MSc in Business Analytics

Course: Business Analytics Practicum I

Associate Professor: A. Zaras

BUSINESS ANALYTICS PRACTICUM I

SAS

GROUP ASSIGNMENT

Chatzimoschou angeliki

baptXXXX

melenikou galateia

bapTXXXX

APRIL 2017

# TABLE OF CONTENTS

[TABLE OF CONTENTS 2](#_Toc479108917)

[CASE STUDY I 3](#_Toc479108918)

[EXECUTIVE SUMMARY 3](#_Toc479108919)

[MAIN BODY 4](#_Toc479108920)

[1- WHAT ARE THE SALES OF EACH BOOK? 4](#_Toc479108921)

[2- WHICH BOOKS SHOULD THE STORE ADVERTISE TO CUSTOMERS? 4](#_Toc479108922)

[3- WHICH ARE THE 3 BOOKS MOST BOUGHT TOGETHER BY CUSTOMERS? 5](#_Toc479108923)

[CASE STUDY II 6](#_Toc479108924)

[EXECUTIVE SUMMARY 6](#_Toc479108925)

[CASE STUDY III 9](#_Toc479108926)

[EXECUTIVE SUMMARY 9](#_Toc479108927)

[MAIN BODY 10](#_Toc479108928)

[QUESTION 1 10](#_Toc479108929)

[QUESTION 2 11](#_Toc479108930)

[QUESTION 3 11](#_Toc479108931)

[QUESTION 4 12](#_Toc479108932)

[QUESTION 5 12](#_Toc479108933)

[QUESTION 6 14](#_Toc479108934)

[QUESTION 7 15](#_Toc479108935)

[QUESTION 8 16](#_Toc479108936)

[QUESTION 9 17](#_Toc479108937)

[QUESTION 10 18](#_Toc479108938)

[QUESTION 11 19](#_Toc479108939)

[QUESTION 12 20](#_Toc479108940)

[QUESTION 13 20](#_Toc479108941)

[QUESTION 14 21](#_Toc479108942)

[QUESTION 15 22](#_Toc479108943)

[QUESTION 16 23](#_Toc479108944)

[QUESTION 17 23](#_Toc479108945)

[QUESTION 18 24](#_Toc479108946)

[QUESTION 19 25](#_Toc479108947)

[QUESTION 20 25](#_Toc479108948)

# CASE STUDY I

## EXECUTIVE SUMMARY

Since Buy-books-online.com is already a famous and successful on-line book store in the academic community, the sales department nowadays wants to expand its transactions, so as to sell as many books as possible. In order to achieve this target, it needs to exploit cross-selling opportunities by providing wise recommendations derived by applied association rules.

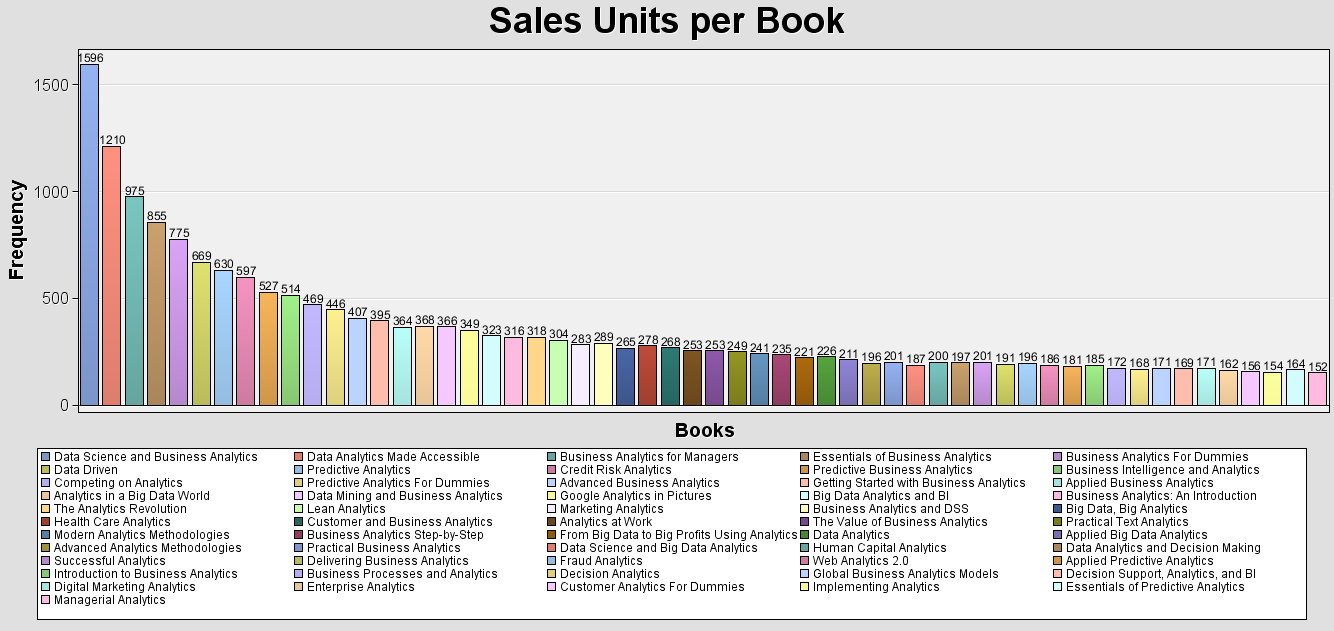
The tool used to conduct this analysis is SAS Enterprise Miner. The technique used is association rule learning which is a rule-based machine learning method for discovering interesting relations between book transactions and for creating further buying recommendations with total correlation. With the application of a system of associations mapping between book titles, the analysis provides valuable insight concerning the combinations of books that customers prefer to buy together.

Thus, the marketing department will have the necessary information to be in a position to apply targeted promotion techniques in order to increase not only sales numbers, but also the customers’ satisfaction. Useful insights follow below:

MAIN BODY

## 1- WHAT ARE THE SALES OF EACH BOOK?

Beginning with our analysis, we examine the sales units of each book, related to the category “Business Analytics”. Using SAS Enterprise Miner tool, the following bar chart shows, for each of 56 books, the relative total sales number, in descending order.



*Figure 1 – Bar Chart for Sales Units per Book*

## 2- WHICH BOOKS SHOULD THE STORE ADVERTISE TO CUSTOMERS?

In order to do next best offer propositions, we apply association rules on the data set and we focus on the following 4 books:

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

Depending on customers’ initial selection concerning one of the books mentioned above, we create a list of different association rules, taking into account the metric: lift, so as to recommend suitable books.

* The customers interested in “Managerial Analytics” are more likely to buy also the books “Web Analytics 2.0” & “Implementing Analytics”. The lift of this rule is equal to 11.47, which is the largest.
* The customers interested in “Implementing Analytics” are more likely to buy also the books “Managerial Analytics” & “Data Science and Big Data Analytics”. The lift of this rule is equal to 11.33.
* The customers interested in “Customer Analytics For Dummies” are more likely to buy also the books “Essentials of Predictive Analytics” & “Enterprise Analytics”. The lift of this rule is equal to 11.19.
* The customers interested in “Enterprise Analytics” are more likely to buy also the books “Managerial Analytics” & “Customer Analytics For Dummies”. The lift of this rule is equal to 11.07.

For more information, the analysis report is attached below:



## 3- WHICH ARE THE 3 BOOKS MOST BOUGHT-TOGETHER BY CUSTOMERS?

In order to find the 3 most bought-together books, we use the metric: support, which expresses the probability of the combinatorial transactions we want, divided by the total transactions. These 3 books are: “Implementing Analytics”, “Data Science” and “Big Data Analytics & Applied Predictive Analytics”.

