Business Analytics Practicum I

Assignment

Name

Helga Ruci (f2822219)

Lefteris Souflas (f2822217)

Case Study 1

Executive Summary

The analysis conducted focused on identifying cross-selling opportunities for an online bookstore called "Buy-books-on-line.com". This online store specializes in books related to science and information technology. The category of books that the analysis is focused on is related to "Business Analytics", which is popular among customers. A Market Basket Analysis (MBA) was performed on a dataset of 19,805 past sales transactions to identify cross-selling opportunities. To prepare for the analysis, a figure was constructed that presented the number of book sales by title, which helped to identify the best and worst performing titles. The analysis consisted of two parts. The first part identified the strongest relationships between four specific books. For each book, two other books were identified to be advertised to customers who have already purchased or searched for them. In the second part of the analysis, the set of three books that were most frequently purchased together by customers was identified. A recommendation was made that customers who purchased one of these books be targeted with advertising proposals for the other two. SAS Visual Data Mining and Machine Learning on SAS Viya software was used to conduct the analysis.

1) The Executive Summary is presented on the previous page.

2) Illustrated in Figure 1, there is a record of sales in units for each book. The y-axis is arranged in descending order according to the sales of the 56 books, while the x-axis represents the count of sales. Each bar represents a book, with the exact number of sales of that book adjacent to it. Notably, the book with the highest sales is "Data Science and Business Analytics" which sold 1596 copies, while the one with the lowest sales is "Managerial Analytics" with only 152 copies sold.

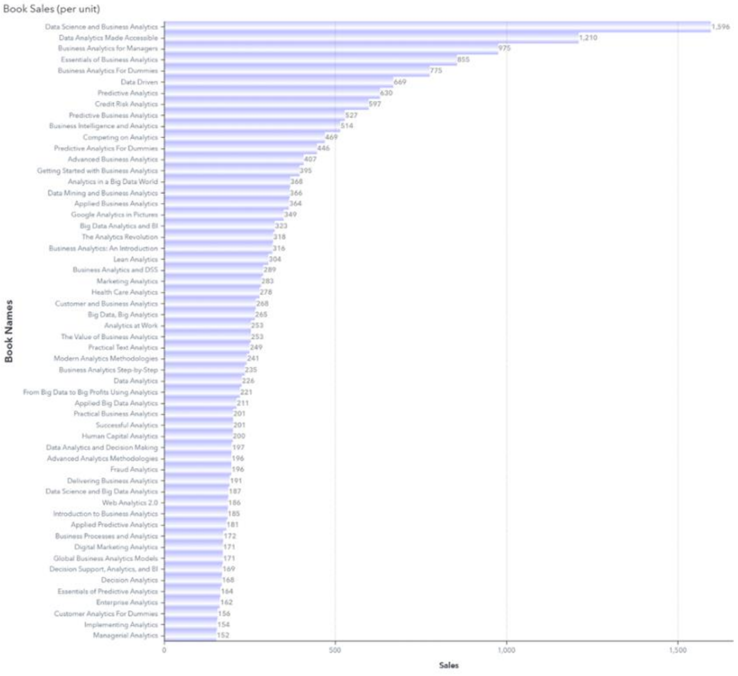


Figure 1 - Number of Sales per Book

3) To identify the most relevant books to advertise to customers who have purchased or shown interest in Managerial Analytics, Implementing Analytics, Customer Analytics for Dummies, and Enterprise Analytics, the lift metric was utilized and filtered the Market Basket Analysis table for association rules that contained each of the four books on the left side of the rule. Then the resulting association rules were sorted by their lift value in descending order and identified the top two books for each of the four books of interest. These recommended books are the ones that are most likely to be purchased along with the initial four books by customers.

Specifically, two books were identified which should be proposed to every user that is interested in buying one of these analytics books.

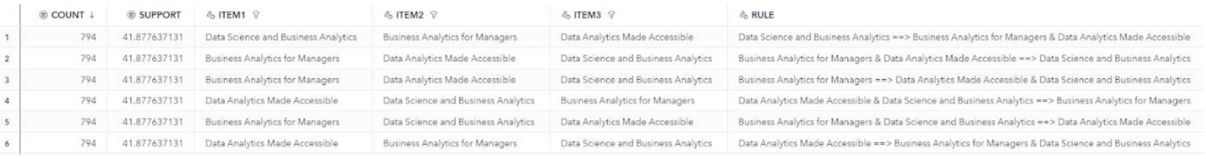
For customers who have shown interest in “Managerial Analytics”, a good recommendation would be to consider purchasing both “Web Analytics 2.0” and “Implementing Analytics”. This is due to the high lift value of 11.47, indicating that if a customer purchases the “Managerial Analytics” book, they are 11.47 times more likely to also purchase the other two books, compared to a random customer who has not made a purchase yet.

Similarly, for customers who have shown interest in “Implementing Analytics”, it would be beneficial to suggest “Managerial Analytics” and “Data Science and Big Data Analytics”. The lift value for this rule is 11.33, indicating that customers who purchase “Implementing Analytics” are 11.33 times more likely to purchase the other two books compared to a random customer.

For customers interested in “Customer Analytics for Dummies”, it would be wise to recommend “Enterprise Analytics” and “Decision Analytics”. The lift value for this rule is 11.19, indicating that customers who purchase “Customer Analytics for Dummies” are 11.19 times more likely to purchase the other two books compared to a random customer.

Lastly, customers searching for “Enterprise Analytics” may benefit from considering the books “Managerial Analytics” and “Customer Analytics for Dummies”. The lift value for this rule is 11.07, indicating that customers who purchase “Enterprise Analytics” are 11.07 times more likely to purchase the other two books compared to a random customer.

4) When limiting the number of items in a rule to 3, the 3 books that appear to be frequently purchased together are “Business Analytics for Managers”, “Data Analytics Made Accessible”, and “Data Science and Business Analytics”. These books have been observed in 794 transactions, indicating that they are frequently bought together by customers. In Figure 2, all the possible rules that are created with these three items are recorded.



	COUNT ↓	SUPPORT	ITEM1	ITEM2	ITEM3	RULE
1	794	41.877637131	Data Science and Business Analytics	Business Analytics for Managers	Data Analytics Made Accessible	Data Science and Business Analytics ==> Business Analytics for Managers & Data Analytics Made Accessible
2	794	41.877637131	Business Analytics for Managers	Data Analytics Made Accessible	Data Science and Business Analytics	Business Analytics for Managers & Data Analytics Made Accessible ==> Data Science and Business Analytics
3	794	41.877637131	Business Analytics for Managers	Data Analytics Made Accessible	Data Science and Business Analytics	Business Analytics for Managers ==> Data Analytics Made Accessible & Data Science and Business Analytics
4	794	41.877637131	Data Analytics Made Accessible	Data Science and Business Analytics	Business Analytics for Managers	Data Analytics Made Accessible & Data Science and Business Analytics ==> Business Analytics for Managers
5	794	41.877637131	Business Analytics for Managers	Data Science and Business Analytics	Data Analytics Made Accessible	Business Analytics for Managers & Data Science and Business Analytics ==> Data Analytics Made Accessible
6	794	41.877637131	Data Analytics Made Accessible	Business Analytics for Managers	Data Science and Business Analytics	Data Analytics Made Accessible ==> Business Analytics for Managers & Data Science and Business Analytics

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

The support metric is a measure of how often a set of items appear together in transactions. In this case, the itemset consists of the books “Data Science and Business Analytics”, “Data Analytics Made Accessible” and “Business Analytics for Managers”. The support metric, which is defined as the probability of intersections of those 3 books, for this item set is 41.87%, which means that these three books were bought together in 41.87% of all customer purchases. This calculation is done by dividing the number of purchases that contain all three books (794) by the total number of purchases (1896). A high support metric indicates that the itemset occurs frequently, and rules with a high support are preferred because they are likely to be relevant to many transactions. For example, a rule with a support of 50% would be applicable to half of all transactions, while a rule with a support of 5% would only be applicable to a small subset of transactions. Therefore, the support metric is an important factor to consider when analyzing transaction data and identifying patterns in customer behavior.

Case Study 2

Executive Summary

Sports-OnLine.com, an online retailer specializing in sports clothing and shoes, aims to leverage the electronic data collected over the previous years to gain deeper insights into the market and is willing to conduct an RFM (Recency – Frequency – Monetary) analysis, to identify and prioritize valuable customers by segmenting them. As a result of the RFM analysis four customer groups were identified. "Bad Customers", who, on average, have not made a purchase for an extended period compared to the overall customer base and make less frequent purchases and have a lower total expenditure. "First Time Customers", who on average have made more recent but less frequent purchases compared to the average customer and their total spending is also lower. "Churners", who on average have not made a purchase for an extended period, but their purchasing frequency is higher than the average customer and have also a higher total expenditure. "Good Customers", who on average, have made a more recent purchase, engage in more frequent transactions, and have a higher total spending compared to the average customer. Based on these customer groups several strategies were proposed including gathering customer feedback, enhancing services, boosting sales, offering personalized offers, and introducing loyalty programs with additional rewards.

Technical Analysis

The analysis began by importing the data to the software. Firstly, some plots are created to visualize the data and understand them better.

