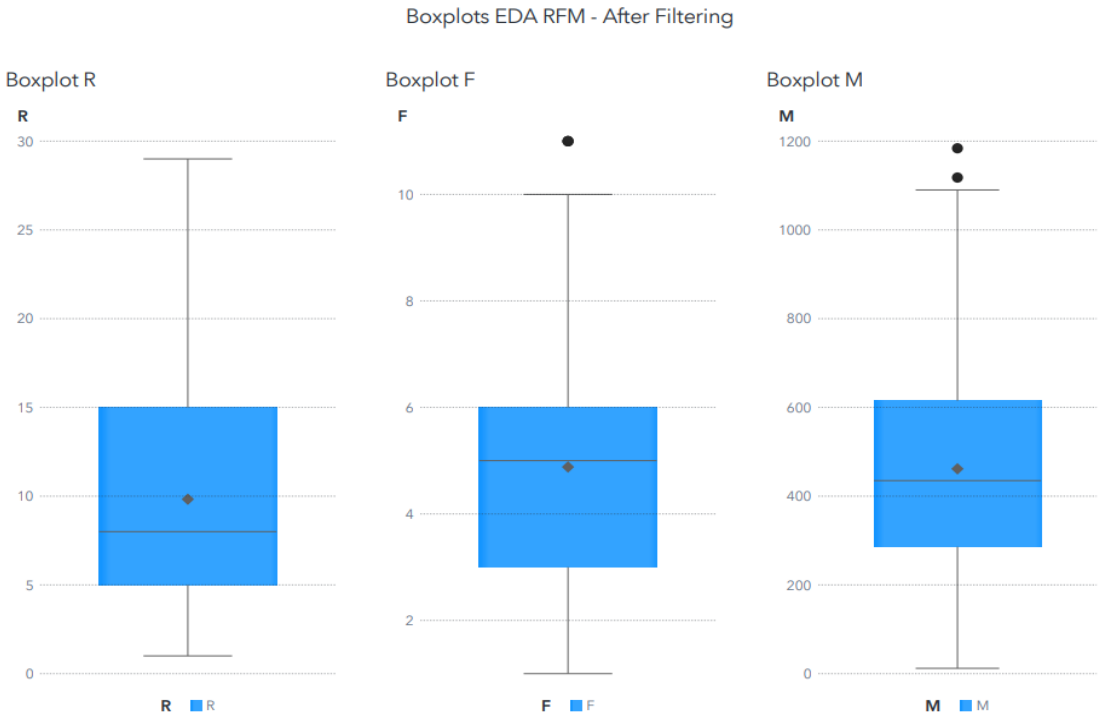


Figure 5: Boxplots for RFM - After Filtering most of the Outliers out.

In order to continue our analysis, after we implemented a filtering process to exclude the outlier observations from our dataset (based on the upper whisker - as mentioned earlier). Additionally, we applied a log-transformation to the input variables to enhance the distributional properties of the data. Our initial dataset consisted of 995 customers, but after the filtering process, 16 outlier observations were removed, resulting in a streamlined dataset of 979 customers for further analysis.

To create customer segments, we employed the K-Means algorithm utilizing the Euclidean distance, which aims to generate clusters that maximize the similarity within each segment while maximizing dissimilarity between different clusters. The clustering process utilized the transformed variables to create distinct segments based on customer behavior. The resulting segmentation model yielded four identifiable group of customers (segments - clusters), as depicted in the Table 4, which presents the crosstable RFM analysis table. To provide a visual representation of the segmentation process and its associated steps, we have included a pipeline diagram (see Figure 6) that outlines the key stages of data preprocessing, filtering, transformation, imputation, and clustering.

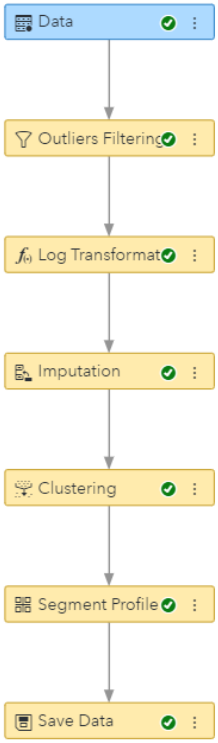


Figure 6: The Pipeline Diagram.

RFM Crosstable

Cluster ID	Segment Names	Frequency	Frequency Percent	R	F	M
1	Lost 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	High Value Customers	217	22.17%	4.2073732719	7.2903225806	715.13824885
Total		**979**	**100.00%**	**9.8273748723**	**4.8815117467**	**461.58835546**

Table 6: RFM Crosstable Analysis.

Table 6 provides a detailed overview of the 4 clusters created using the k-means algorithm, including their Recency, Frequency, and Monetary values. This information was utilized to gain a better understanding of our customers, and to develop specific marketing strategies and promotions for each cluster based on their individual characteristics.

The Frequency column displays how often customers from each cluster make purchases from our company, while the Monetary column indicates the amount of money that the typical customer in each segment spends. The Recency column represents the number of months that have passed since the last purchase made by each cluster's typical customer. The values in these columns depict the mean characteristics of each cluster's ideal customer.

To better identify and differentiate between each segment, we assigned them appropriate names based on the performance of their ideal customer in the three main categories (R, F, and M) compared to the **mean** characteristics of the individuals in our data. To do this, we colored the values in the R, F, and M columns green or red depending on whether they were better or worse than the mean value in the last row of the table depending on the category. For instance, a cluster with a lower Recency value and higher Frequency and Monetary values was named "High Value Customers" while a cluster with lower values in Frequency and Monetary while maintain high Recency was labeled "Churners".

Based on the above information, we have segmented its customer base into four categories or clusters based on their purchasing behavior: Lost Customers, First Time Customers, Churners, and High Value Customers.

The **Lost or Bad Customers** segment refers to customers who have a lower frequency of purchases (the lowest one), lower monetary value (the lowest one), and higher recency than the average customer (the highest one). This means that they have not purchased many items from the company, they don't spend much money, and it has been a long time since they last made a purchase.

The **First Time Customers** segment refers to customers who have a lower frequency of purchases, lower monetary value, and lower recency than the average customer. This is because they are new customers who have not yet made many purchases from the company.

The **Churners** segment refers to customers who have a higher frequency of purchases, higher monetary value, and higher recency than the average customer. However, they used to be good customers but have not made a purchase in a long time.

Lastly, the **High Value Customers** segment refers to customers who have a higher frequency of purchases, higher monetary value, and lower recency than the average customer. This means that they have made many purchases from the company, they spend a lot of money, and it has not been long since they last made a purchase.

Furthermore, we have determined the size of each segment and their respective proportions within the customer base (without the 16 outliers). Additionally, we have examined the average representative customer of each segment based on their purchasing behavior. Moreover, we have calculated the total purchases made by each segment and the corresponding percentage of the overall amount spent in the store during the study period. The above mentioned are presented on the following Figure 7.

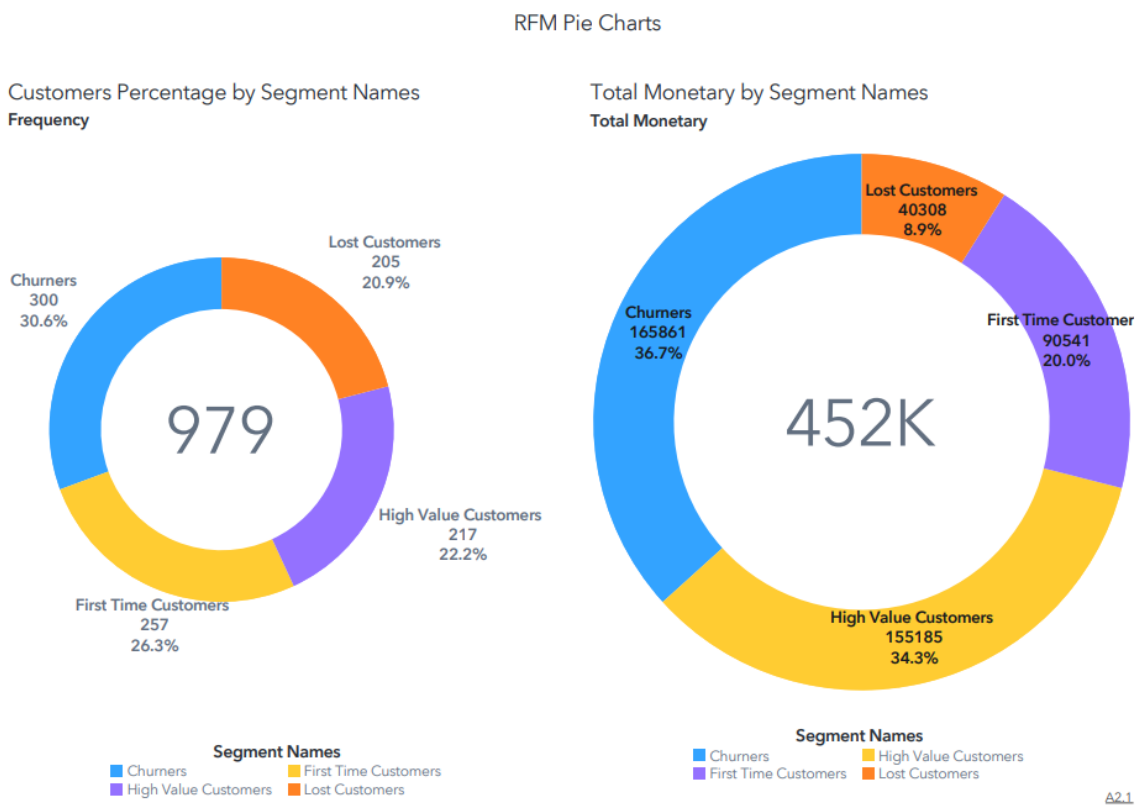


Figure 7: RFM Cluster Analysis Pie Chart.