We concluded to these 3 books because our highest support association rules are the ones given below:

* “Implementing Analytics” ==> “Data Science and Big Data Analytics” & “Applied Predictive Analytics” (Support: 7.28%)
* “Data Science and Big Data Analytics” & “Applied Predictive Analytics” ==> “Implementing Analytics” (Support: 7.28%)

Thus, the occurrences (No of transactions) of this 3 books set is calculated by the support (0.0728) multiplied by the total number of sales (19805).

No of transactions = 0.0728\*19805 ⇔ No of transactions = 1442 transactions

# CASE STUDY II

## EXECUTIVE SUMMARY

As a well-established online retailer on sports clothes and shoes, Sports-OnLine.com, needs a deeper understanding of its market. In this analysis, we have a big dataset containing information about 4906 sales transactions of 995 customers from year 2001 to 2006. The decision taken by the marketing department, in order to achieve this target, is to do a customer segmentation analysis and more specifically an RFM (Recency- Frequency- Monetary) analysis.

The goal of RFM Analysis is to segment customers based on their historical buying actions. Afterwards, we rank them based on each individual RFM factor, and finally pull all the factors together to create RFM segments for targeted marketing. Recency is the number of months since the customer’s last purchase. Frequency is the number of orders placed in a given time period. Monetary Value is the total amount of money spent by the customer over a certain time period.

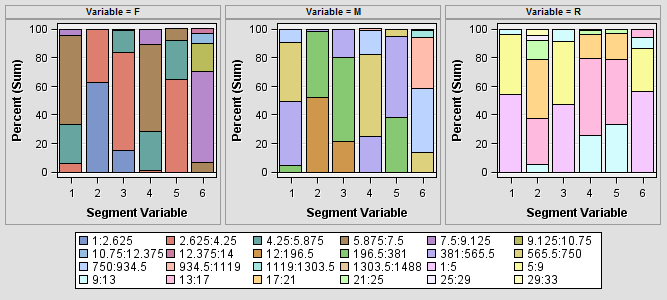
There are three basic steps to RFM analysis:

1. Sort all customers in ascending order based on Recency, Frequency and Monetary Value.

2. Split customers into quartiles for each factor.

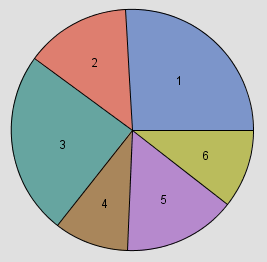
3. Combine factors to group customers into RFM segments for targeted marketing.

The results of our RFM analysis are shown below. More details are available through RFM table.



*Figure 2 – Bar Chart for each RFM factor*

**CUSTOMER SEGMENTS**



*Figure 3 – Pie Chart for RFM Customer Segmentation*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster | R  values | R  mean | Description | F  values | F  mean | Description | M values | M  mean | Description |
| 1 | 1 to 10 | 5.32 | High | 4 to 8 | 5.94 | Low | 313 to 895 | 570.45 | Medium/Low |
| 2 | 11 to 33 | 19.26 | Low | 1 to 4 | 2.2 | Low | 12 to 388 | 185.41 | Low |
| 3 | 1 to 12 | 5.68 | High | 1 to 6 | 3.56 | Low | 23 to 492 | 288 | Low |
| 4 | 10 to 29 | 15.47 | Low | 4 to 9 | 6.15 | High/Medium | 435 to 937 | 644.5 | High/Medium |
| 5 | 10 to 23 | 15.13 | Medium/Low | 3 to 7 | 4.24 | Low | 222 to 632 | 412 | Low |
| 6 | 1 to 16 | 5.63 | High | 7 to 14 | 9.1 | High | 595 to 1488 | 905 | High |

*Table 1 – RFM Table*

Now that all customers are divided into RFM segments, we have identified a number of different groups that need to be approached by very different marketing strategies. Here are the customer groups using this technique:

Cluster **1** and **3** contain the first time customers who recently bought from our shop and spent medium to low amount of money on their first orders. Since they are new comers, their Frequency rate is rather low. This is the kind of customers we want to convert into loyal, regular customers that love our products and brand. It is proposed to welcome them and thank them for making a first purchase, and follow it up with unique offers to lure them back again. Consider branding the email with a special note from the CEO, and include a survey to ask about their experience.

Cluster **6** contains the best customers who have bought recently, buy often and spend a lot. It’s likely that they will continue to do so. Since they already like us so much, consider marketing to them without price incentives to preserve your profit margin. It is proposed to tell these customers about new products we carry, how to connect on social networks, and any loyalty programs or social media incentives we run.

Cluster **2** contains the worst customers, in other words the ones who spent very little, have bought very few times, and last ordered quite a while ago. They are unlikely to be worth much time, so it is suggested to put them in our general house list and consider a re-opt-in campaign.

Cluster **4** contains lost customers (churners), who used to buy frequently from us, and at one point they spent a lot with us, but they’ve stopped. Now it’s time to win them back. They might be lost to a competitor; they might not have need of our products anymore, or they might have had a bad customer service experience with us. Regardless, they were an extremely valuable customer that should be approached differently.

Similar to lost customers is Cluster **5**. For these almost lost customers, it has been time since their last purchase. These customers might need more aggressive discounts so that we can win them back before it’s too late, since it is much less expensive to maintain them compared to attracting new ones.

Many times segments can be combined to receive an offer. For example, you may combine “Lost Customers” and “Almost Lost” for a win-back campaign. “Going silent” or creating “contact fatigue” have negative consequences and should be managed appropriately.

# CASE STUDY III

EXECUTIVE SUMMARY

Although XYZ is a respectable organization in the field of retail banking, it has been granting consumer loans to its customers by using a generic scorecard developed by a credit agency using external data, coupled with the intuition and experience of its officers. The past year the regulator of the banking system has obliged the credit institutions to modernize by adopting objective methods for credit operations and more specifically to develop in house statistically based credit risk systems using their own data. This way the decisions will be more fact-based and less by intuition.

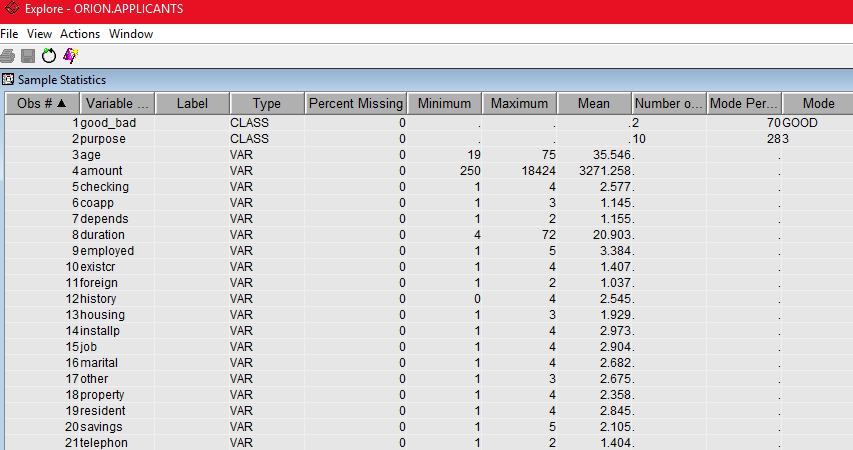
With the analysis below we will provide the bank with a new scoring system that based on historical data, concerning previous transactions. We will build several models that will help setup the framework according to which a loan applicant will be classified as “good” or “bad” candidate and going one step further, by designing a profit matrix that will provide accurate information about unseen possibilities either to increase profitability or help to avoid potentially harmful decisions.

Decision trees, regression analysis and neural network models were deployed and their results were thoroughly tested and compared in order to reach an optimal solution.