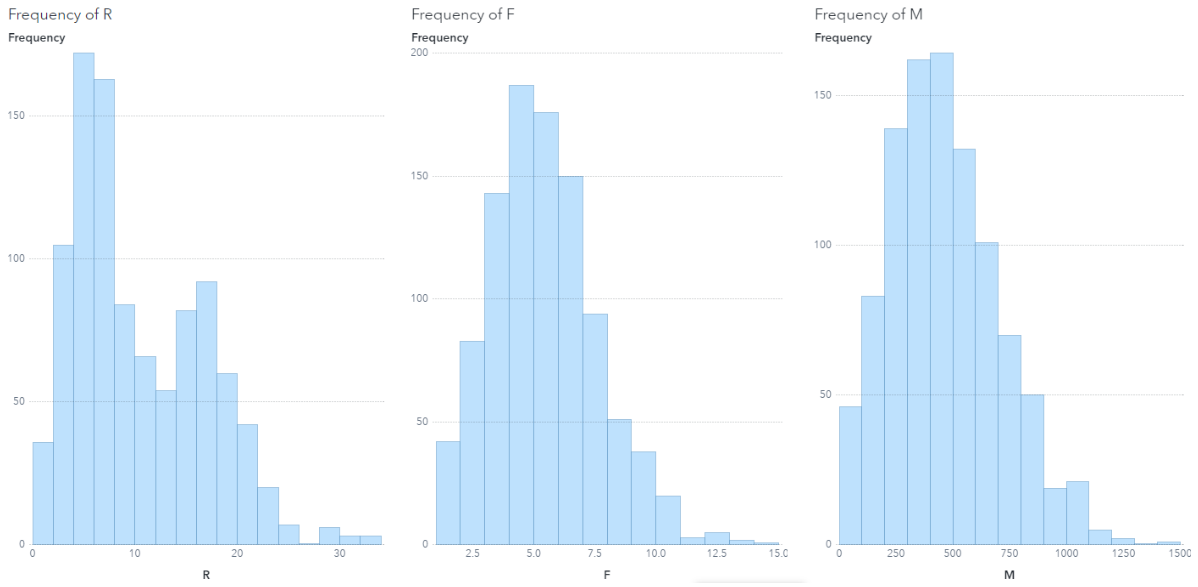


Figure 3 - Histograms for R, F, M

The histograms presented in Figure 3 provide insights into customer behavior based on recency, frequency, and monetary metrics. The Recency histogram reveals a bimodal distribution, indicating that the customer base can be divided into two groups: those who make a purchase within the last 2-7 months and those who take 15 or more months to make a purchase. The Frequency histogram demonstrates that the majority of the customers have made 3-7 purchases from the company. The Monetary histogram displays that most customers have spent between 200-800€ on the company's products or services. These metrics are essential for understanding customer purchasing behavior and can be used to inform marketing and sales strategies. Furthermore, based on the above histograms, it can be observed that the distributions of R, F and M exhibit a positive skewness, so it is needed to consider a log transformation later.

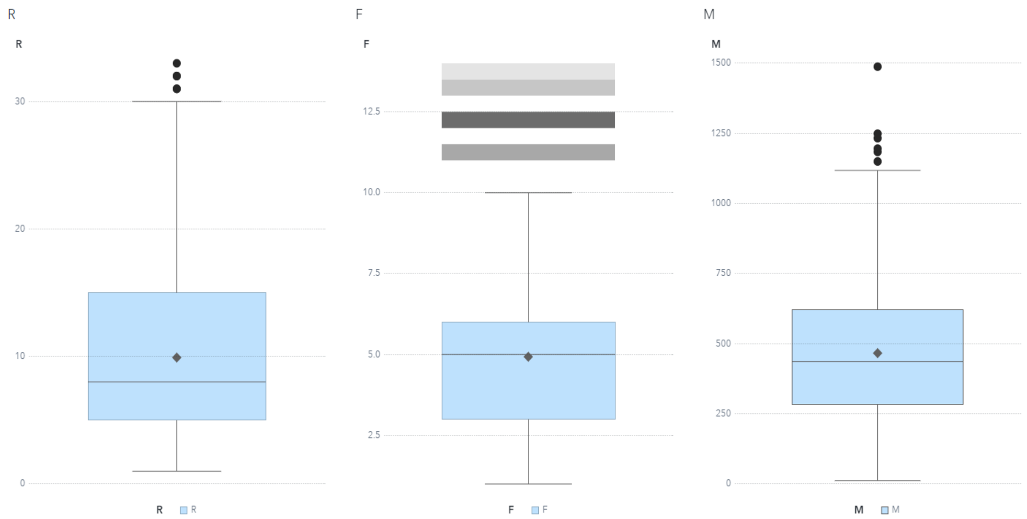
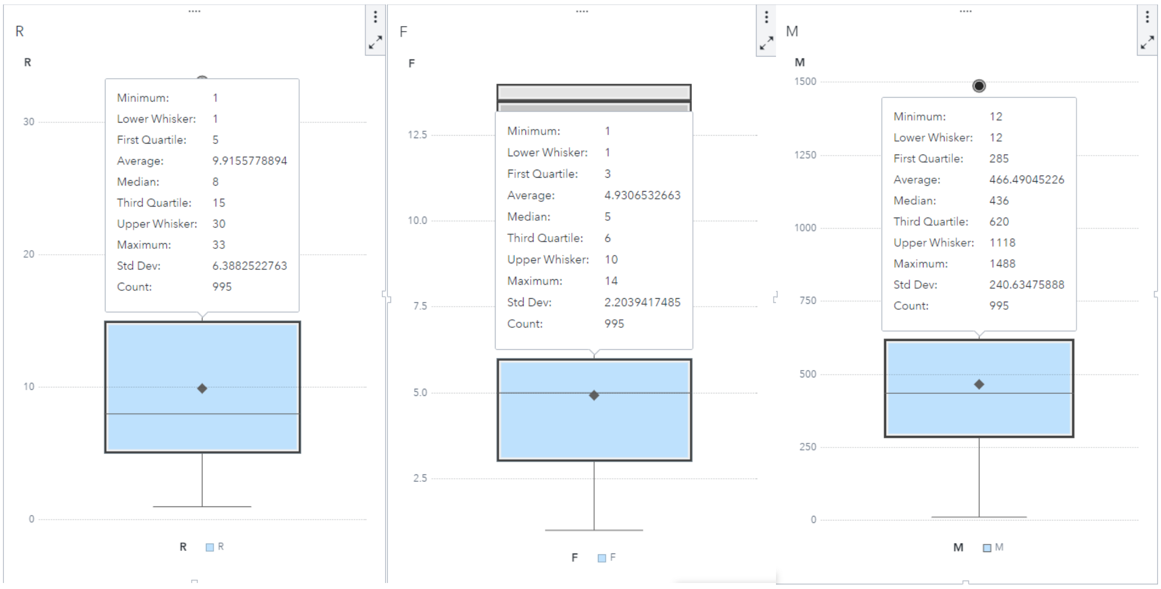


Figure 4 - Boxplot for R, F, M variables

Next, an analysis is conducted for detecting any possible outliers in the data. Three boxplots for the three variables R, F, and M, are produced, as shown in Figure 4 and Figure 5. The boxplots revealed the presence of outliers, which are represented by dots located above the whiskers in the plots. Hence, it is required to remove the outliers of the input variables and it is derived that for the variable R the outlier values are greater than 30, for the variable F the outlier values are greater than 10 and for the variable M the outlier values are greater than 1118.

The median for variable R is 8 months, for variable F is 5 purchases, and for variable M is 435€.



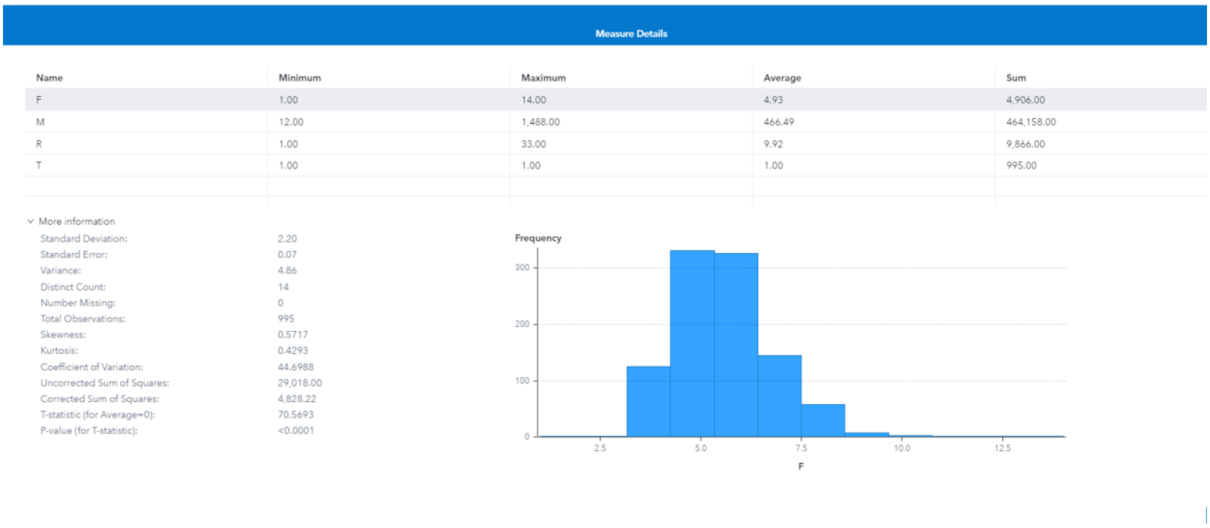
Minimum:	1
Lower Whisker:	1
First Quartile:	5
Average:	9.9155778894
Median:	8
Third Quartile:	15
Upper Whisker:	30
Maximum:	33
Std Dev:	6.3882522763
Count:	995