To summarize this case study, based on the analysis of the customer segments, the following business actions can be proposed to the company:

Lost Customers Segment:

The Lost Customers segment consists of 205 customers, representing 20.9% of the customer base. They have contributed a total of $40,308, accounting for 8.9% of the company's income. The objective for this segment is to convert them into more active customers. To achieve this, it is recommended to initiate direct communication with these customers to gather feedback and understand their perceptions of the company. This feedback can help identify areas of improvement. Additionally, offering targeted promotions and discounts can incentivize them to make more purchases.

First Time Customers Segment:

The First Time Customers segment comprises 257 customers, making up 26.3% of the customer base. They have contributed a total of $90,541, representing 20% of the company's income. The focus for this segment is to encourage them to become repeat buyers rather than one-time purchasers. Given their higher monetary contribution compared to Bad Customers, it is important to nurture their loyalty. Special offers and promotions tailored to their preferences can be effective in driving their

engagement. For example, offering discounts on specific products that are likely to appeal to this segment can increase their frequency of purchases.

Churners Segment:

The Churners segment consists of 300 customers, accounting for 30.6% of the customer base. They have contributed a total of $165,861, which corresponds to 36.7% of the company's income. The primary objective for this segment is to reactivate these customers and regain their loyalty. Since they were previously good customers, it is crucial to make efforts to bring them back to that level. Regularly providing special offers and discounts can entice them to make purchases again. It is also important to establish direct contact with these customers, both via email and phone, to understand the reasons for their disengagement and address any concerns they may have. By making them feel valued and addressing their feedback, the company can strive to regain their trust and loyalty.

High Value Customers Segment:

The High Value Customers segment comprises 217 customers, representing 22.2% of the customer base. They have contributed a total of $155,186, accounting for 34.3% of the company's income. This segment, along with the Churners segment, holds significant importance for the company, as their combined income represents 70.1% of the total. To maintain the loyalty of these customers, it is crucial to make them feel appreciated and valued. Regular communication through emails, providing updates on sales and promotions, can keep them engaged. Additionally, offering personalized benefits such as exclusive discounts or loyalty rewards can enhance their experience. The company should also consider cross-selling and upselling strategies to maximize their purchases and further increase their monetary contributions.

Case Study 3

This case study focuses on a fictitious company called XYZ, operating in the motor insurance industry for the past seven years. As the insurance market becomes more competitive due to deregulation, XYZ faces mounting pressure to reduce losses, increase profits, and enhance customer service. To achieve these goals, XYZ's management has decided to invest in analytics-based decision making, particularly by developing a machine learning-based fraud detection process. By leveraging historical claims data and implementing a predictive model, XYZ aims to identify potentially fraudulent claims accurately. If the system predicts a claim to be fraudulent, it can be referred to the fraud prevention department for further investigation, ensuring that the company does not incur losses. On the other hand, if a claim is predicted to be legitimate, the claims department can promptly provide compensation, thereby improving customer service.

To kickstart this initiative, the fraud prevention department, in collaboration with the IT department, has collected a dataset named "Historical_Claims_Final." This dataset comprises various claim characteristics, and a target variable coded as 1 or 0, indicating whether a claim has been proven fraudulent or not. As a machine learning engineer hired by XYZ, the task is to develop a mathematical model using the historical claims data from May to September 2017. Once the model is developed, it can be applied to new claims (issued after October 1st, 2017) to predict their likelihood of being fraudulent. Claims deemed more likely to be fraudulent by the model can then be directed to the investigation department for further scrutiny. The characteristics of the new claims to be evaluated for October 2017 are stored in the "New_Claims_Final" dataset. In the subsequent sections, we will outline the steps involved in developing the predictive model and address the relevant questions pertaining to this case study.

Executive Summary

1. The purpose of this report is to develop and present a binary predictive classification tool for an insurance company that detects fraudulent claims and ensures the interests of the insurance company and its clients. The objective is to accurately predict fraudulent actions to minimize losses and provide better customer service, thereby gaining a competitive advantage in the market. The dataset analyzed consists of 3,000 claims from May 1, 2017, to August 30, 2017, including whether they were fraudulent or not and other inputs that provide crucial information about the claims' status.

The first step of the analysis involved conducting an Explanatory Analysis, which revealed important details about the dataset. It was observed that the proportion of fraudulent claims to non-fraudulent

claims converged to 30% over 70% of the available sample. Furthermore, claims involving vehicles with higher age or claim value 1.2 or more times higher than the vehicle value had an increased proportion of fraudulent efforts compared to general cases.

The second step involved forming four different models using distinct predictive modeling methods, namely Decision Tree, Logistic Regression, and Neural Network models. Hyperparameter tuning optimization was implemented in the Decision Tree model. After evaluating the models on unseen data partition, the Decision Tree Model was selected as the optimal solution for fraud detection, using Misclassification Error as the main performance metric.

Finally, the optimal model was deployed in a test dataset consisting of 200 claims to detect potential fraudulent claims. Based on the probabilistic threshold set by the company's acceptable profit ranges, 49.50% of the claims were classified as fraudulent, with Vehicle Age, Base Policy, and Fault of a third party in an accident case being the most important parameters for the claims classification.

2. The profit matrix provided represents the financial outcomes associated with investigating and compensating fraudulent and non-fraudulent cases. Based on the following profit matrix (Table 5) and the assumption to maintain the negative values since the company loses money in these cases and these cases should be avoided.

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

Table 7: Case's Profit Matrix.

From the above profit matrix, we can conclude to four cases:

Investigating Fraudulent Cases (1500$ profit): When a case is identified as fraudulent and an investigation is conducted and found out that the case is actual a fraudulent case, there is a profit of 1500$. This profit may come from preventing further financial losses, or recovering funds that were about to be lost due to the fraud.

Compensating Fraudulent Cases (-1500$ profit): If a case is wrongly classified as non-fraudulent and compensation is provided, resulting in a loss of 1500$, this indicates a financial loss. This loss origins from compensating cases that were, in fact, fraudulent.

Investigating Non-Fraudulent Cases (-200$ profit): In situations where a case is initially suspected as fraudulent but is eventually found to be non-fraudulent after investigation, there is a loss of 200$. This

loss could be associated with the costs incurred during the investigation process, such as personnel expenses, resources utilized, or any other associated expenses.

Compensating Non-Fraudulent Cases (0 profit): Finally, when a case is determined to be non-fraudulent and compensation is provided, there is no profit or loss.

In conclusion, the profit matrix for investigating fraudulent and non-fraudulent cases presents several outcomes with varying financial implications. It is essential to carefully manage and analyze cases to minimize losses and maximize profits. To effectively market the investigation services, the following strategies can be mentioned. For starters, a comprehensive fraud detection, which will emphasize to the ability to identify and investigate potential fraudulent cases with a high level of accuracy. Highlight the expertise and experience of the investigation team in employing advanced techniques and tools to uncover fraudulent activities, will help immediately our company. In addition, timely and efficient investigations are needed in order to highlight the importance of swift and efficient investigations to minimize financial losses.

Regarding the clients it will be needed to improve and establish transparent and trustworthy services by demonstrating transparency in the investigation process. Communicate the steps involved, the methodologies used, and the reporting mechanisms to keep clients informed throughout the investigation, will help to gain trust with the clients of the company. A final strategy it could be a proactive client education and prevention program, which will offer educational materials and resources to help clients understand common fraud schemes and develop preventive measures.

By adopting these marketing strategies and highlighting the value of the investigation services, clients will recognize the importance of proactive fraud detection, prevention, and accurate case classification.