MAIN BODY

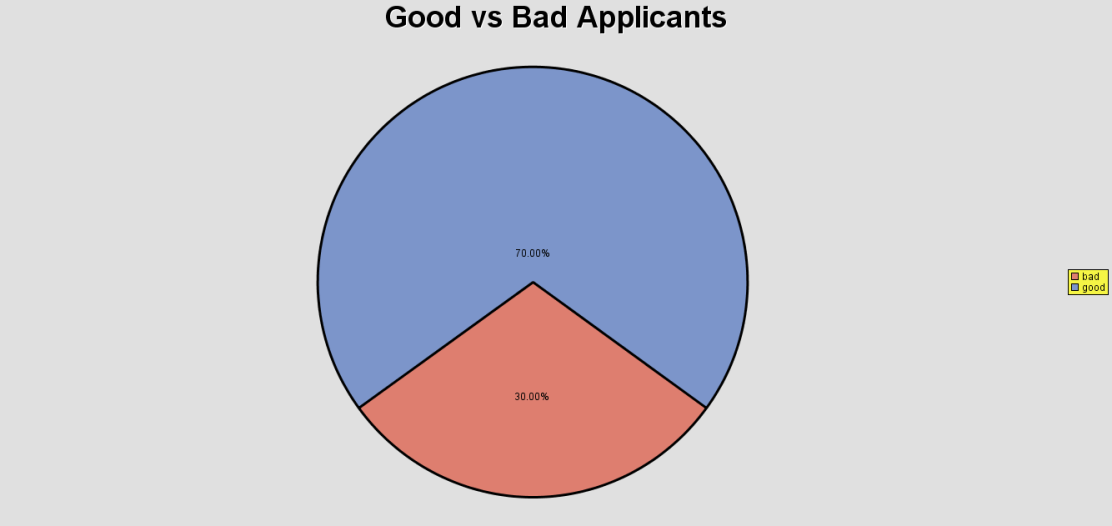
## QUESTION 1

In order to explore the “Applicants” dataset in SAS Enterprise Miner, we run the Stat Explore node where we can see various details about the dataset used. There are no missing values, as indicated by the column “Percent Missing” in the following picture:



*Figure 4 – Data Exploration*

In addition, the proportion of good clients is significantly larger in comparison to the proportion of bad clients which equals to the 30% of the total, as shown at the pie chart below.



*Figure 4 – Pie Chart for Applicants*

## QUESTION 2

The original proportion of good and bad clients was 90% - 10% which may cause inefficiencies in the model development process and requires adjustment. So, it is often the case to prepare a more balanced data sample using a user-defined fixed proportion of good and bad clients. Depending on the proportion of the rare event, we apply separate sampling (rare event < 5%) or oversampling (5% < rare event < 10%). In this case, we follow the rule of thumb-separate sampling, since the proportion of bad clients is 10, and we re-partition the dataset in a 70%- 30% proportion.

In order to calculate the number of applicants in the original data set, it is calculated in the following table:

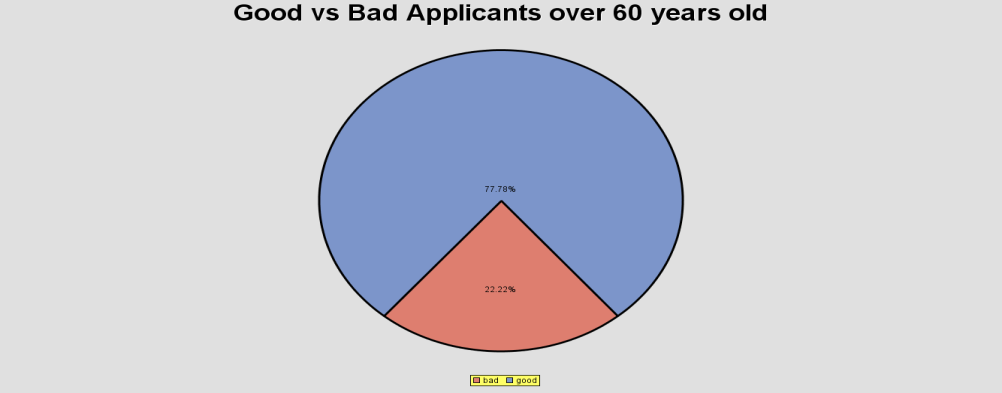
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Good clients | Bad clients | Total |
| Adjusted data set | Proportion | 70% | 30% | 100% |
| Absolute | 700 | 300 | 1000 |
| Original data set | Proportion | 90% | 10% | 100% |
| Absolute | 700 | 78 (700\*0.1/0.9) | 778 |

*Table 2 – ‘Applicants’ Data set*

The sample was created after implementing the balancing method of oversampling to the minority class (bad applicants).

QUESTION 3

The proportion of the good-bad applicants who are over 60 years old equals to 77,8% good vs 22.22% bad as it is shown on the below graph. This analogy in percentages is almost the same as the total sample which leads to the inference that individuals over 60 years old tend more to be good applicants and do not differ from the behavior of the entire population.



*Figure 5 – Pie Chart for Applicants over 60 years old*

QUESTION 4

By dropping in the diagram the “Stat Explore” node and setting as “target” the variable “good\_bad”, we conclude that the average age of bad applicants is 34, while the average age of good applicants is 36.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Interval Variable Summary Statistics by Class Target  (maximum 500 observations printed) | | | | | | | | | | | | |
| Data Role=TRAIN Variable=age | | | | | | | | | | | | |
| Target | **Target**  **Level** | **Median** | **Missing** | **Non**  **Minimum** | **Minimum** | **Maximum** | **Mean** | **Standard Deviation** | **Skewness** | **Kurtosis** | **Role** | **Label** |
| \_OVERALL\_ |  | 33 | 0 | 1000 | 19 | 75 | 35.546 | 11.37547 | 1.020739 | 0.59578 | INPUT | age |
| good\_bad | bad | 31 | 0 | 300 | 19 | 74 | **33.96333** | 11.22238 | 1.155186 | 0.787579 | INPUT | age |
| good\_bad | good | 34 | 0 | 700 | 19 | 75 | **36.22429** | 11.38114 | 0.981607 | 0.574085 | INPUT | age |

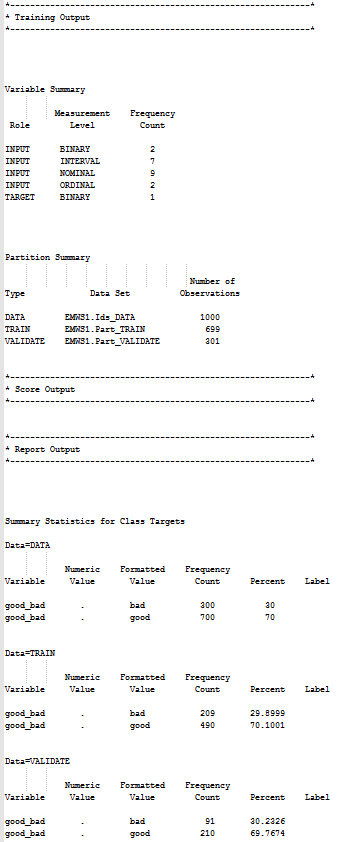
*Table 3 – Summary Statistics*

The results above indicate that the classification groups actually share similar age range. The average ages of bad and good applicants differ from each other only by 2 years.

QUESTION 5

By adding a data partition node on the diagram and assigning 70% of the initial dataset for training and 30% for validation, the software output is in the following page.