Minimum:	1
Lower Whisker:	1
First Quartile:	3
Average:	4.9306532663
Median:	5
Third Quartile:	6
Upper Whisker:	10
Maximum:	14
Std Dev:	2.2039417485
Count:	995

Minimum:	12
Lower Whisker:	12
First Quartile:	285
Average:	466.49045226
Median:	436
Third Quartile:	620
Upper Whisker:	1118
Maximum:	1488
Std Dev:	240.63475888
Count:	995

Figure 5 - Boxplot for R, F, M variable with details

From the insights gained from the graphs, occurred some preprocessing steps to improve the usefulness of the data. This involved filtering out the outlier observations, by excluding values that deviated more than three standard deviations from the mean and applying a log-transformation to the input variables, because of the skewness of their distributions. Finally, the missing values were extracted from the dataset. The initial data set consisted of 995 customers. After filtering out 19 outlier observations, 976 customers left to be used for the analysis. Finally, a check was conducted for any missing value. Since there were not any missing values, as can be observed in the following figure, it was not required for any imputation to be made.

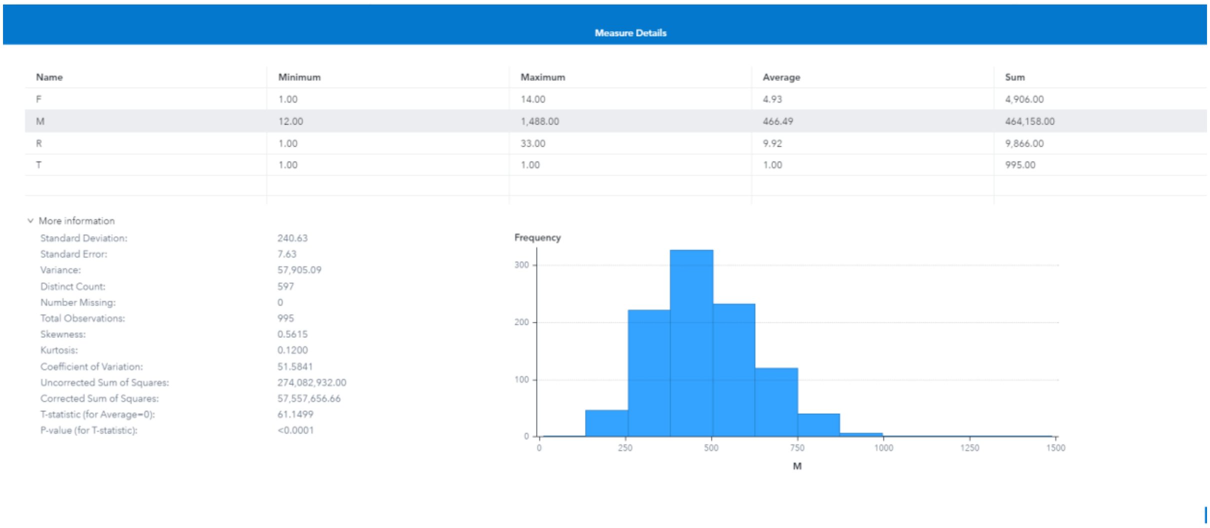


Measure Details

Name	Minimum	Maximum	Average	Sum
F	1.00	14.00	4.93	4,906.00
M	12.00	1,488.00	466.49	464,158.00
R	1.00	33.00	9.92	9,866.00
T	1.00	1.00	1.00	995.00

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
T-statistic (for Average=0):	70.5693
P-value (for T-statistic):	<0.0001

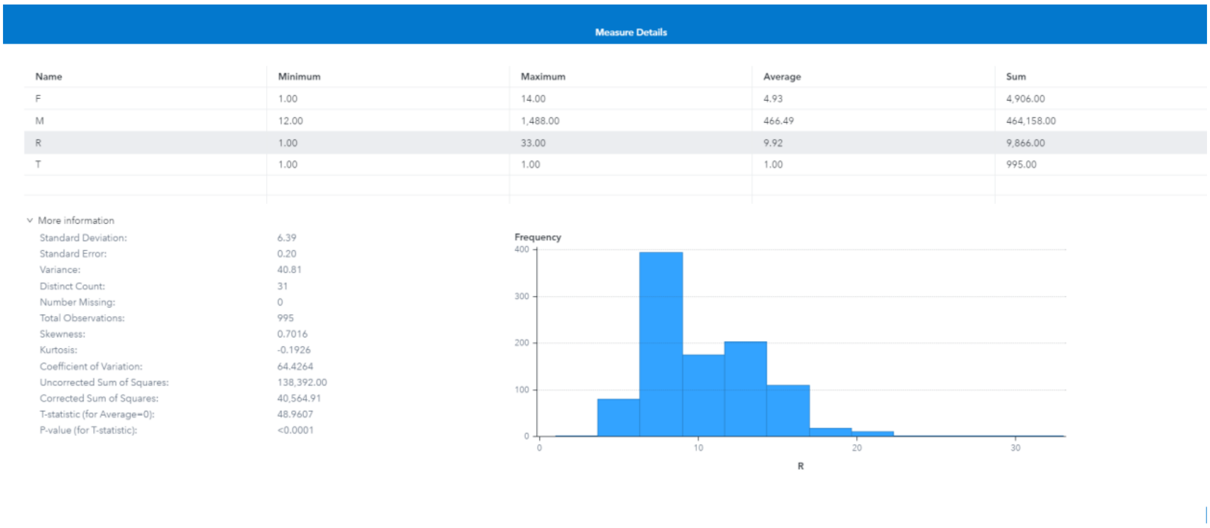


Measure Details

Name	Minimum	Maximum	Average	Sum
F	1.00	14.00	4.93	4,906.00
M	12.00	1,488.00	466.49	464,158.00
R	1.00	33.00	9.92	9,866.00
T	1.00	1.00	1.00	995.00

More information

Standard Deviation:	240.63
Standard Error:	7.63
Variance:	57,905.09
Distinct Count:	597
Number Missing:	0
Total Observations:	995
Skewness:	0.5615
Kurtosis:	0.1200
Coefficient of Variation:	51.5841
Uncorrected Sum of Squares:	274,082,932.00
Corrected Sum of Squares:	57,557,656.66
T-statistic (for Average=0):	61.1499
P-value (for T-statistic):	<0.0001



Measure Details

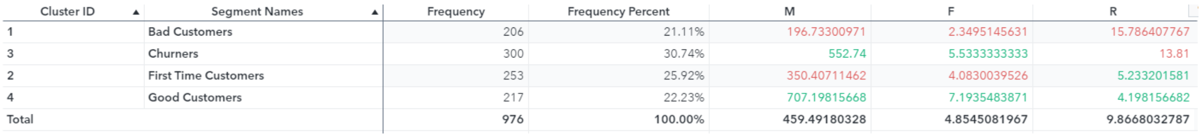
Name	Minimum	Maximum	Average	Sum
F	1.00	14.00	4.93	4,906.00
M	12.00	1,488.00	466.49	464,158.00
R	1.00	33.00	9.92	9,866.00
T	1.00	1.00	1.00	995.00

More information

Standard Deviation:	6.39
Standard Error:	0.20
Variance:	40.81
Distinct Count:	31
Number Missing:	0
Total Observations:	995
Skewness:	0.7016
Kurtosis:	-0.1926
Coefficient of Variation:	64.4264
Uncorrected Sum of Squares:	138,392.00
Corrected Sum of Squares:	40,564.91
T-statistic (for Average=0):	48.9607
P-value (for T-statistic):	<0.0001

Figure 6 - View measure details for variables R, F, M

The final dataset was, then, used to perform the RFM analysis, which involved clustering customers using k-means and Euclidean distance. The analysis was set to identify four distinct customer groups, each with unique characteristics. As depicted in Figure 7, the algorithm identified four distinct segments in the dataset.



Cluster ID ▲	Segment Names ▲	Frequency	Frequency Percent	M	F	R
1	Bad Customers	206	21.11%	196.73300971	2.3495145631	15.786407767
3	Churners	300	30.74%	552.74	5.5333333333	13.81
2	First Time Customers	253	25.92%	350.40711462	4.0830039526	5.233201581
4	Good Customers	217	22.23%	707.19815668	7.1935483871	4.198156682
Total		976	100.00%	459.49180328	4.8545081967	9.8668032787

Figure 7 - RFM Analysis

Figure 7 displays details of 4 clusters and their respective values for Recency, Frequency, and Monetary. This information was gathered to improve our understanding of our customers and gain insights into each cluster. With this knowledge, the Business Analytics department can tailor marketing strategies and promotions to meet the needs of each cluster. The values in the table represent the average characteristics of the ideal customer in each cluster, and there is a row at the bottom of the table that provides the mean characteristics of a typical customer.