3. To determine the minimum probability (cut-off point) for considering a claim as fraudulent and redirecting it for investigation, we need to analyze the profit matrix and find the threshold at which the expected profit from investigating a fraudulent case exceeds the expected profit from compensating a non-fraudulent case. To do so, we will calculate the expected profit for investigating a case and the expected profit for compensating a case:

Expected Profit for Investigating a Case =

*= (Probability of Fraudulent Case) * (Profit from Investigating) + (Probability of Non-Fraudulent Case) * (Profit from Investigating)*

*= p1 * 1500 - 200 * (1 - p1)*

$$= 1700 * p1 - 200 \quad (1)$$

Expected Profit for Compensating a Case =

*= (Probability of Fraudulent Case) * (Loss from Compensating) + (Probability of Non-Fraudulent Case) * (Loss from Compensating)*

$$= - 1500 * p1 + 0 * (1 - p1)$$

$$= - 1500 * p1 \quad (2)$$

To find the minimum probability (cut-off point), we need to set the expected profit for investigating a case greater than the expected profit for compensating a case:

Expected Profit for Investigating a Case > Expected Profit for Compensating a Case $\Leftrightarrow$ *combing (1), (2)*

$$1700 * p1 - 200 > - 1500 * p1 \Leftrightarrow$$

$$3200 * p1 > 200 \Leftrightarrow$$

$$p1 > \frac{1}{16} = \mathbf{0.0625}$$

Therefore, the minimum probability (cut-off point) for considering a claim as fraudulent and redirecting it for investigation is p1 > 0.0625 or 6.25%. Any claim with a probability of fraud higher than 6.25% should be considered fraudulent and undergo further investigation based on the given profit matrix.

4. Partitioning the historical data set into training and validation sets using the 70% - 30% rule of thumb is a common practice in machine learning and predictive modeling. This process is essential for model development, evaluation, and testing. To be more precise, the process of partitioning the historical dataset into training and validation sets is crucial for effective model development and evaluation. It ensures that the model learns the underlying features and patterns present in the data while being able to generalize well to unseen data samples in the future.

The **training dataset** is utilized to train the model by exposing it to a substantial portion of the available data. During training, the model learns the relationships, patterns, and features within the dataset, enabling it to make accurate predictions.

The **validation dataset** is used to evaluate and validate the model's performance during the training process. By assessing the model's predictions on the validation data, we can gauge its ability to generalize to new, unseen data samples.

The decision to choose the 70% - 30% rule was decided to prevent the overfitting phenomenon and maintain the generalization. More specifically, splitting the dataset into training and validation sets helps prevent overfitting, a phenomenon where the model becomes excessively tailored to the training data and fails to perform well on new data. By evaluating the model's performance on the validation set, we can identify signs of overfitting and take necessary measures to mitigate it, such as adjusting model complexity or regularization techniques. Furthermore, the purpose of training a model is to develop a generalizable representation of the underlying data patterns. The training process aims to expose the model to a diverse range of scenarios, allowing it to capture various features and relationships that contribute to accurate predictions. The validation set serves as an indicator of how well the model generalizes to new, unseen data samples, providing a reliable estimation of its performance.

The sampling in the data partition to be stratified is employed to ensure that the distribution of the target variable (fraudulent vs. non-fraudulent cases) remains balanced in both the training and validation sets. This sampling technique helps maintain the representative nature of the original dataset, ensuring that the model is exposed to an equitable proportion of each class during training and evaluation. It mitigates the risk of biased model performance due to imbalances in the target variable distribution. Overall, it is used to ensure that the same distribution of classes exists on both training and validation datasets. Thus, the training and validation datasets will contain observations corresponds to 70% for non-fraudulent and 30% for fraudulent observations.

Finally, to evaluate the performance of the model, the Misclassification Rate (Event) can be employed as the performance criterion. This metric quantifies the rate at which fraudulent cases are incorrectly classified by the model. In line with the previously calculated cut-off point of p1 > 0.0625 or 6.25% for identifying fraudulent claims, this threshold was inputted into the software during the model evaluation process. By comparing the model's predicted probabilities against the cut-off point, we can effectively measure the misclassification rate and assess the model's ability to accurately classify cases.

5. In Table 8, the number of missing values for each variable in the dataset is presented. However, it is evident that there are no missing values in the dataset that is currently available. This implies that all variables in the dataset have complete and valid data without any missing entries.

Variable Name	Missing	Label	Type	Role	Level
AccidentArea	0.0000	AccidentArea	Character	Input	Nominal
Age_OF_Vehicle	0.0000	Age_OF_Vehicle	Numeric	Input	Interval
AgentType	0.0000	AgentType	Character	Input	Ordinal
BasePolicy	0.0000	BasePolicy	Character	Input	Nominal
Claim_Value_Div_Vehicle_Value	0.0000	Claim_Value_Div_Vehicle_Value	Numeric	Input	Interval
Days_Accident_End_Of_Policy	0.0000	Days_Accident_End_Of_Policy	Character	Input	Ordinal
Days_Policy_Claim	0.0000	Days_Policy_Claim	Character	Input	Ordinal
DriverRating	0.0000	DriverRating	Numeric	Input	Ordinal
Fault	0.0000	Fault	Character	Input	Nominal
FraudFound_P	0.0000	FraudFound_P	Numeric	Target	Binary
Make	0.0000	Make	Character	Input	Nominal
NumberOfCars	0.0000	NumberOfCars	Character	Input	Ordinal
Partition_Indicator	0.0000	Partition_Indicator	Numeric	Input	Binary
PastNumberOfClaims	0.0000	PastNumberOfClaims	Character	Input	Ordinal
PoliceReportFiled	0.0000	PoliceReportFiled	Character	Input	Binary
PolicyID	0.0000	PolicyID	Character	ID	Nominal
PolicyType	0.0000	PolicyType	Character	Input	Nominal
Vehicle_Category	0.0000	Vehicle_Category	Character	Input	Nominal
Witness_Present	0.0000	Witness_Present	Character	Input	Binary

Table 8: Number of Missing Values by Variable in the Dataset.

Furthermore, the proportion of fraudulent and non-fraudulent is presented in Figure 8, where it can be observed that the fraudulent cases are the 30% of the total cases in the database (923 cases). While the non- fraudulent are the 70% of the total cases (2134 of the 3057 cases).

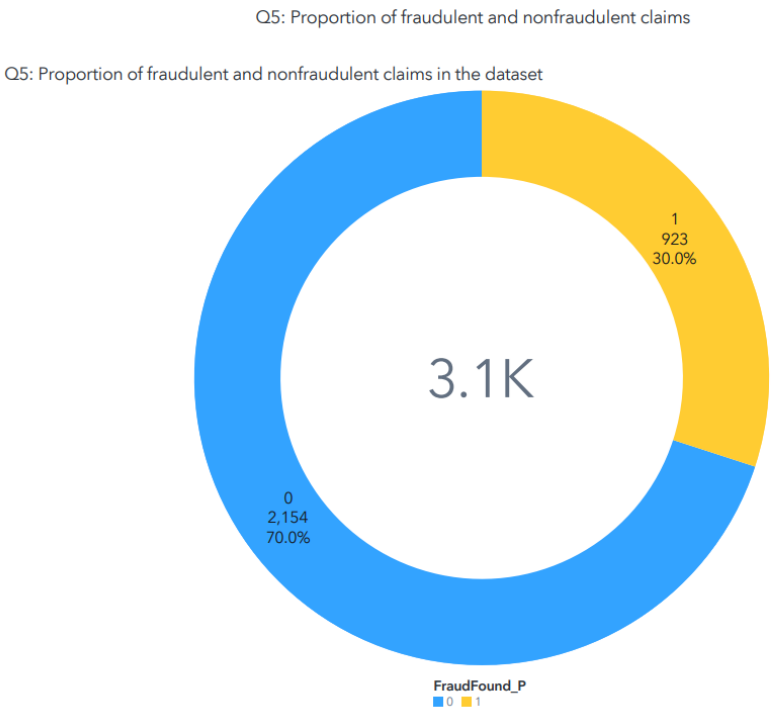


Figure 8: Proportion of fraudulent and non-fraudulent claims cases pie chart.

6. Based on an analysis of the proportion of fraudulent and non-fraudulent claims in our historical data set, it is evident that the dataset is a balanced one. Nevertheless, if the proportion of fraudulent and non-fraudulent claims in the historical data set were 10% - 90%, it would indicate an imbalanced dataset. Imbalanced datasets can present challenges for predictive modeling, as most machine learning algorithms are designed with the assumption of an equal number of examples for each class. This often leads to poor predictive performance, particularly for the minority class.

To address this issue, one approach would be to rebalance the proportions of each class in the dataset. In this case, we would need to select all the observations from the minority class (fraudulent claims) and a subset of the majority class (non-fraudulent claims) to achieve a desired balance. Based on the above, where the goal is to have a dataset with 30% fraudulent claims and 70% non-fraudulent claims, we would follow an empirical rule for undersampling. This would involve creating a separate sample that includes 100% of the fraudulent claims (10% of the original dataset) and an appropriate number of non-fraudulent claims to achieve the desired balance. The resulting separate sample would contain 30% fraudulent claims and 70% non-fraudulent claims.