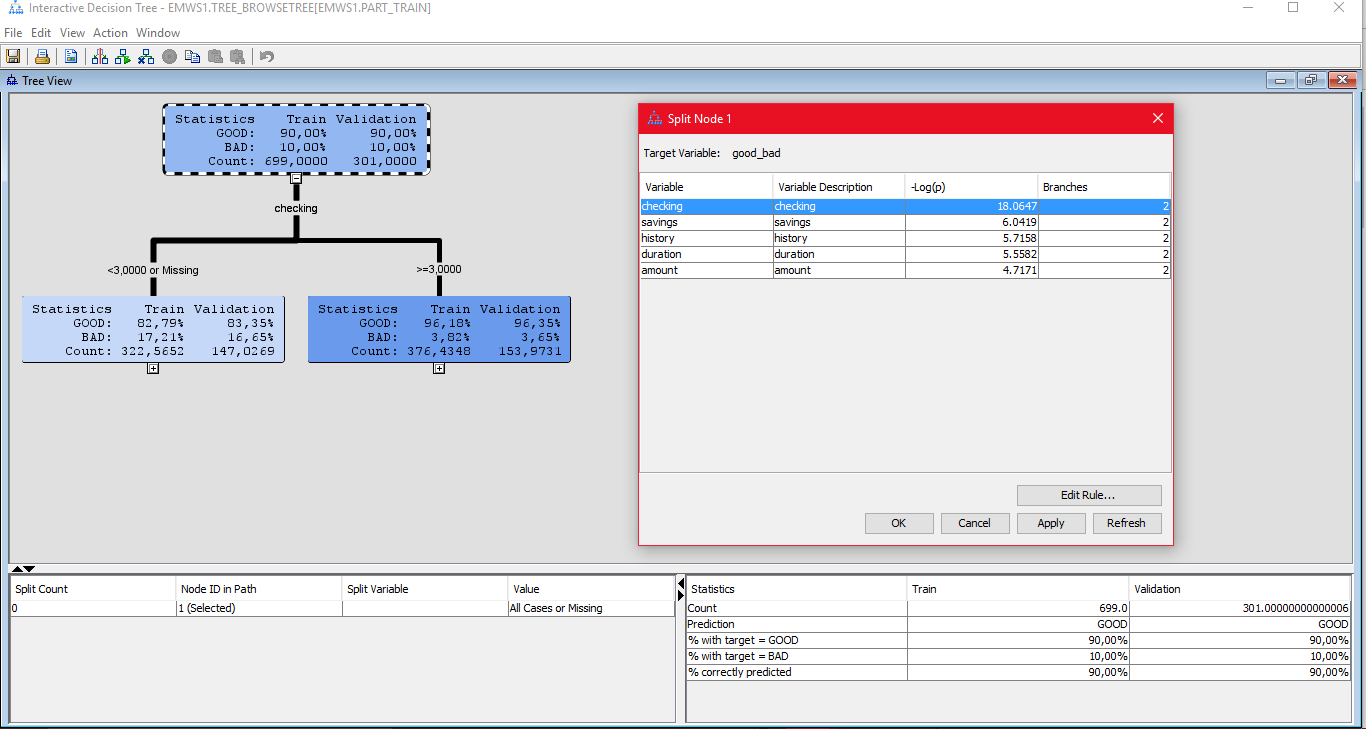
The sample in the data portion was re-balanced, which means that the initial set is partitioned into groups that share certain characteristics and from these groups a sample is being drawn randomly in a proportion of 70%-30%. SAS creates by default stratified samples and this can be seen by the table above where we see that the proportions of good - bad customers remains 70%-30% for both the train and the validate partitions, just like the split of the initial dataset.



*Table 4 – Summary Statistics*

QUESTION 6

After setting prior probabilities, we created a Decision Tree interactively, as shown in the images.

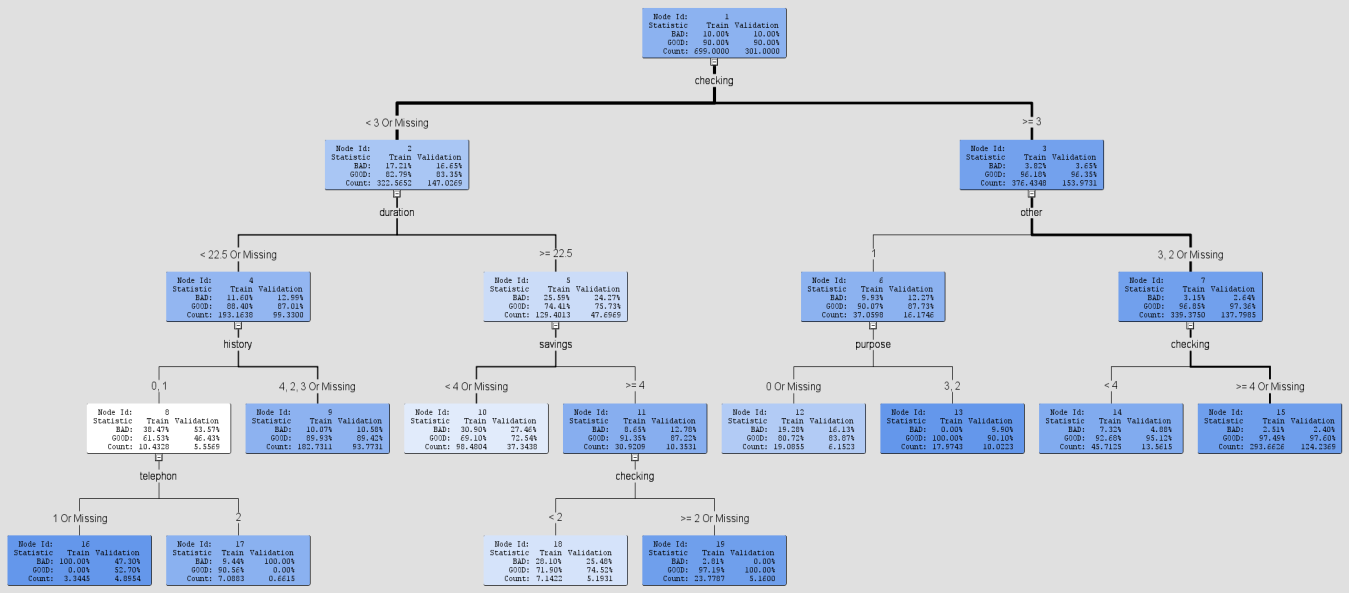


*Figure 6 – Decision Tree*

We observe that the variable used for the first split is “checking” with a logworth equal to 18.0647. Cases with checking less than 3 or missing values are directed to the left node, while cases bigger than 3 are directed to the right node.