The column labeled Frequency indicates how many times customers in each cluster have purchased from the company the last five years.

The Monetary column shows the amount of money the ideal customer in each segment has spent on his transactions.

The Recency column denotes the number of months that have elapsed since the last purchase made by the typical customer in each cluster.

To categorize each group, considering the R, F, M values of their ideal customer, the values were compared to the average values of individuals in the dataset. The green and red color scheme was used to distinguish whether each value in the R, F, M columns is above or below the mean value in the last row. This allowed the naming of each cluster to take place based on their customers' performance in these three main categories. In terms of the recency (R) category, we aim for a lower value than the mean recency. Conversely, a higher value for both the frequency (F) and monetary (M) categories is desired.

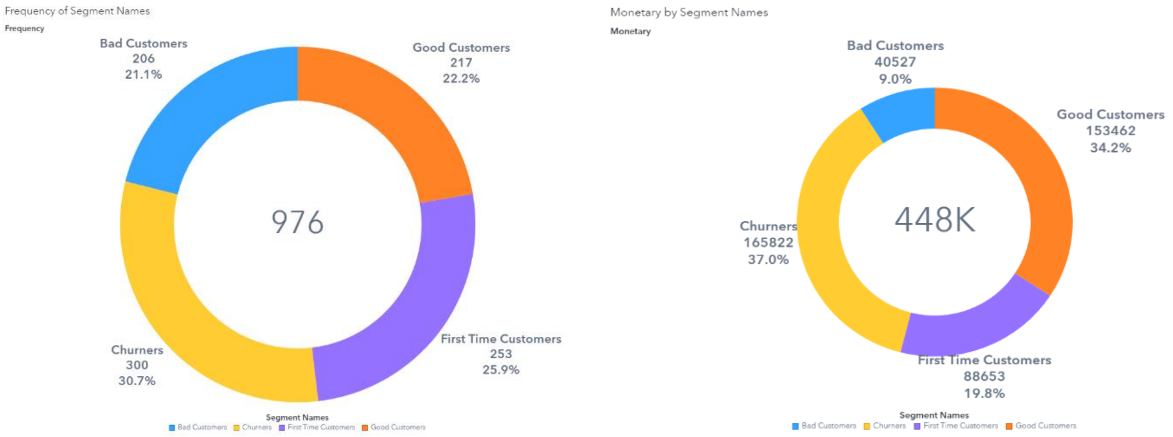


Figure 8 - RFM Cluster Analysis Graph

The first group was named "Bad Customers" and represented 21.1% of the total customer base, with 206 customers. The typical customer in this group had not made a purchase in almost 16 months, made an average of slightly more than 2 purchases, and spent an average of 196.7€. This group accounted for 9% of the total amount spent on purchases during the study period.

The second group was named "Churners" and represented 30.74% of the total customer base, with 300 customers. The typical customer in this group had not made a purchase in almost 14 months, made an average of 5 and a half purchases, and spent an average of 552.74€. This group accounted for 37% of the total amount spent on purchases during the study period.

The third group was named "First Time Customers" and represented 25.92% of the total customer base, with 253 customers. The typical customer in this group had not made a purchase in approximately 5 months, made an average of nearly 4 purchases, and spent an average of 350.4€. This group accounted for 19.8% of the total amount spent on purchases during the study period.

The fourth group was named "Good Customers" and represented 22.2% of the total customer base, with 217 customers. The typical customer in this group had not made a purchase in 4 months, made an average of around 7 purchases, and spent an average of 707.19€. This group accounted for 34.2% of the total amount spent on purchases during the study period.

The analysis showed that the "First Time Customers" group was the most profitable for Sports-OnLine.com, as these customers made more frequent purchases and spent more money per transaction than customers in the other groups.

After analyzing the information from both charts, there are some suggestions for some business actions based on our findings. These are the proposals for each segment:

Bad Customers – Low Value Customers: They spent a total of 40527€, representing 9% of our total income. Our aim is to turn them into active customers. To achieve this, we could contact them and gather feedback on their opinion of our company. There could also be offers for them and some promotions that may incentivize them to engage with our company.

First Time Customers – New Customers: They spent a total of 88653€, which is 19.8% of our total income. Our goal is to prevent them from becoming "one-time buyers." If they become regular buyers, the frequency of purchases would increase, resulting in more revenue for our company. To achieve this, frequent or special offers can be made, such as a small discount on specific products that may interest them.

Churners – Inactive Customers: They spent a total of 165822€, which corresponds to 37% of our total income. Our priority is to reactivate them as good customers, as they used to be. To achieve this, there could be frequent or special discounts for them on each selling item for the next three purchases in a specific period. Also, contact, not only via email, but also via call, could be useful to gather honest feedback about our company. This way, the company can evaluate why they stopped being good buyers and make them feel valued.

Good Customers – High Value Customers: They spent a total of 153462€, representing 34.2% of our total income. Undoubtedly, this segment, along with the churners, is the most important for our company, as they account for 71.2% of our total income. Our priority is to make them feel valued and keep them loyal. To achieve this, the company could send them frequent emails with items that are on sale and offer them a standard small discount at checkout. Also, enhanced cross-selling and upselling activities can increase their purchase frequency.

Case Study 3

Executive Summary

The problem under consideration tackled at this case study was the development of a mathematical model to categorize claims issued at XYZ insurance company to being fraudulent or not. The company provided historical data to aid the machine learning engineer create the model, which would suggest directing for further investigation claims that are more likely to be fraudulent and compensating the legitimate ones without delays, thus providing fewer losses and better customer service. The problem was addressed by the development of various supervised learning models, which were evaluated and compared through a profit matrix, provided by the management team of the fraud prevention department of the company and various statistics and graphs. SAS Visual Data Mining and Machine Learning on SAS Viya software was used for building and assessing various models and SAS Visual Analytics was utilized for exploring and visualizing data and supporting decisions taken. The best model of the forementioned comparison was a Decision Tree, which was finally applied to new "unseen" data provided by the company to evaluate the model developed.

1) The Executive Summary is presented on the previous page.

2) In the case study under consideration there are four outcome/action combinations: Investigate a truly fraudulent claim; investigate a falsely identified as fraudulent claim (i.e., non-fraudulent); compensate a truly non-fraudulent (i.e., legitimate) claim; compensate a fraudulent claim. The management team of the fraud prevention department informed us that it creates a profit of 1500 monetary units to investigate a claim and find out that it truly is fraudulent, because it will save the company from compensating it. If a claim is decided to be investigated and from the investigation is derived that it is a legitimate claim, the company will have a negative profit (cost) of 200 monetary units, because of the delay in compensating it. If it is decided to compensate a truly legitimate claim, then there will be no profit or cost for the company, but if it is decided to firstly compensate a claim that it is then found to be fraudulent, then the company will have a cost of 1500 monetary units.

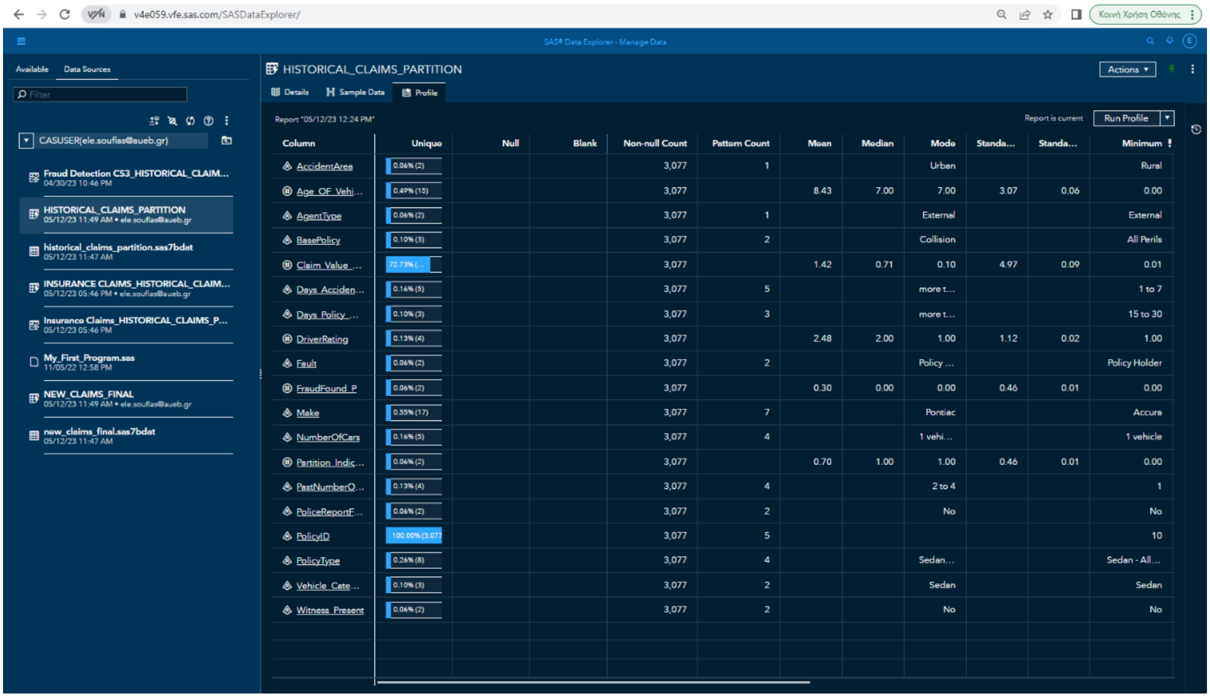
3) Let p be the probability of a claim being fraudulent and 1-p the probability of it being non-fraudulent (legitimate). So, if p is above a certain threshold (cut-off point), it should be considered as fraudulent, and it should be redirected for investigation. Based on the profit matrix provided by the management team of the fraud prevention department:

$$(p \times 1500) + (1 - p) \times (-200) > p \times (-1500) + (1 - p) \times 0$$
$$1500 \times p + 200 \times p - 200 > -1500 \times p$$
$$3200 \times p > 200$$
$$p > \frac{1}{16} = 6.25\%$$