Finally, it's important to note that the specific undersampling technique used may vary depending on the dataset and the problem at hand. The goal is to create a balanced dataset that allows the machine learning algorithms to learn effectively from both classes and improve predictive performance for the minority class (the fraudulent claims).

7. Based on the provided graph (Figure 9), the pie chart shows that **54.4% (466)** of the claims with Claim Value Divided by Vehicle Value greater than 120% are **fraudulent**, while **47.7% (391) are non-fraudulent**. This indicates that a slightly higher proportion of the claims with higher ratios (120%) of Claim Value to Vehicle Value fall under the fraudulent category compared to the non-fraudulent category. We need to note that only 857 out of the total of 3057 cases (about 28%) fulfil the desired condition.

The above suggests a potential relationship between the excessive Claim Value in relation to the Vehicle Value and the likelihood of fraud. However, it's important to note that this analysis only considers one factor, and there is only a slight difference between the two categories.

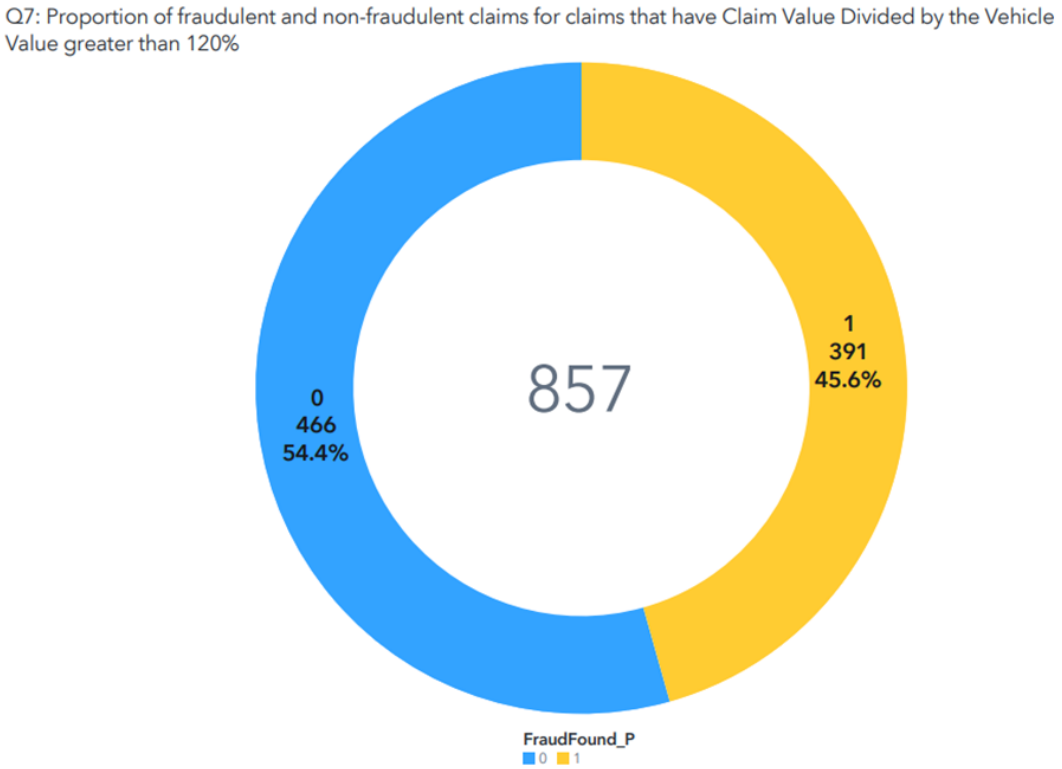


Figure 9: Proportion of fraudulent and non-fraudulent claims for claims that have Claim Value Divided by the Vehicle Value greater than 120% pie chart.

8. As it is presented in the following Figure 10, the average AgeOfVehicle for **fraudulent claims is 10.47** (notated as 1). This means that, on average, the vehicles involved in fraudulent claims have an age of 10.47 units (e.g., years, months, etc.). This information provides insights into the typical age range of vehicles associated with fraudulent claims.

The average AgeOfVehicle for **non-fraudulent claims is 7.56** (notated as 0). This means that, on average, the vehicles involved in non-fraudulent claims have an age of 7.56 units. This information indicates the average age range of vehicles for non-fraudulent claims.

More precisely and with respect to the target, these average values can help in understanding the relationship between the age of the vehicle and the likelihood of fraud. It suggests that, on average, fraudulent claims involve slightly older vehicles compared to non-fraudulent claims.

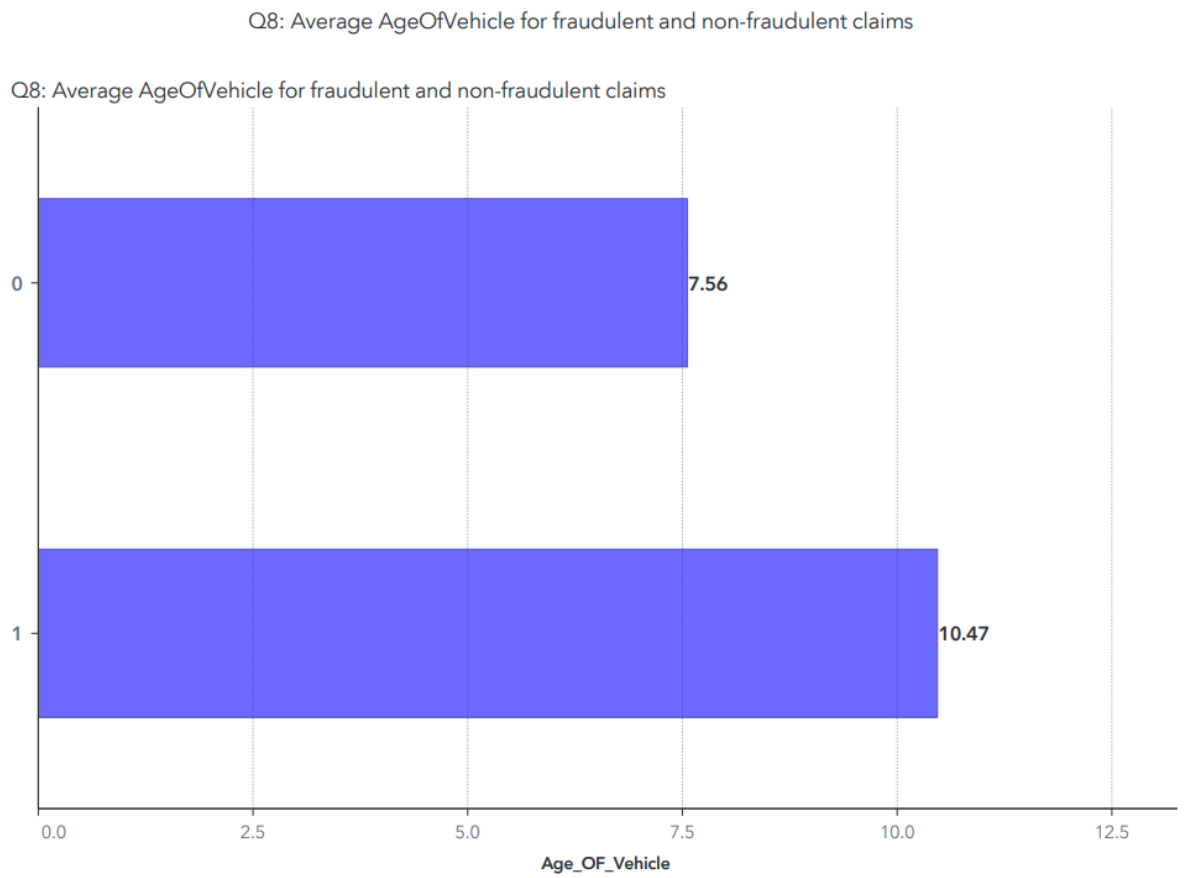


Figure 10: Average AgeOfVehicle for fraudulent and non-fraudulent claims barchart.