QUESTION 7

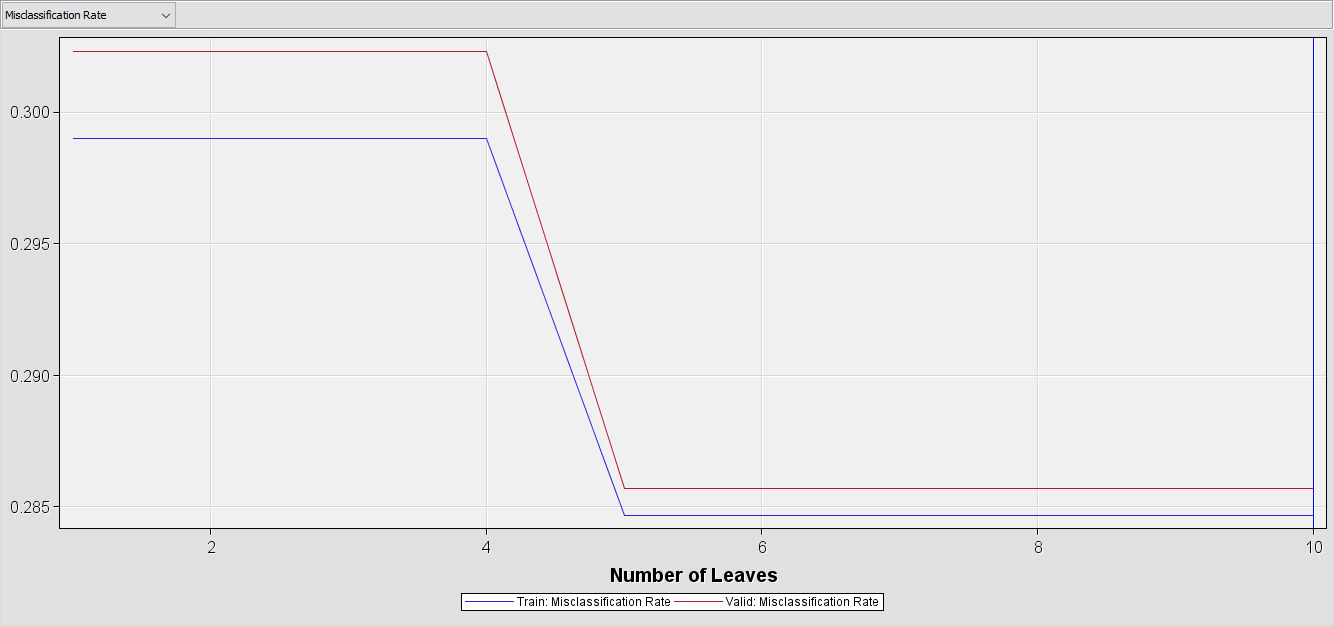
By adding and running a new Decision Tree node in the diagram, we observe that the “largest tree” has 10 terminal leaves and it is called “maximal tree”. It is shown below:



*Figure 7 – Maximal Tree*

We also checked the performance of the training and validation datasets using the misclassification rate as assessment criterion. The results are shown in the relevant graph below. According to the train trendline, the misclassification error drops significantly after the first 4 splits and then around 5 leaves and as the number of leaves increases, it stabilizes.

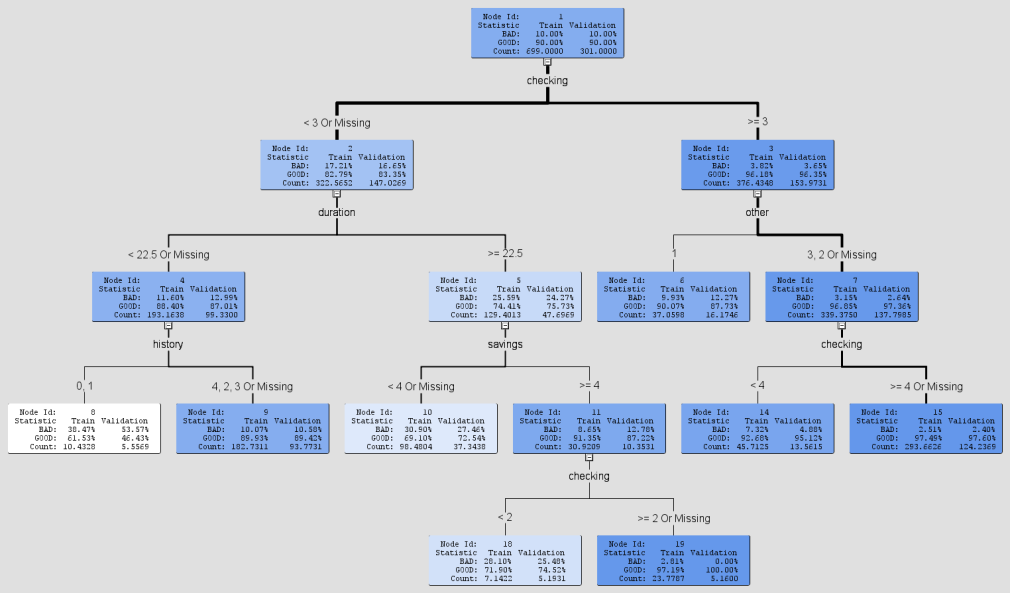
For the current number of nodes that the maximal tree has, there doesn’t seem to be a problem of overfit. The more complicated the tree gets, the more the misclassification rate approaches 0. This phenomenon is called overfitting, which means that the model, due to its complexity, describes mostly random error and noise. In order to avoid overfitting, we prune the maximal tree until the validation function’s misclassification rate is minimized.



*Figure 8 – Subtree Assessment plot (misclassification rate)*

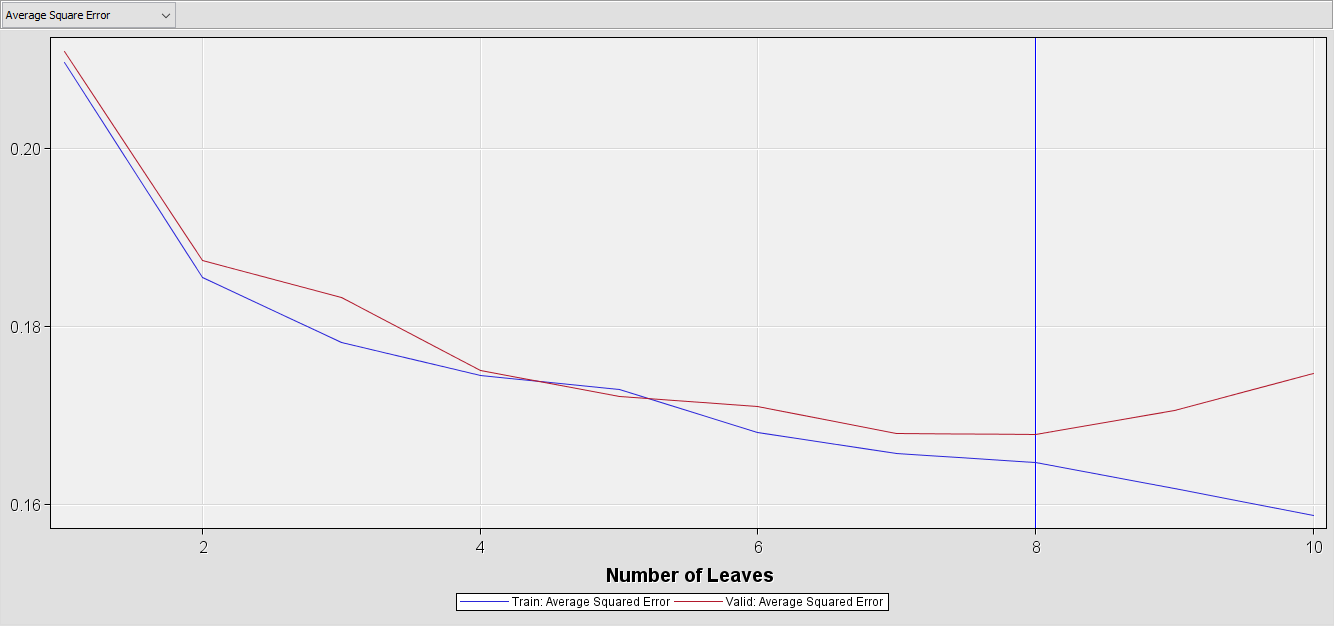
QUESTION 8

The optimal tree has 8 leaves, fewer than those of maximal tree, as shown in the figure below.



*Figure 9 – Optimal Tree*

By observing the Assessment Plot with the criterion of Average Squared Error, we confirm that the maximal tree was pruned sequentially based on the average squared error. In other words, the pruning criterion is the lowest average squared error on the validation sample among all other tree candidates. In the graph, the performance of the validation sample steadily decreases, until it reaches the 8 leaves and then, starts to gradually increase.