4) The historical data set is partitioned to training and validation using the 70% - 30% rule of thumb. In typical machine learning tasks, data is divided into different sets (partitions): some data for training the model and some data for evaluating the model. Fitting a model to data requires searching through the space of possible models. Constructing a model with good generalization requires choosing the right complexity. Selecting model complexity involves a trade-off between bias and variance. An insufficiently complex model might not be flexible enough, which can lead to *underfitting*—that is, systematically missing the signal (high bias). An overly complex model might be too flexible, which can lead to *overfitting*—that is, accommodating nuances of the random noise in the sample (high variance). A model with the right amount of flexibility gives the best generalization. The first partition of the data, the *training set*, is used to build models. Usually, for each modeling algorithm, a series of models is constructed, and the models increase in their complexity. The idea behind constructing a series of models is that some will be too simple (underfit) and others will be too complex (overfit). Each of the models is assessed for its performance on the second partition of the data, the *validation set*. In this way, the validation data is used to “optimize complexity” of the model and find the sweet spot between being underfit and being overfit. Validation data is used to fine tune the models built on training data and determine whether additional training is required.

The sampling in the data partition is stratified. Stratified Sampling is a sampling method that reduces the sampling error in cases where the population can be partitioned into subgroups. We perform Stratified Sampling by dividing the population into homogeneous subgroups, called strata, and then applying Simple Random Sampling within each subgroup. As a result, the test set is representative of the population, since the percentage of each stratum is preserved. The strata should be disjointed; therefore, every element within the population must belong to only one stratum. When the input data has a partition variable or a class target (or both), the sample is stratified using them. Otherwise, a simple random sample is used.

5) There are no missing values in the variables of the data set. Proof of it is the following screenshot from SAS Data Explorer (Figure 9), where each variable has a non-null count of 3,077.


v4e059.vfe.sas.com/SASDataExplorer/
Κοινή Χρήση Οθόνης
SAS® Data Explorer - Manage Data
Available
Data Sources
Filter
CASUSER(ele.souflas@aueb.gr)
Fraud Detection CS3_HISTORICAL_CLAIM...
HISTORICAL_CLAIMS_PARTITION
historical_claims_partition.sas7bdat
INSURANCE CLAIMS_HISTORICAL_CLAIM...
Insurance Claims_HISTORICAL_CLAIMS_P...
My_First_Program.sas
NEW_CLAIMS_FINAL
new_claims_final.sas7bdat
HISTORICAL_CLAIMS_PARTITION
Actions
Details
Sample Data
Profile
Report "05/12/23 12:24 PM"
Report is current
Run Profile

Column	Unique	Null	Blank	Non-null Count	Pattern Count	Mean	Median	Mode	Standa...	Standa...	Minimum
AccidentArea	0.06% (2)			3,077	1			Urban			Rural
Age_OF_Vehi...	0.49% (15)			3,077		8.43	7.00	7.00	3.07	0.06	0.00
AgentType	0.06% (2)			3,077	1			External			External
BasePolicy	0.10% (3)			3,077	2			Collision			All Perils
Claim_Value_...	72.73% (...			3,077		1.42	0.71	0.10	4.97	0.09	0.01
Days_Acciden...	0.16% (5)			3,077	5			more t...			1 to 7
Days_Policy_...	0.10% (3)			3,077	3			more t...			15 to 30
DriverRating	0.13% (4)			3,077		2.48	2.00	1.00	1.12	0.02	1.00
Fault	0.06% (2)			3,077	2			Policy ...			Policy Holder
FraudFound_P	0.06% (2)			3,077		0.30	0.00	0.00	0.46	0.01	0.00
Make	0.55% (17)			3,077	7			Pontiac			Accura
NumberOfCars	0.16% (5)			3,077	4			1 vehi...			1 vehicle
Partition_Indic...	0.06% (2)			3,077		0.70	1.00	1.00	0.46	0.01	0.00
PastNumberO...	0.13% (4)			3,077	4			2 to 4			1
PoliceReportF...	0.06% (2)			3,077	2			No			No
PolicyID	100.00% (3,077			3,077	5						10
PolicyType	0.26% (8)			3,077	4			Sedan...			Sedan - All...
Vehicle_Cate...	0.10% (3)			3,077	2			Sedan			Sedan
Witness_Present	0.06% (2)			3,077	2			No			No

Figure 9 - Historical Claims Screenshot (SAS Data Explorer)

The proportion of fraudulent (1) and non-fraudulent (0) claims in the data set is 30% – 70% respectively, as shown in Figure 10.

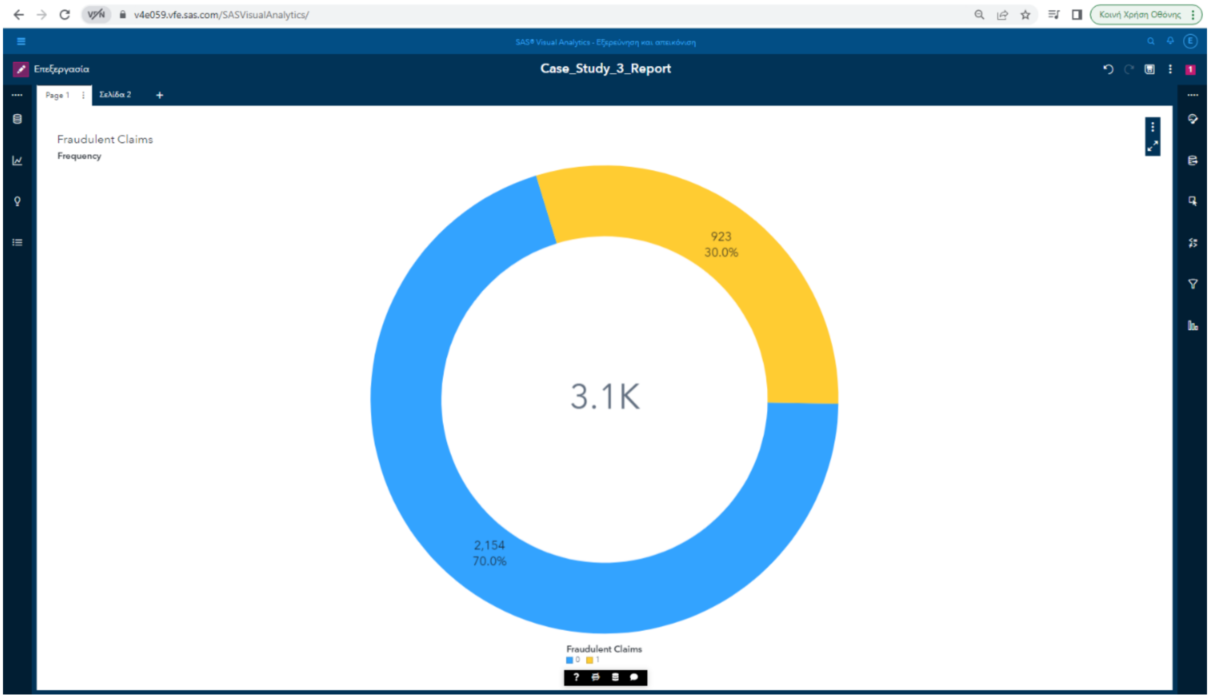


Figure 10 - Pie Chart of the target variable's frequency

6) The proportion of fraudulent and non-fraudulent claims in the historical data set is 30% - 70%. If this proportion was 10% - 90%, then we would have, what is called, imbalanced data, because the target of interest is characterized as a rare event in relation to the total number of samples. Fitting a model to such data without

accounting for the extreme imbalance in the occurrence of the event will provide a model that is extremely accurate at telling absolutely nothing of value. One of the widely adopted class imbalance techniques for dealing with highly unbalanced datasets is called resampling. It consists of removing samples from the majority class (under-sampling the 90% non-fraudulent claims) and/or adding more examples from the minority class (over-sampling the 10% fraudulent claims). The simplest implementation of over-sampling is to duplicate random records from the fraudulent claims, which can cause overfishing. In under-sampling, the simplest technique involves removing random records from the majority class (the 90% non-fraudulent claims), which can cause a loss of information. Under-sampling is provided by the Graphical User Interface of SAS, whereas over-sampling can be implemented through SAS Code e.g., Synthetic Minority Oversampling Technique (SMOTE) algorithm.