9. The most appropriate variable that led to the best possible split in the begging of the process is the Vehicle Age and the set split point is the vehicle age of 8 years. In order to select the best variables and the best split point of the selected variable in the Decision Trees nodes an algorithmic process called split search take place. In particular, the split search starts with selecting and input variable as candidate variable for the split, a value of the variable or the average category in case of categorical value is selected as a split point and two groups are generated the values below the point are called the left branch and the values of above the point are called the right branch. The two groups combined with the target outcome form a 2x2 contingency table where columns represent the left and right branch of the tree and the rows represent the two possible target outcomes of the model. A Pearson's chi-square distribution is used in order to quantify the independence of counts in table's columns. Large values for chi-square statistics suggest that the proportion of 0 and 1 in the left branch is different than the proportion on the right branch. A large difference in outcome proportions indicates

a good split. The Pearson chi-square statistic is then converted to a probability value, a p-value. The p-value indicates the probability of obtaining the observed value of the statistic assuming identical target proportions in each branch direction (left, right). For large datasets these p-values can be very close to zero and for this reason the quality of split is defined by *logworth* where $logworth = 1 - \log(chi - square\ p - value)$. In an iterative procedure all inputs are tested as possible split variables in each node and for each input variable all possible split points are tested as well. Finally, the input variable with the split point for which the *logworth* is maximized is selected as the final split choice for an internal node.

To sum up, the Vehicle Age is the root node selected split input with the split point equal to 8. The claims with vehicle age below 8 or with missing the Vehicle Age value are directed to the right branch and the claims with vehicle Age above or equal to 8 are directed to the left branch. In this case we see that the software presents a different convention than the general one and the observations with input values below the split point or missing values are directed to the right and the observations with input values above or equal to the split point are directed to the left branch of the Decision Tree produced by the software.

10. The Decision Tree produced by the system using the *Largest* method as a pruning method is called a Maximal Tree and it contains all the terminal nodes that are initially produced by the software during the training phase of the model. The Maximal Decision Tree is depicted in Figure 11 and contains **13 terminal leaves.**

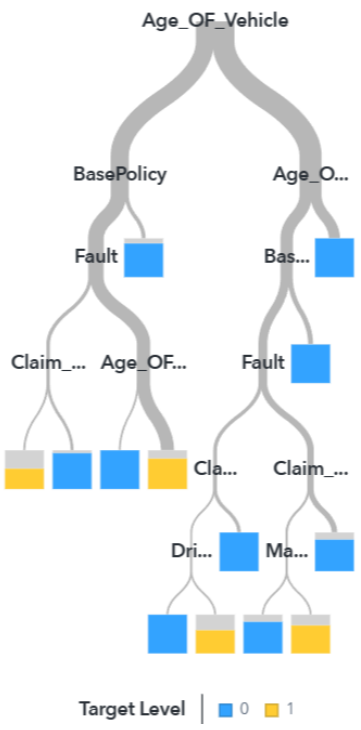


Figure 11: Maximal Decision Tree

In Figure 12 the Subtree Assessment Plot or Reduced Error Plot is provided. This plot shows the misclassification error curve for decision tree models with different number of terminal nodes. These trees are produced through the pruning process of the Maximal Tree that has been trained initially and presented in figure 11, by cutting further terminal nodes. The Subtree Assessment Plot or Reduced Error Plot contains all number of nodes from one node that represents the simple minimum tree with only the root node and the maximal number of nodes that represent the maximal tree, the most complex model created by the training process. For each subtree the misclassification error is calculated as a measure of the subtree's performance.

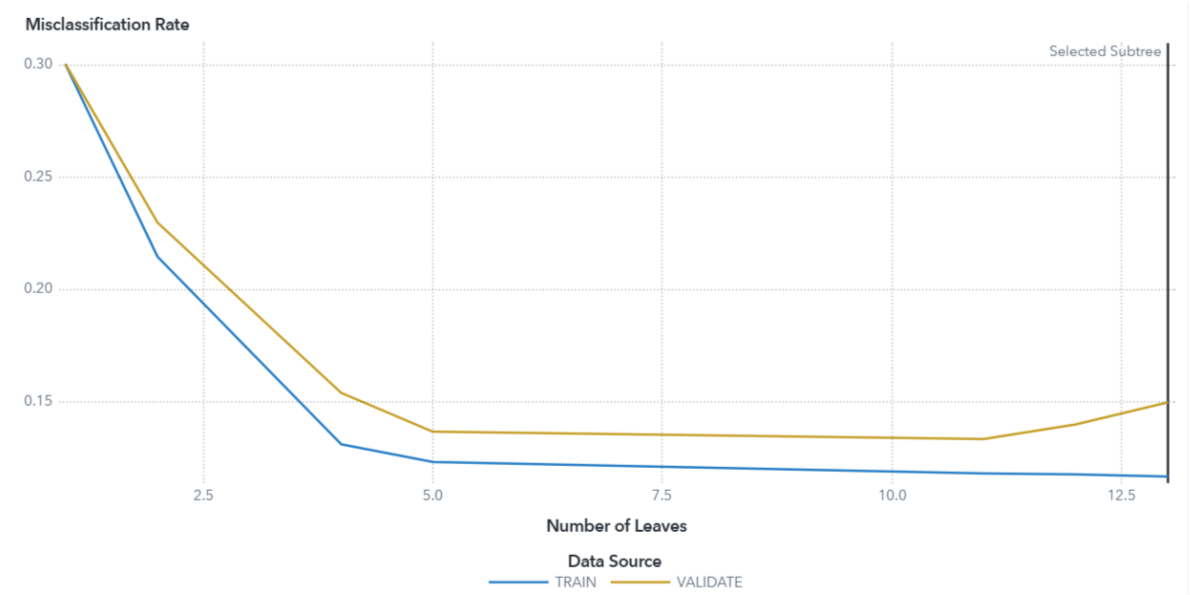
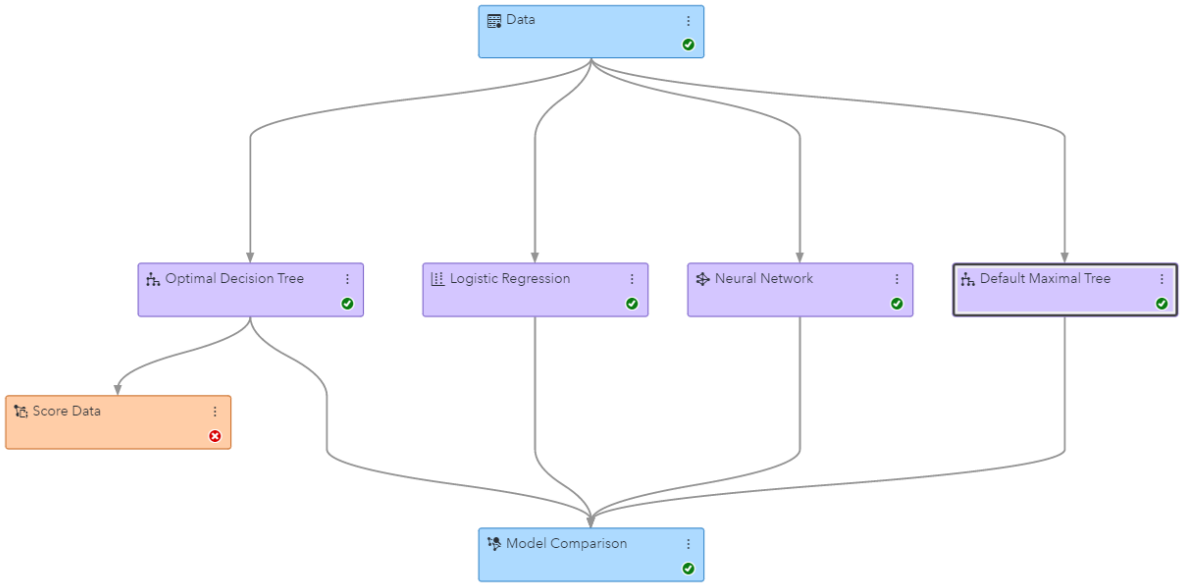


Figure 12: Subtree Assessment Plot for the Maximal Decision Tree Model.

Figure 12 shows the phenomenon of ***overfitting***. This phenomenon is presented by the blue line that expresses the error values for the different subtree values calculated for the prediction made using the training data. As the number of nodes augments and the model's complexity is augmented the misclassification error on the train data set is persistently reduced. This implies that the model presents high complexity and it captures all the training data set's noise especially after a certain point of complexity, number of nodes, is exceeded. In order to tackle this phenomenon, we use techniques such as the data partitioning so as to be able to test the model's performance in unseen data that have not been exposed during the training phase and proceed to the hyper tunning process during which we choose the optimal value of model's parameters and especially model's complexity (number of nodes) in a point where the misclassification error on the unseen data predictions is minimized. In this case of the Maximal Tree that is the most complex tree with the highest number of nodes and

most vulnerable to the overfitting phenomenon, has misclassification error on the train data set equal to 10% and the misclassification on the validation data set equal to 15%.

NOTE: In case we made the Maximal Tree with the default properties and tuning parameters and not with the suggested parameters (the one used for the Optimal Tree and the Largest method), the terminal nodes are **31** and not 13. We made a second pipeline to prove that, since we were not sure of what the exercise was asking regarding the Maximal Tree (Default Maximal Tree), to construct one with the initial default settings or one with the predefined settings of the Optimal Tree. A pipeline constructed with the default Maximal Tree properties is presented below.