*Figure 10 – Subtree Assessment plot (average squared error)*

QUESTION 9

The decision tree model consists of the terminal nodes of the optimal tree. In our case, we have 8 nodes and depending on the route we follow, it allows us to decide with a specific probability whether an applicant is “good” or “bad”.

For the first terminal leaf the decision is “good” with posterior probabilities (good-bad) 46.43%- 53.57% and the English rule is “WHERE checking < 3 Or Missing AND duration < 22.5 Or Missing AND history 0, 1”.

For the second terminal leaf the decision is “good” with posterior probabilities 89.42%- 10.58% and the English rule is “WHERE checking < 3 Or Missing AND duration < 22.5 Or Missing AND history 4, 2, 3 Or Missing”.

For the third terminal leaf the decision is “good” with posterior probabilities 72.54%- 27.46% and the English rule is “WHERE checking < 3 Or Missing AND duration >= 22.5 AND savings < 4 Or Missing”.

For the forth terminal leaf the decision is “good” with posterior probabilities 74.52%- 25.48% and the English rule is “WHERE checking < 3 Or Missing AND duration >= 22.5 AND savings >= 4 AND checking < 2”.

For the fifth terminal leaf the decision is “good” with posterior probabilities 100.0%- 0% and the English rule is “WHERE checking < 3 Or Missing AND duration >= 22.5 AND savings >= 4 AND checking >= 2 Or Missing”.

For the sixth terminal leaf the decision is “good” with posterior probabilities 87.73%- 12.27% and the English rule is “WHERE checking >= 3 AND other 1”.

For the seventh terminal leaf the decision is “good” with posterior probabilities 95.12%- 4.88% and the English rule is “WHERE checking >= 3 AND other 3, 2 Or Missing AND checking < 4”.

For the eighth terminal leaf the decision is “good” with posterior probabilities 97.60%- 2.40% and the English rule is “WHERE checking >= 3 AND other 3, 2 Or Missing AND checking >= 4 Or Missing”.

QUESTION 10

The classification of an applicant as “good” or “bad” depends on a specific combination of features that each applicant should have. The most important combination for an applicant to be considered as “good” is to have at least 200 DM annual salary assignments or no checking account at all, as well other installment plans in stores or not at all. In this case, the applicant is classified as good with 97.60% accuracy.

Another effective combination of features that leads to a classification as “good” with accuracy of 89.42% is: having annual salary assignments from 0 to 200 DM, loan duration less than 22.5 months and as dealing history with the Bank one of the following: existing credits paid back duly until now, delay in paying off in the past or critical account/ other credits existing in other banks.

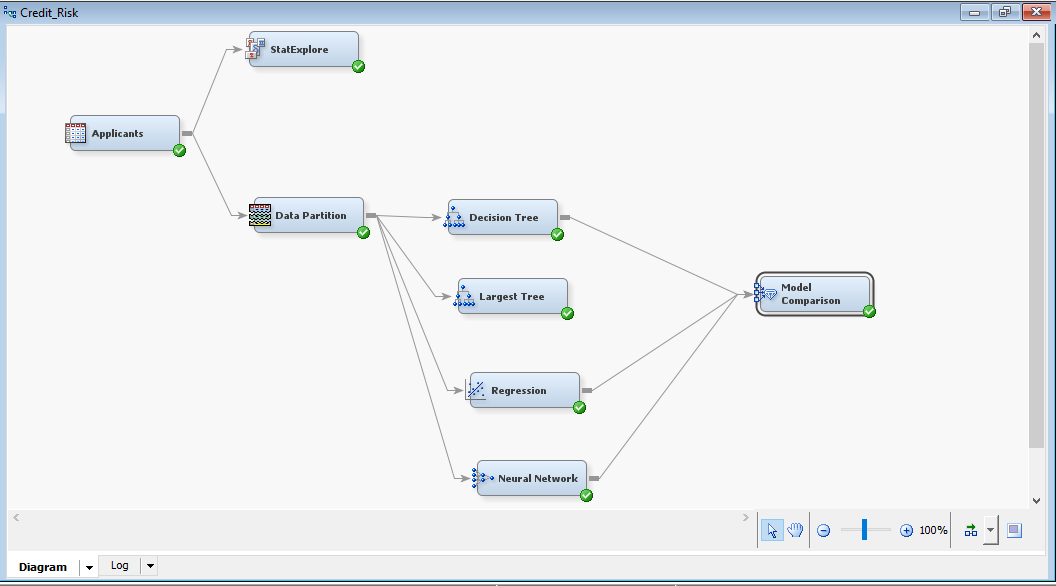
Finally, another possible combination of features that leads to a classification as “good” with accuracy of 72.54% is: having annual salary assignments from 0 to 200 DM, loan duration more than 22.5 months and savings from 100 DM to 1000 DM.

In conclusion, the most important features that separate our groups into “good” and “bad” applicants are

* the status of existing checking account,
* loan duration,
* savings and
* each client’s history.

QUESTION 11

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Prediction | |
|  |  | Good ---- > Accept | Bad ---- > Reject |
| Actual | Good | *2000* | *-2000* |
| Bad | *-12000* | *0* |

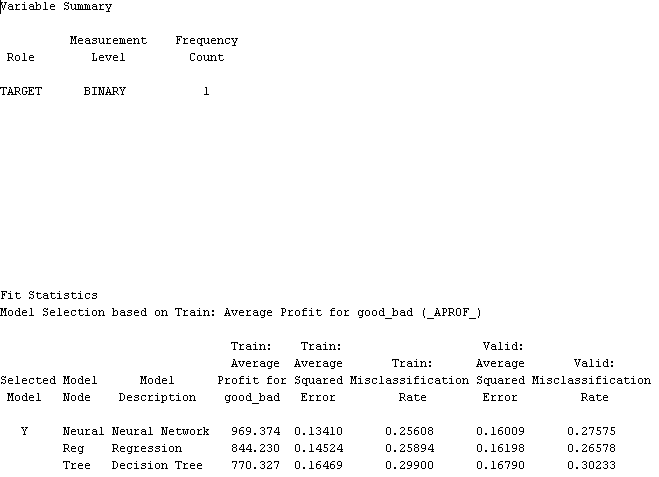


*Figure 11 – Process flow with profit matrix input*

According to the profit matrix provided, if an applicant is predicted as a “good” one, he is offered a loan and then proves to be a “good” payer (True positive), then the bank’s profit equals to 2000 MU (Monetary Units). If the applicant instead proves to be a “bad” (False positive) then the bank loses 12000 MU.

In addition, if the bank rejects a “good” applicant because they assume he is a “bad” one (False negative), then the bank loses 2000 MU. Finally, if an applicant is rejected as a “bad” applicant and he is indeed a “bad” applicant (True negative), then the bank suffers no loss.