7) We can observe that for those claims that have Claim Value Divided by Vehicle Value greater than 120%, the proportion of fraudulent claims increased from 30% of the whole dataset to 46.3%. So, a significant variable for capturing fraudulent claims is to measure if the claim's value is bigger than 1.2 times the vehicle's value.

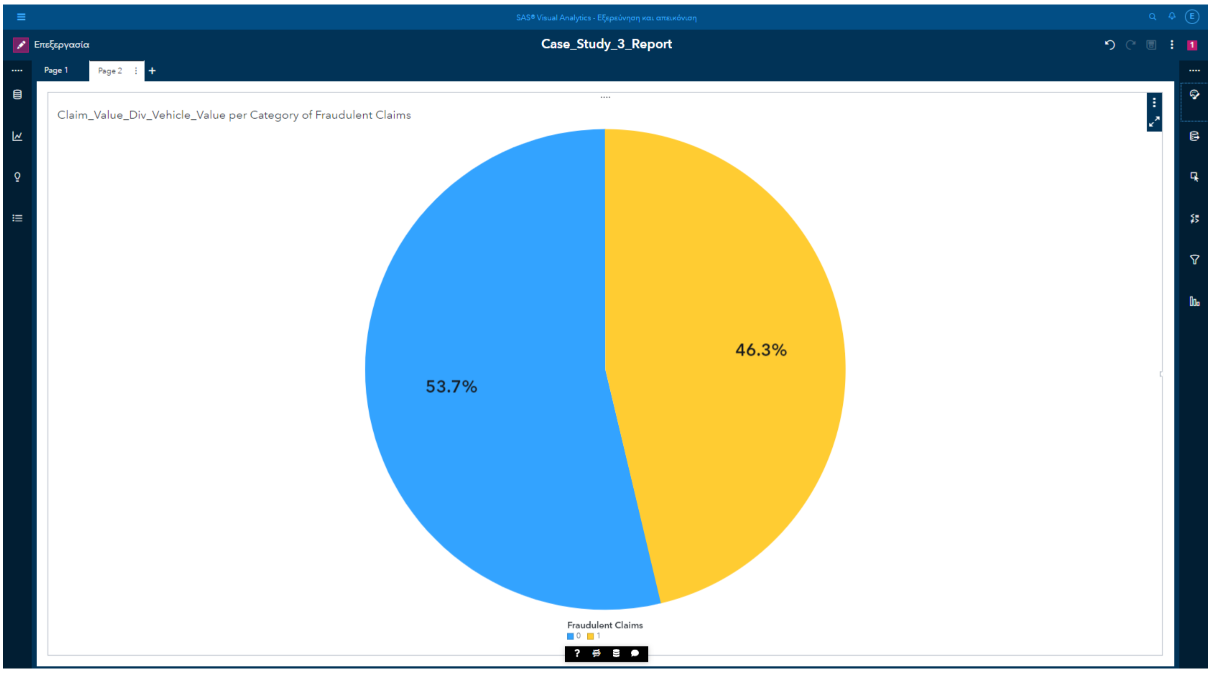

SAS® Visual Analytics - Εξερεύνηση και απεικόνιση
Επεξεργασία
Case_Study_3_Report
Page 1
Page 2
Claim_Value_Div_Vehicle_Value per Category of Fraudulent Claims
53.7%
46.3%
Fraudulent Claims
0
1

Figure 11 - Pie Chart of Claim Value Divided by Vehicle Value per Category of Fraudulent Claims

8) The average age of vehicle for fraudulent claims is 10 years, whereas for legitimate (non-fraudulent) claims is 7.6 years. Thus, we can derive that if a claim is for a car that is 10 years old or older, it is more possible to be a fraudulent claim than a legitimate one.

9) The variable used for the first split is the age of the vehicle. To select useful inputs for splitting a node in the tree, a split-search algorithm is used. The split-search starts by selecting an input for partitioning the available training data and generating two groups. The groups, combined with the target outcomes, form a 2x2 contingency table with columns specifying branch direction (left or right) and rows specifying the target values (0 or 1). A Pearson's Chi-Squared Statistic is then used to quantify the independence of counts in the table's columns, which is then converted to a probability p-value. The p-value indicates the likelihood of obtaining the observed value of the statistic assuming identical target proportions in each branch direction. Because these p-values can be very close to zero for large data sets, the quality of a split is reported by $logworth = -\log(chi\_squared\ p\_value)$. Then, Bonferroni adjustments are applied to the $logworth$

calculations for an input to penalize those with many split points in order to enable a fairer comparison of inputs with many and few levels. The best split for an input is the split that yields the highest $logworth$ and after the best split for every input is determined, the tree algorithm compares each best split 's corresponding $logworth$. The split with the highest adjusted $logworth$ is deemed best, which in the case of the first split of our data set is the age of the vehicle.

The optimal split for each of the next inputs considered is the one that maximizes the $logworth$ function for that input. At each tree split there exists a condition based upon an input variable. If the condition is satisfied (True), then the cases are directed to the left node, whereas if the condition is not satisfied (False), the cases are directed to the right node.

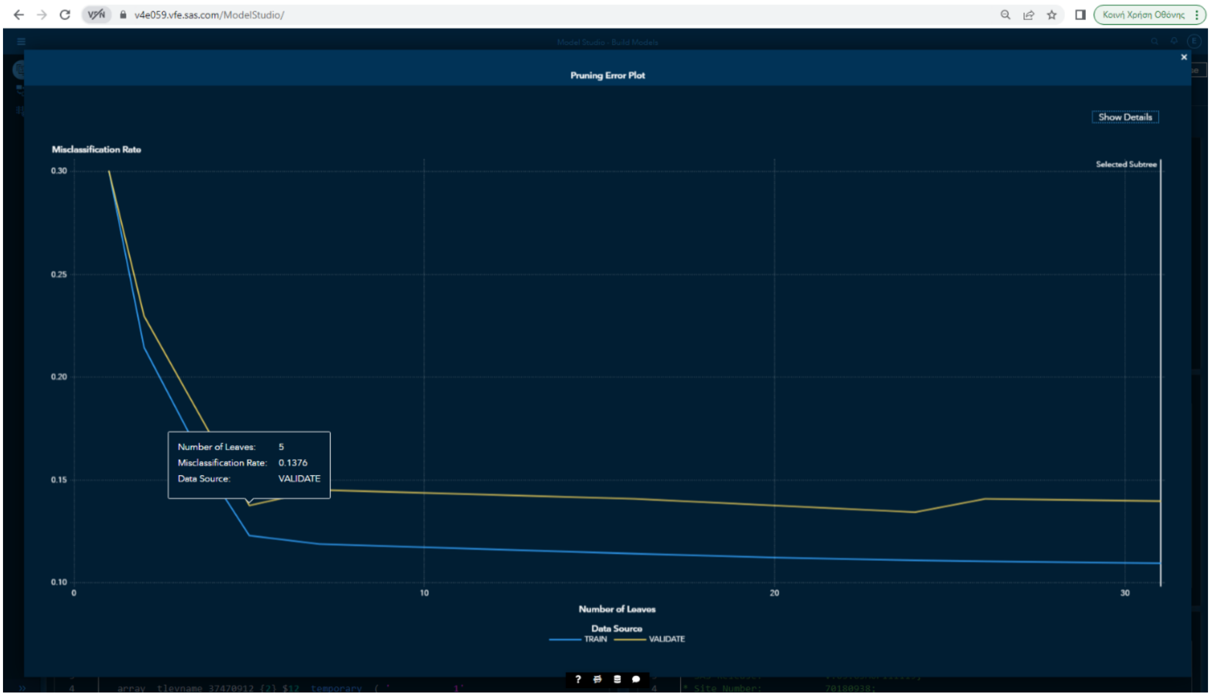
Because we have used the option "use in search" for missing values, missing values are used as a value. For nominal inputs (e.g., Make) the missing values are treated as a separate level; for ordinal inputs (e.g., Past number of Claims) it is required modification of the split search strategy by adding a separate branch adjacent to the ordinal levels; for interval inputs (e.g., Age of Vehicle) missing values are treated as having the same unknown non-missing value. For inputs with missing values two sets of Bonferroni-adjusted $logworths$ are generated and the best split is selected from the set of possible splits with the missing values in the left and right branches. In our case missing values are directed at the right branches. However, as we have seen from question 5, there are no missing values in the variables of the data set.

10) The second decision tree that we added has 31 terminal nodes (leaves) and it is called Maximal Tree. The Subtree Assessment Plot below (Figure 12) shows how the misclassification rate changes for subtrees, which are created by pruning the full decision tree to various numbers of leaves. The training error decreases as the number of leaves increases. For this decision tree model, the selected subtree based on the pruning options has 31 leaves with a misclassification rate of 0.1398 for the VALIDATE partition. This is referred to as overfitting and it is the phenomenon presented for the training data set.

A model that is complex enough to perfectly fit the existing data will not generalize well when used to score new observations. It might provide accurate answers for some cases by chance, but in general it does not represent the trend of the data. If the tree can continue to split the data all the way down to each observation being in its own leaf, it will be 100% accurate for every observation in the training data. But after a certain depth, the tree is not providing any information that can be applied in general. The maximal tree is the result of overfitting.