11. The optimal tree is selected from the software, and it presents the tree with the optimal performance in terms of the misclassification error in out-of-sample data after the pruning process. In Figure 13 the Subtree Assessment Plot the selected optimal tree is depicted. The optimal tree includes 11 leaves and 11.7% misclassification error on train data and 13.3% misclassification error on validation data. The optimal number of nodes is selected depending on the validation error curve as the **11 nodes** value on the x-axis is the value of the validation data error curve minimization spot, indicating a model with the best complexity in terms of out-of-sample performance. After the 11 nodes point the validation error starts to rise and this indicates that more complex models would be overfitted to the train dataset and less generalized and consequently less capable of making predictions in unseen data. In addition, we clarify that the optimal tree presents smaller misclassification error on the validation data set (13,3%) in comparison with the maximal tree (MSE 15%).

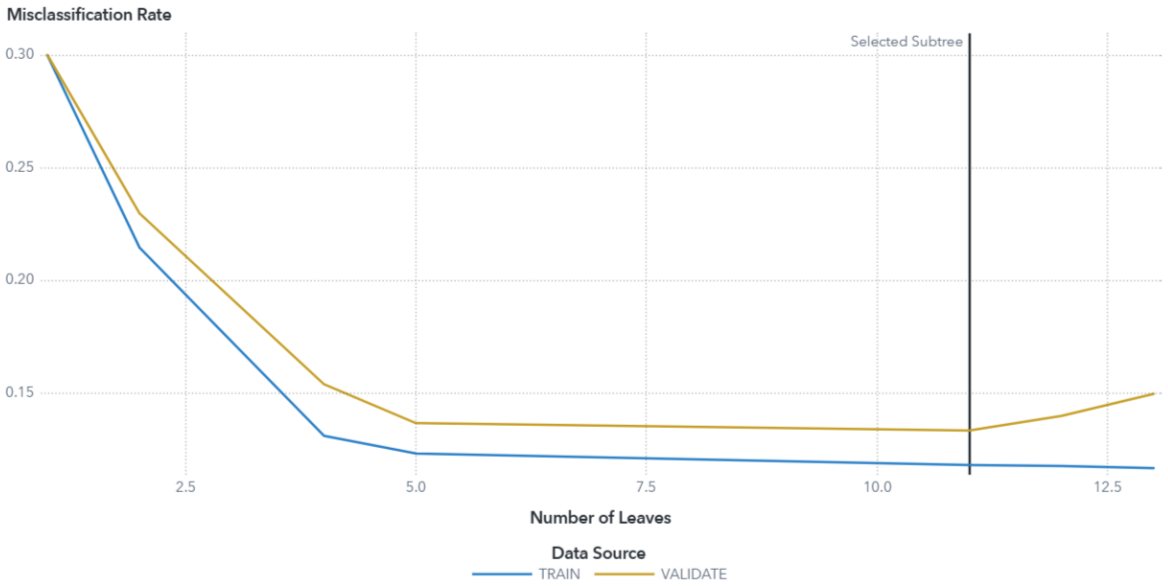


Figure 13: Subtree Assessment Plot of the Optimal Tree.

12. In the following Table 9 we have the technical interpretation of five terminal nodes of the optimal Decision Tree model.

Number of Terminal Node	Posterior Probabilities	Decision depending on the cut-off-point (P1=0.0625)	Rules
Node 4	P0=86.81% P1=13.19%	1	Vehicle Age >= 8, Base Policy = Liability
Node 6	P0=100.0% P1=0%	0	Vehicle Age < 7 or missing
Node 7	P0= 78.91% P1=21.09%	1	Vehicle Age >= 8, Base Policy = Collision, All Perils, Fault = Third Party
Node 10	P0=99.51% P1=0.49%	0	Vehicle Age >= 7, Base Policy = Liability
Node 12	P0=21.69% P1=78.31%	1	Vehicle Age >= 8, Base Policy = Collision, All Perils, Fault = Policy Holder, 0 <= Vehicle Age <15

Table 9: Optimal Decision Tree Model Technical Interpretation.

13. The decision tree model we have developed aims to help us distinguish between fraudulent and non-fraudulent cases in the insurance organization. The first example referring to older vehicles, 8 years or more, and are associated with a specific type of insurance coverage known as "Liability." Liability insurance primarily focuses on covering damages or injuries caused to third parties by the policyholder's vehicle. Then, the probability that the case will not be fraudulent case is 86.81%, while the probability that the case is fraudulent is 13.19%. Thus, since the probability of the case is over the cut-off point 6.25%, then the case will be categorized as fraudulent.

The second example is referring to cases that are vehicles less than 7 years old (or their age is not known - missing, there are not missing values in the dataset). In this case, the non-fraudulent case is 100%, while the probability that the case is fraudulent is 0%. Thus, since the probability of the case is less than the cut-off point 6.25%, then the case will be categorized as non-fraudulent.

In the third example, when a claim case meets the following conditions: the vehicle age is equal to or greater than 8 years, the base policy is Collision or All Perils, and the fault lies with a third party, it implies the occurrence of a specific type of claim situation. Then, the probability of being a non-fraudulent case is 78.91%, while the probability of being fraudulent is 21.09% and is characterized as fraudulent case, since the probability of the case is over the cut-off point of 6.25%.

The fourth example is referring to cases that contain older vehicles, 7 years or more, and are associated with a specific type of insurance coverage known as "Liability." Thus, the probability that the case will not be fraudulent case is 99.51%, while the probability that the case is fraudulent is 0.49%. Thus, since the probability of the case is less than the cut-off points of 6.25%, then the example case-study will be categorized as non-fraudulent case.

The final fifth example is referred to claim cases that containing vehicle, with age equal to or greater than 8 years but less than 15 years, the base policy is Collision or All Perils, and the fault lies with Policy Holder. Then, the probability of being a non-fraudulent case is 21.69%, while the probability of being fraudulent is 78.31% and is characterized as fraudulent case, since the probability to be fraudulent of the case is over the cut-off point of 6.25%.

14. In Figure 14 the Cumulative % Response of all models on the validation data set is presented. To be more specific, if the check the 20% of the most suspicious claims according to the probability that the Optimal Decision Tree model gives them to be fraudulent, the 75.59% of this 20% will be fraudulent claims.

If the check the 20% of the most suspicious claims according to the probability that the Maximal Decision Tree model gives them to be fraudulent, the 75.59% of this 20% will be fraudulent claims.

if the check the 20% of the most suspicious claims according to the probability that the Logistic Regression model gives them to be fraudulent the 66% of this 20% will be fraudulent claims.

if the check the 20% of the most suspicious claims according to the probability that the Neural Network model gives them to be fraudulent the 30% of this 20% will be fraudulent claims.

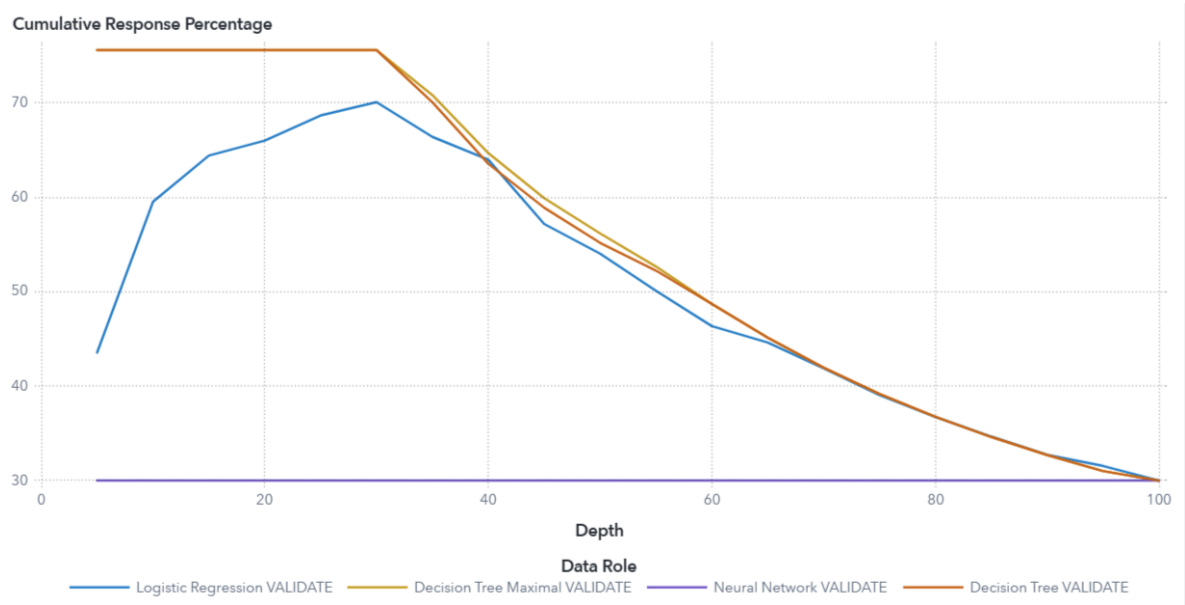


Figure 14: Comparative Cumulative Response Percentage on validation dataset.