We introduced the profit matrix in the flow by tuning the Decision option of the Applicants node. We set the Decision Name to “Accept” and “Reject” and in the Decision\_Weights, we used as an input the values from the given profit matrix. Finally, in the Model Comparison node, we set selection statistic to Average Profit/Loss and after running the flow, we got the output below where you can see that the profit matrix is included:



*Figure 12 – Fit Statistics with profit matrix*

QUESTION 12

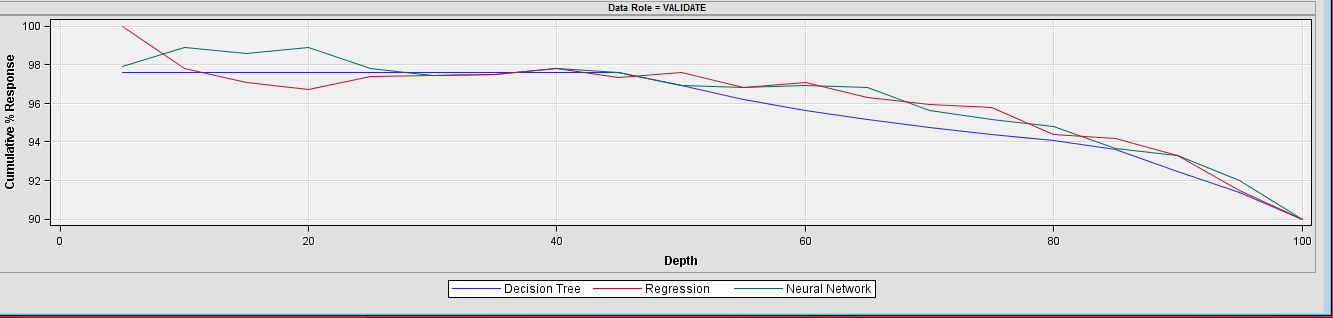
According to our profit matrix, the expected profit can be calculated as following:

* p1 is the probability an applicant is a “good” one,
* p0 is the probability an applicant is a “bad” one

Expected Profit Accept: (2000\*p1 ) – (12000\*p0) > 0

Expected Profit Reject: (-2000\*p1) + (0\*p0) > 0

QUESTION 13



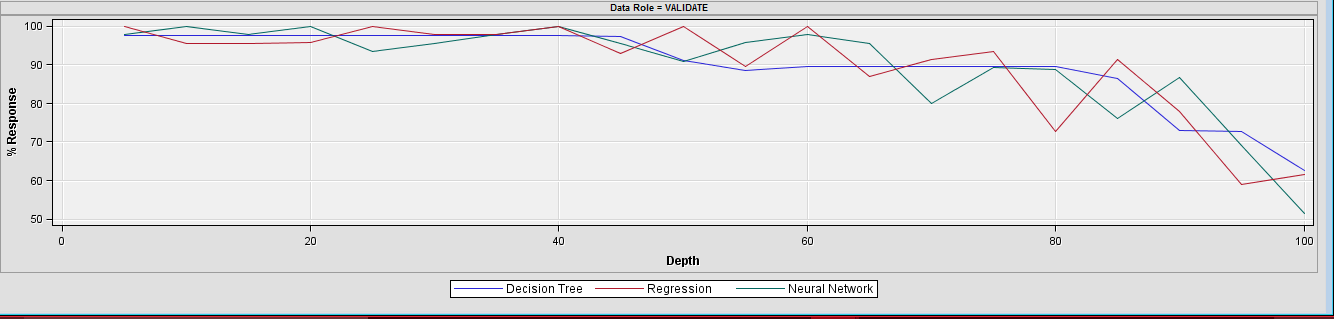
*Figure 13 – Cumulative % response chart*

Observing the cumulative percentage response plot, we may conclude that if we select 20% of the highest ranked applicants, then:

* according to the Decision Tree model, the 97.60% of them will be “good” applicants,
* according to the Logistic Regression model, the 96.71% of them will be “good” applicants and
* according to the Neural Network model, the 98.90% of them will be “good” applicants.

In addition, when we select 100% (all of them) of the highest ranked applicants, then the cumulative percentage response will reach a rate of 90% (90% of them will be “good” applicants).

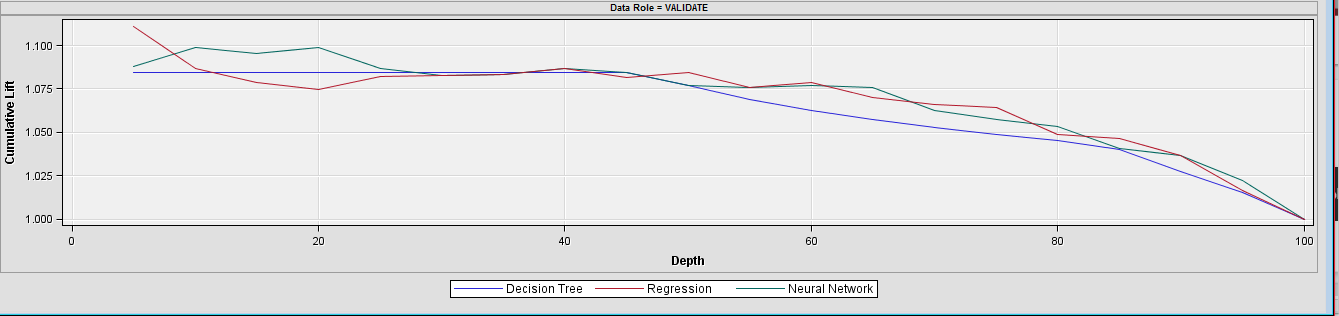
QUESTION 14



*Figure 14 – % response chart*

The % Response plot sorts applicants by their probability of response (as predicted by the model), and then groups them into 20 buckets, each one containing 5% of the highest ranked applicants (1st bucket: 1st 5% highest ranked applicants, 2nd bucket: 2nd 5% highest ranked applicants and so on until all sample is covered). The chart then plots the actual percentage of respondents in each bucket, using the value of “good” applicants. Choosing the results for half of the sample (1 to 10th bucket: 50% of the highest ranked applicants), we note that the 90.88% of them are predicted to be “good” applicants by both the Neural Network and the Decision Tree model, while the Logistic Regression model predicts a percentage of 100%.

QUESTION 15



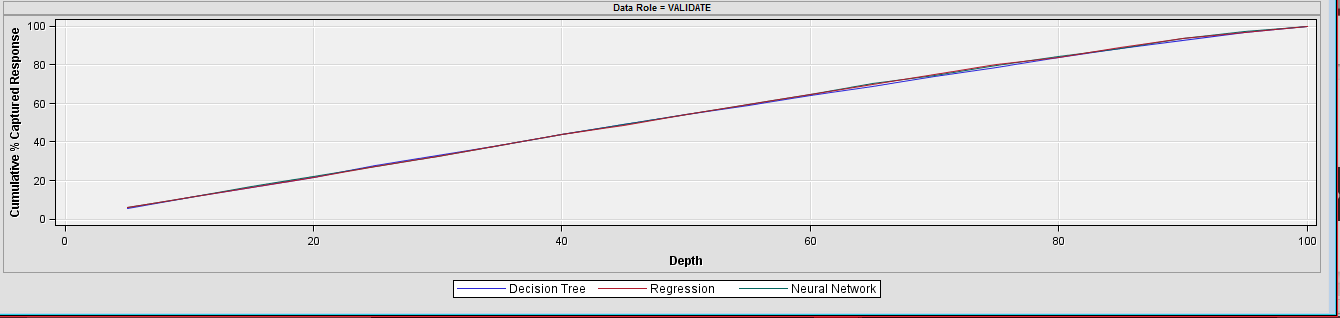
*Figure 15 – Cumulative Lift chart*

Observing the Cumulative Lift plot, we may conclude that if we select the 20% of the highest ranked applicants, then according to the Neural Network model, we will have 1.099 times more “good” applicants compared to the case when the selection is random.

In addition, if we select the 20% of the highest ranked applicants, then according to the Decision Tree model, we will have 1.084 times more “good” applicants compared to the case when the selection is random.

Finally, if we select the 20% of the highest ranked applicants, then according to the Logistic Regression model, we will have 1.074 times more “good” applicants compared to the case when the selection is random.