The solution to avoid overfitting the decision tree is pruning, a mechanism for adjusting model complexity. We can utilize the VALIDATE partition to prune the tree in order to prevent overfitting. For our case presented below (Figure 12), the best solution provided is to prune the tree at 5 leaves, which provide a better misclassification rate (0.1376) compared to the previous solution provided by the maximal tree.

A screenshot of the largest tree in the report is presented below (Figure 13).



Number of Leaves: 5
Misclassification Rate: 0.1376
Data Source: VALIDATE

Figure 12 - Subtree Assessment Plot of the Maximal Tree

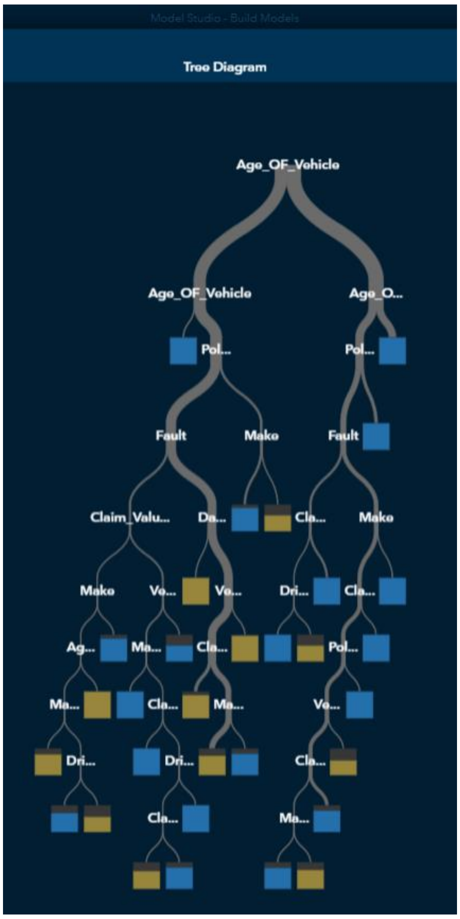


Figure 13 - Largest Tree (Maximal Tree)

11) The optimal tree has 11 leaves (terminal nodes), as shown in Figure 14. The Subtree Assessment Plot, when Misclassification Rate is selected (Figure 15), shows how the misclassification rate changes for subtrees,

which are created by pruning the full decision tree to various numbers of leaves. The training error decreases as the number of leaves increases, so the VALIDATE partition can be used to prune the tree to prevent overfitting. For this decision tree model, the selected subtree based on the pruning options has 11 leaves with a misclassification rate of 0.1333 for the VALIDATE partition.

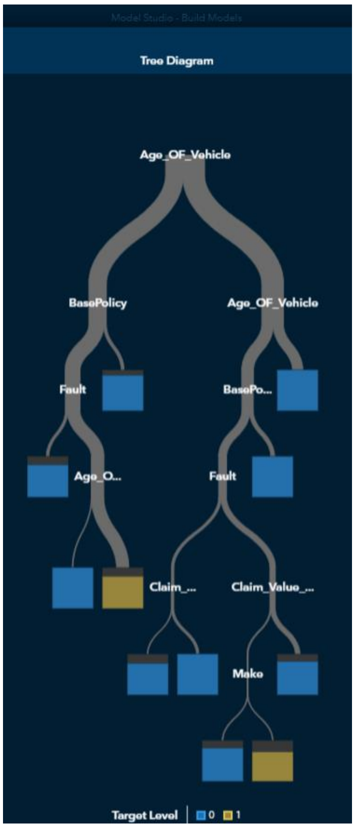


Figure 14 - Optimal Decision Tree

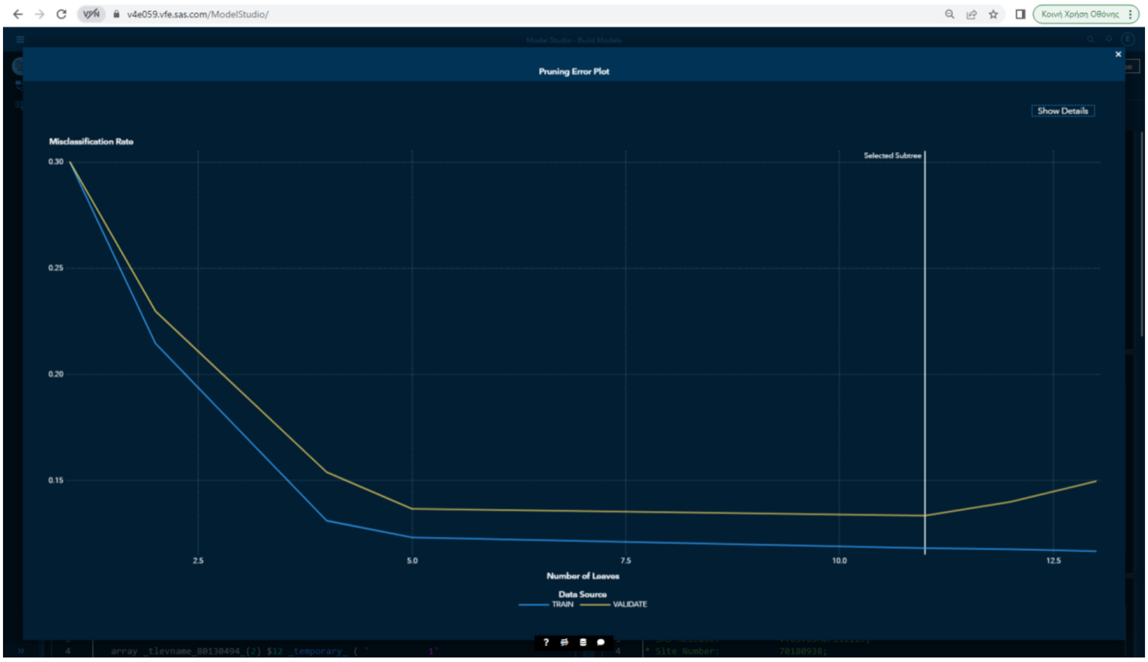


Figure 15 - Subtree Assessment Plot of the Optimal Decision Tree

12) The Decision Tree Model that we have trained and evaluated on the Pipeline so far is a model that has the following parameters:

a) It uses a chi-square statistic to split each variable and then uses the p-values that correspond to the resulting splits to determine the splitting variable with Bonferroni Adjustment and Significance Level of 0.2.
b) The maximum branches per Node are 2 (binary tree).
c) The maximum depth at which the tree can grow is 10.
d) The minimum number of cases that a node can incorporate is 5 (a node having 5 cases cannot be split further).
e) The handling of missing values is done via the “use in search” option.
f) The number of interval bins is 50 and the Interval bin method is quantile.
g) As a pruning method the smallest subtree with the best assessment value is chosen (reduced error).
h) The way to construct the subtree in terms of selection methods is automatic.

Interpretation:

- If we have a claim that the Age of the Vehicle is greater than or equal to the threshold of 8 years and the Base Policy is Liability then the model classifies the claim as fraudulent, because it exceeds the cut-off point of 6.25% probability of being fraudulent (13.19%).
- If we have a claim that the Age of the Vehicle is less than 8 years and in fact is even less than 7 years or it has a missing value, then the model classifies the claim as non-fraudulent, because the probability cut-off point of being fraudulent is not exceeded (0%).
- If we have a claim that the Age of the Vehicle is greater than or equal to the threshold of 8 years and the Base Policy is Collision or All Perils and the Fault is Policy Holder and the Age of the Vehicle is greater than or equal to 0 and less than 15 years, then the model classifies the claim as fraudulent, because it exceeds the cut-off point of 6.25% probability of being fraudulent (78.31%).
- If we have a claim that the Age of the Vehicle is greater than or equal to the threshold of 8 years and the Base Policy is Collision or All Perils and the Fault is Third Party, then the model classifies the claim as fraudulent, because the probability cut-off point of being fraudulent is exceeded (21.09%).
- If we have a claim that the Age of the Vehicle is less than 8 years and greater than or equal to 7 and the Base Policy is Liability, then the model classifies the claim as non-fraudulent, because the probability cut-off point of being fraudulent is not exceeded (0.49%).

13) Explanation of the decision tree to the management team of the insurance organization:

The mathematical model developed analyzes the historical data from the period May – September 2017 and can be applied to new claims issued after 1st of October 2017 in order to predict whether they are fraudulent or not. From the developed model we derive that:

- Claims that the vehicle is 8 years old or older and the Base Policy is Liability are more likely to be fraudulent according to the model and therefore should be directed to the investigation department for further checks.
- Claims that the vehicle is less than 7 years old are more likely to be legitimate according to the model and therefore should be compensated.
- Claims that the vehicle is between 8 and 15 years old and the Base Policy is Collision or All Perils and the Fault is Policy Holder are more likely to be fraudulent according to the model and therefore should be directed to the investigation department for further checks.
- Claims that the vehicle is 8 years old or older and the Base Policy is Collision or All Perils and the Fault is Third Party are more likely to be fraudulent according to the model and therefore should be directed to the investigation department for further checks.
- Claims that the vehicle is between 7 and 8 years old and the Base Policy is Liability are more likely to be legitimate according to the model and therefore should be compensated.

The most important variables that separate fraudulent from legitimate claims are in ascending order of importance the following: the age of the vehicle, the fault, the base policy, the claim value divided by vehicle value, and the Car's Make.

14) The Cumulative % Response for the validation data set shows that if for example we send for investigation the 20% of the claims with the highest probability of being fraudulent according to the probabilities that the best model (Decision Tree) gives, the 75.593% of this 20% will be truly fraudulent claims, whereas if we send for investigation all (100%) claims only the 30.011% of them will be truly fraudulent claims.

15) In order to construct the % Response chart we firstly order the claims by their probability of being fraudulent according to the probabilities the best model produces. Then we separate them into 20 buckets, each holding the 5 percent of the total population (claims) of the partition (train, validation). Each bucket is represented in the x axis of the chart (5% the 1st bucket, 10% the 2nd, etc.) and in the y axis we depict the percentage of the truly fraudulent claims that each bucket holds.
If for example we send for investigation the fifth (5th) bucket (20% - 25%) of the claims with the highest probability of being fraudulent according to the probabilities that the best model (Decision Tree) gives, the 75.593% of this bucket will be truly fraudulent claims.

16) The Cumulative Lift chart for the validation data set shows that if we send for investigation the 20% of the most probable fraudulent claims, according to the probability that the best model gives them to be fraudulent, we will capture 2.5653 times more fraudulent claims than if we did the same job without a model i.e., at random.

17) The Cumulative % Captured Response graph for the validation data set shows that if we send for investigation the 40% of the most probable fraudulent claims, according to the probability that the best model gives them to be fraudulent, we will capture the 86.265% of all the responders of the whole validation data set.

18) Below, in Figure 16, we can observe the completed process flow with the Score Data Node. In the new data set i.e., "New_Claims_Final", there exist 200 claims, from which 160 are truly non-fraudulent and 40 are truly fraudulent. From those 200 claims, 52 are predicted as fraudulent whereas 148 are predicted as non-fraudulent, as shown in Figure 17 (cases with predicted probability of being fraudulent are those above 0.5).

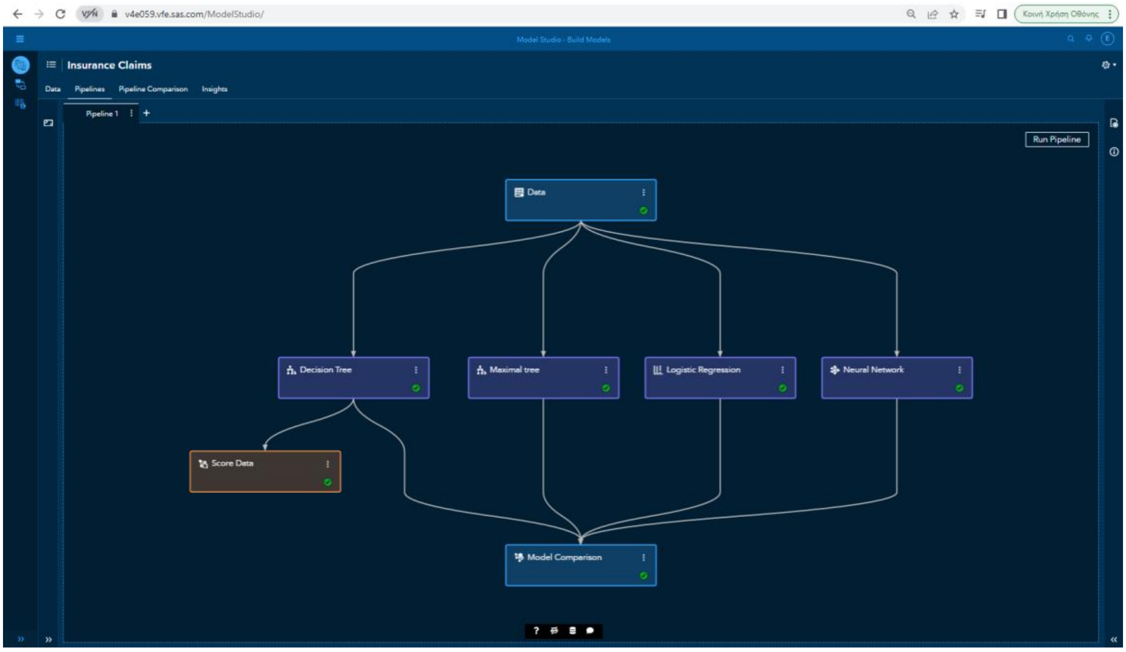
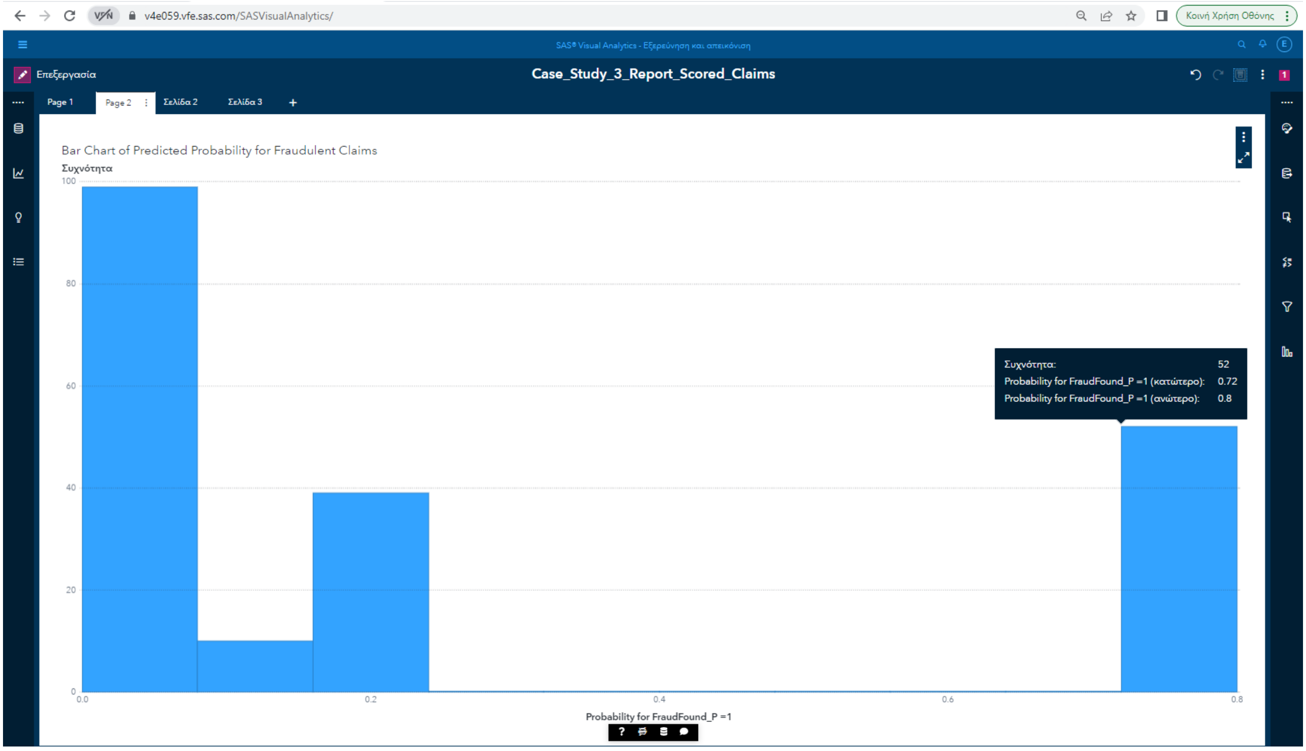


Figure 16 - Completed Process Flow



v4e059.vfe.sas.com/SASVisualAnalytics/

SAS® Visual Analytics - Εξερεύνηση και απεικόνιση

Case_Study_3_Report_Scored_Claims

Bar Chart of Predicted Probability for Fraudulent Claims

Συχνότητα: 52
Probability for FraudFound_P =1 (κατώτερο): 0.72
Probability for FraudFound_P =1 (ανώτερο): 0.8

Figure 17 - Bar Chart of the Predicted Probability for Fraudulent Claims

19) The biggest probability of being fraudulent assigned to a claim is 80% and the smallest one is 72%, as shown in the figure above (Figure 17).

20) The Claim with Policy ID=15 has the following values in the forementioned important variables: Vehicle Age = 7, Fault = Policy Holder, Base Policy = Liability, Division = 7.05%, Make = Honda. Because the vehicle's age is less than 8 and bigger than or equal to 7 and the Base Policy is Liability, it is assigned to the Node with ID: 10, which is a terminal node (leaf) of the decision tree model and thus, it is classified as a non-fraudulent claim (0), because the probability cut-off point of being fraudulent is not exceeded (0.49%) for this leaf.

The Claim with Policy ID=107 has the following values in the forementioned important variables: Vehicle Age = 10, Fault = Policy Holder, Base Policy = Collision, Division = 57.75%, Make = Honda. Because the vehicle's age is greater than or equal to 8 and the Base Policy is Collision or All Perils and the Fault is Policy Holder and the vehicle's age is between 0 and 15, it is assigned to the Node with ID: 12, which is a terminal node (leaf) of the decision tree model and is classified as fraudulent claim (1), because it exceeds the cut-off point of 6.25% probability of being fraudulent (78.31%).