15. In Figure 15 the % Response chart of all models on the validation data set is presented. This graph is constructed after sorting in descending order the claims depending on the posterior probability p1 of a claim to be fraudulent. Then the claims are grouped into 20 bins of equal size in terms of number of claims each bin contains. Therefore, the x-axis shows the bins that contain the observations and specifically, each bin which eventually contains 5% of the data set's observations in descending order in terms of which observations are more probable to be fraudulent claims according to a model.

To be more specific, if we check the fifth bin (20%-25%) of the most suspicious claims according to the probability that the Optimal Decision Tree gives them to be fraudulent, the 75.59% of the claims belonging to this bin will be fraudulent.

If we check the fifth bin (20%-25%) of the most suspicious claims according to the probability that the Maximal Decision Tree gives them to be fraudulent, the 75.59% of the claims belonging to this bin will be fraudulent.

If we check the fifth bin (20%-25%) of the most suspicious claims according to the probability that the logistic Regression model gives them to be fraudulent, the 79.80% of the claims belonging to this bin will be fraudulent.

If we check the fifth bin (20%-25%) of the most suspicious claims according to the probability that the Neural Network Model gives them to be fraudulent, the 30% of the claims belonging to this bin will be fraudulent.

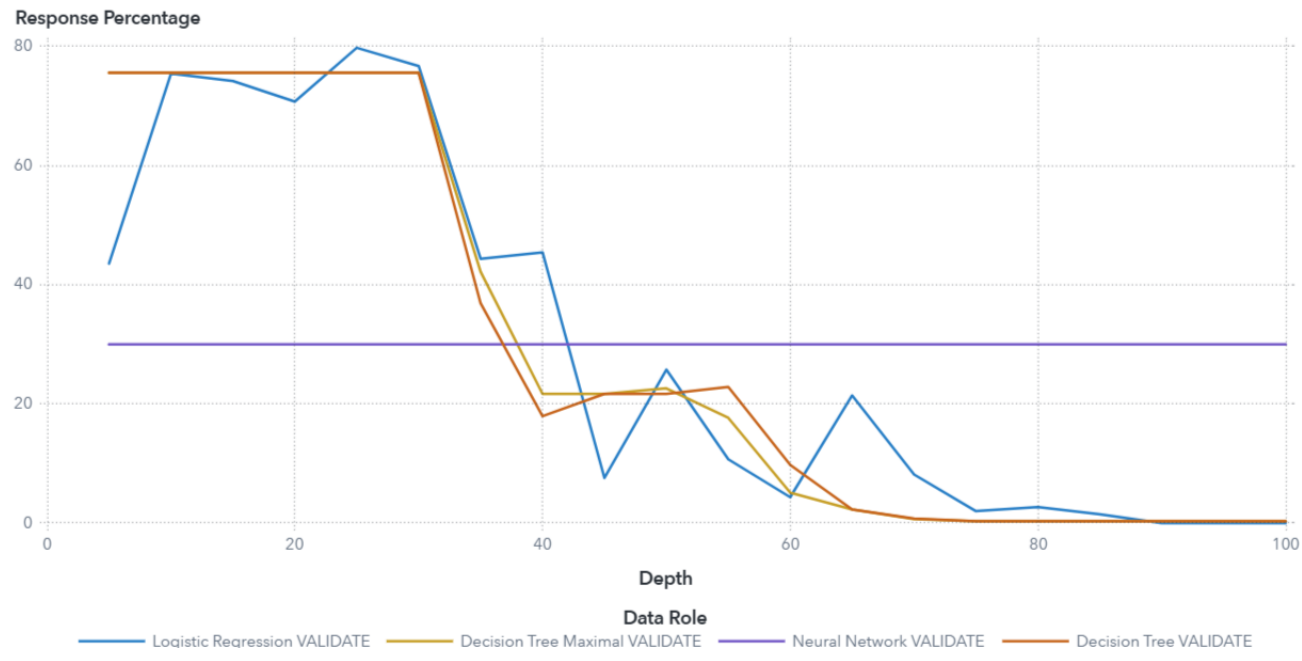


Figure 15: Comparative Response Percentage graph on validation dataset.

16. In figure 16 the Cumulative Lift chart of all models on the validation data set is presented**.** To be more specific, if we check the 20% of the most suspicious claims according to the probability that the Optimal Decision Tree model gives them to be fraudulent, we will capture 2.56 times more fraudulent claims than if we did the same job without a model at random.

If we check the 20% of the most suspicious claims according to the probability that the Maximal Decision Tree model gives them to be fraudulent, we will capture 2.56 times more fraudulent claims than if we did the same job without a model at random.

If we check the 20% of the most suspicious claims according to the probability that the Logistic Regression Tree model gives them to be fraudulent, we will capture 2.23 times more fraudulent claims than if we did the same job without a model at random.

If we check the 20% of the most suspicious claims according to the probability that the Neural Network model gives them to be fraudulent, we will capture 1.01 times more fraudulent claims than if we did the same job without a model at random.

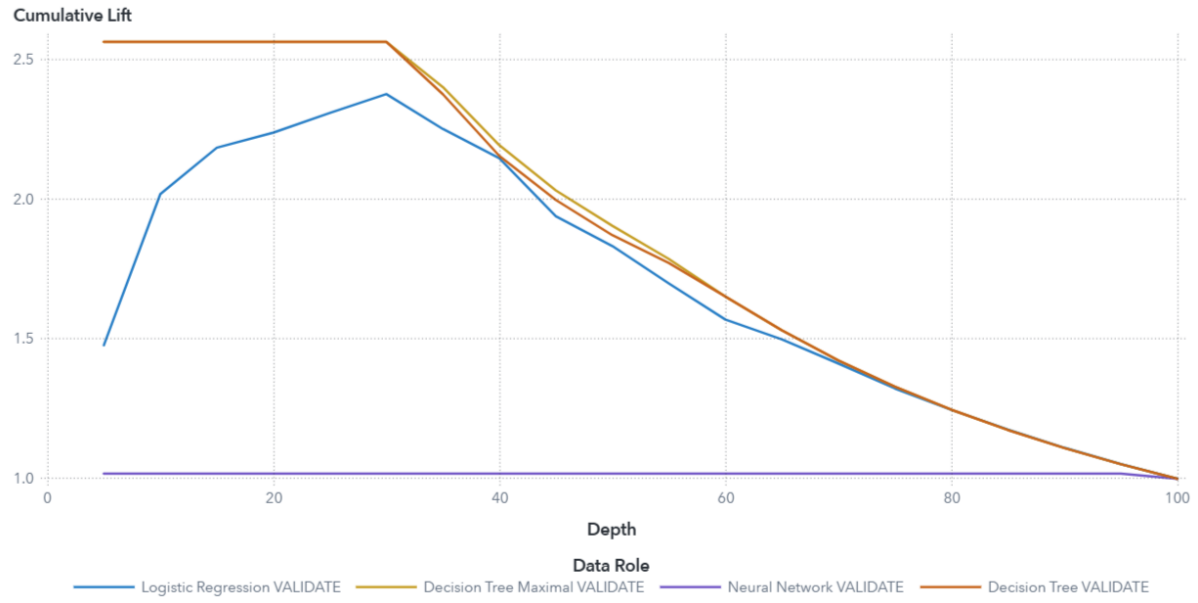


Figure 16: Comparative Cumulative Lift graph on validation dataset.

17. In figure 17 the Cumulative % Captured Response chart of all models on the validation data set is presented**.** To be more specific, if we check the 40% of the most suspicious claims according to the probability that the Optimal Decision Tree model gives them to be fraudulent, we will capture the 86.26% of all fraudulent claims of the whole validation data set.

If we check the 40% of the most suspicious claims according to the probability that the Maximal Decision Tree model gives them to be fraudulent, we will capture the 87.79% of all fraudulent claims of the whole validation data set.

If we check the 40% of the most suspicious claims according to the probability that the Logistic regression model gives them to be fraudulent, we will capture the 85.92% of all fraudulent claims of the whole validation data set.

If we check the 40% of the most suspicious claims according to the probability that the Neural Network model gives them to be fraudulent, we will capture the 40.73% of all fraudulent claims of the whole validation data set.

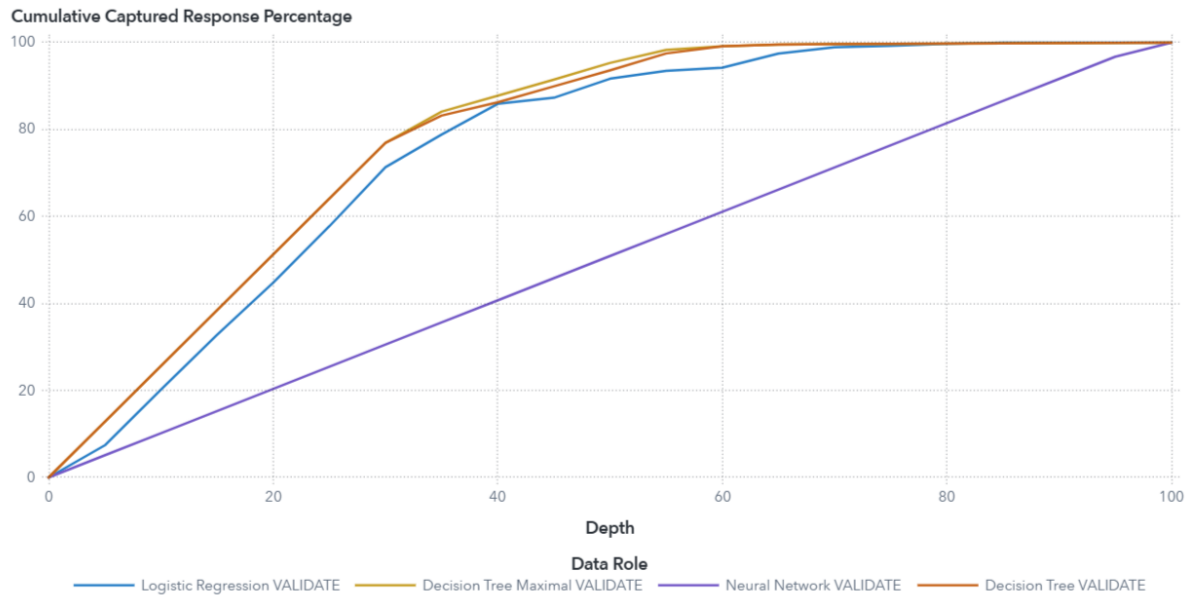


Figure 17: Comparative Cumulative Percentage Captured Response graph on validation dataset.

Finally, the Figure 18 displays the process flow constructed within the SAS software. The process flow diagram illustrates the sequence of activities and their interconnections in a formal manner, representing the logical flow of operations.

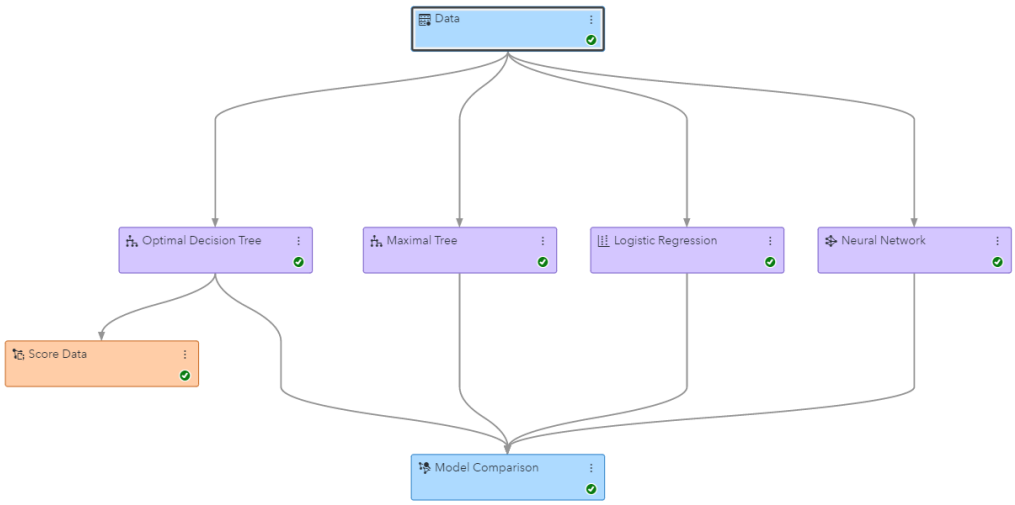


Figure 18: Process Flow Diagram.

18. To answer this question, we present the Figure 19, which exhibits the total count of claims within the "New_Claims_Final" dataset, categorized as either fraudulent or non-fraudulent. The dataset encompasses a total of **200** cases. Among these, the optimal model predicts that **101** cases are non-fraudulent, constituting **50.5%** of the dataset, while **99** cases are predicted to be fraudulent, accounting for **49.5%** of the dataset.

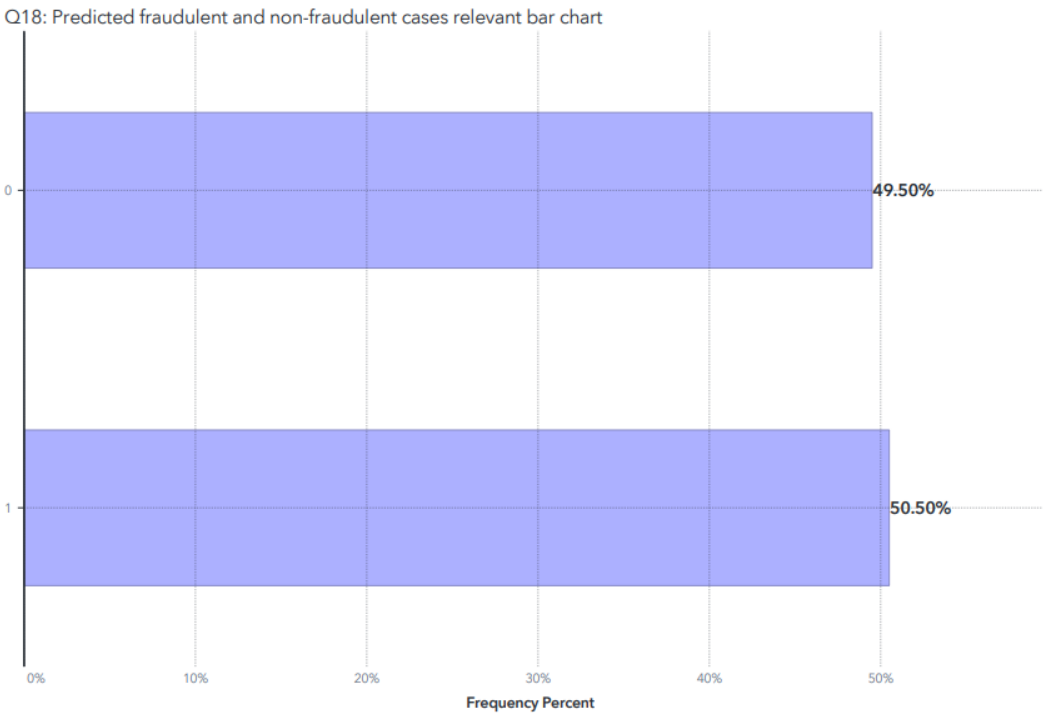


Figure 19: Predicted fraudulent and non-fraudulent cases relevant bar chart.

19. Within the context of the provided information, the claim exhibiting the highest probability of being fraudulent is assigned a value of 78.3%. This indicates a significantly elevated likelihood of fraudulent activity associated with that particular claim. Conversely, the claim with the smallest probability of being fraudulent is assigned a value of 0%, implying an absence of any suspected fraudulent behavior in relation to that claim. The column used to find out these probabilities is the p1 or as mentioned on the dataset, **Probability for FraudFound_P =1**.

20. To determine why the software assigns a value of 1 (fraudulent) or 0 (non-fraudulent) to the claims with PolicyID = 15 and PolicyID = 107, we need to examine the relevant column in the score dataset. To be more precise, the software assigns a value of 0 (non-fraudulent) to the claim with PolicyID = 15 and a value of 1 (fraudulent) to the claim with PolicyID = 107. This assignment is based on the column "Predicted for FraudFound_P" in the score dataset.

For PolicyID = 15, the assigned value of 0 indicates that the predicted probability of fraud (0.0048780488) is below the cutoff point (0.0625). Therefore, the software predicts this claim to be non-fraudulent. In the case of PolicyID = 107, the assigned value of 1 signifies that the predicted probability of fraud (0.7830609212) surpasses the cutoff point (0.0625). Hence, the software predicts this claim to be fraudulent. The software's assignment of 1 or 0 to these claims is based on comparing the predicted probabilities with the specified cutoff point to determine the fraudulent/non-fraudulent classification.

Q20: PolicyID= 15 and PolicyID=107 Crosstable

Predicted for FraudFound_P ▲	0	1
PolicyID ▲	**Predicted: FraudFound_P=1**	**Predicted: FraudFound_P=1**
15	0.0048780488	—
107	—	0.7830609212

Table 10: PolicyID= 15 and PolicyID=107 Crosstable.