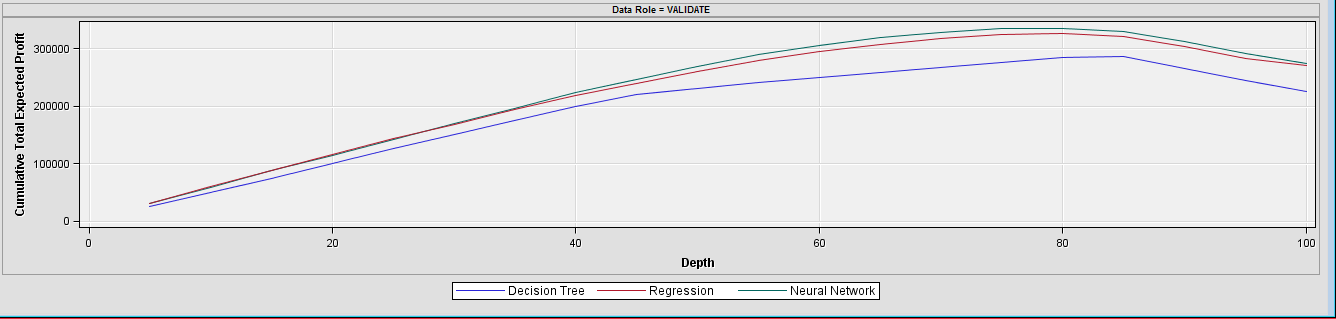
QUESTION 16



*Figure 16 – Cumulative % Captured Response graph*

Observing the Cumulative % Captured Response plot, we may conclude that if we select the 40% of the population of highest ranked applicants, the 43.81% of the validation data set will be “good” applicants, according to all of the three models that we examine.

QUESTION 17

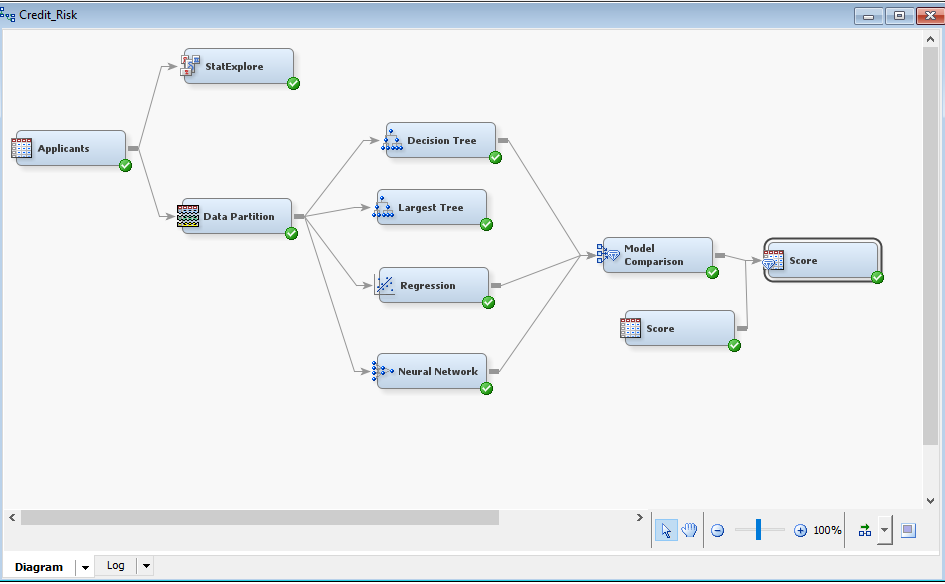


*Figure 17 – Cumulative Total Expected Profit*

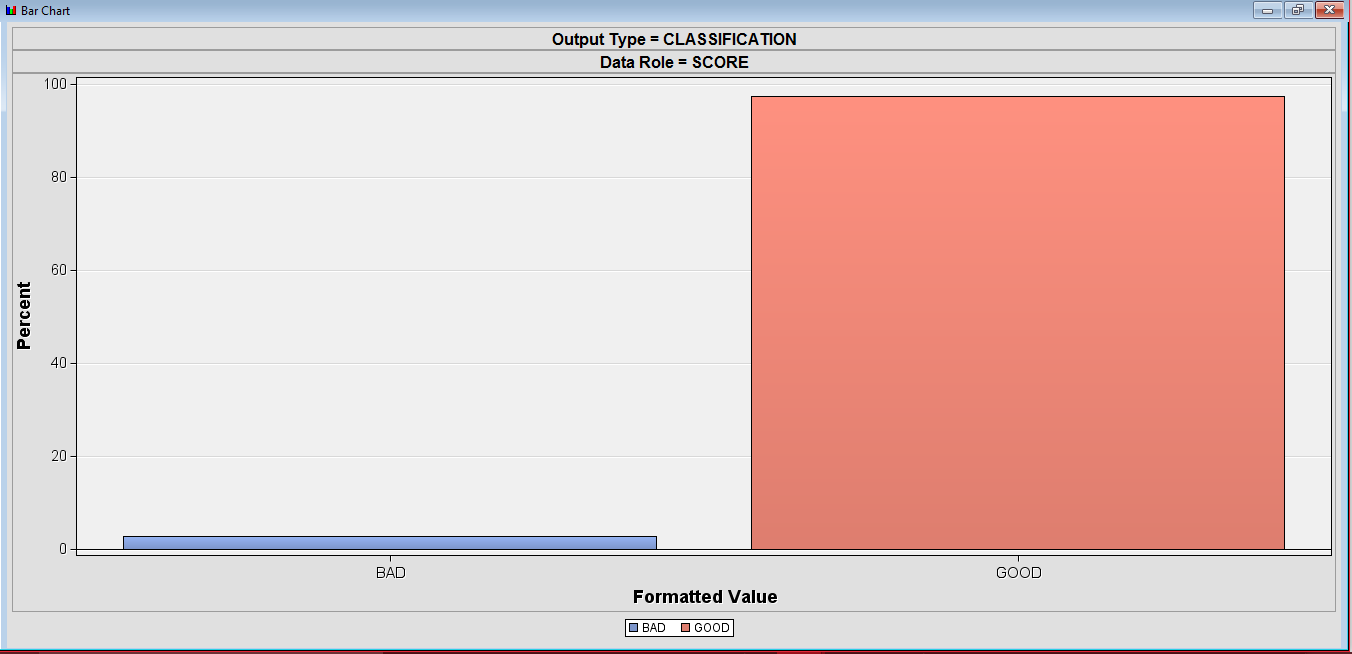
Observing the Cumulative % Expected Profit plot, we may conclude that if we select the 80% of the highest ranked applicants from the validation data set, then the total expected profit is 335897.80 DM when applying the Neural Network, 326777.4 DM when applying Logistic Regression model and 284585.9 DM when applying the Decision Tree model.

The optimal model to pick for scoring new applicants is the neural network model since it is the one that offers the highest cumulative total expected profit.

QUESTION 18



*Figure 18 – Process flow*



*Figure 19 – Bar Chart*

By adding and running the Score node to the diagram, we conclude that the total number of applicants is 75, while the 97.33% among them (73) are predicted as “good” and 2.67% (2) are predicted as “bad” applicants.

QUESTION 19

In order to examine which observation was assigned with the highest and lowest probability of being a good applicant, we drew our information from the exported results’ column “Probability for level GOOD of good\_bad”.

The highest probability of being a good client is equal to 99.75% while the lowest one equals to 36.25%.

The expected profit is simply calculated by using the summing the cell values of “Expected Profit for good\_bad” column, which gives us 57757.34485 DM.

QUESTION 20

The custids of the highest ranked clients (1st bucket) with respect to the probability of repaying the loan to be granted are the following: 160, 682, 211.

Additionally, the custids of the lowest ranked clients (20th bucket) are: 497, 467, 667, 897, 826, 640 and 705.

The .xlsx file with the exported data can be found above:

