German Credit Problem (Question 2)

**1.Data exploration and Visualization:**

**a. Choose 3 numerical predictor variables, compute mean, Standard deviation etc and plot box-plots. Interpret your results**

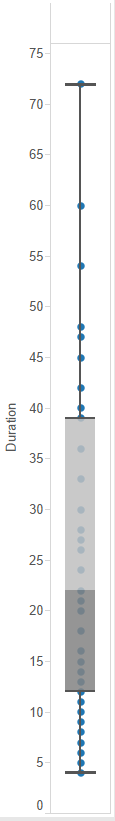
Choosing the three numerical predictor variables: Duration (Duration of credit in months), Age (Age in years) and Amount (Credit amount),

* Duration (Duration of credit in months)

Mean: 20.903

Standard Deviation: 12.05278

Box plot:



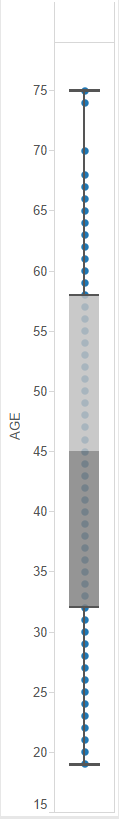
The mean duration of credit is observed to hover around 21 months which is almost close to two years.

* Age (Age in years)

Mean: 35.546

Standard Deviation: 11.36978

Box plot:



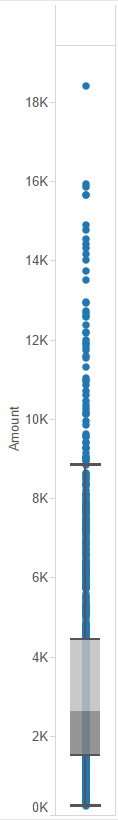
The mean age of the applicants is observed to be around 35 years. A large number is observed in the lower quartile which maybe because the high chance of the taking credit by people who are at the beginning of their careers.

* Amount (Credit amount)

Mean: 3271.258

Standard Deviation: 2821.325

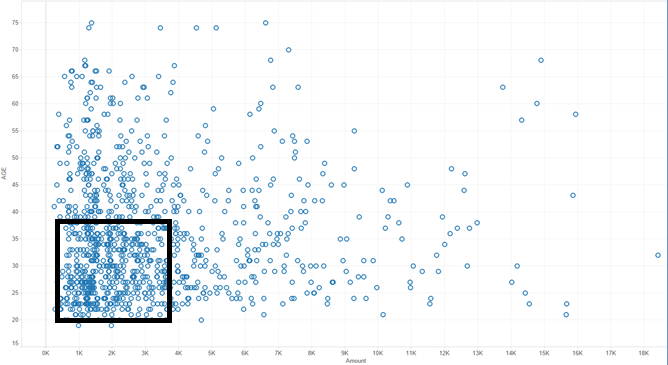
Box plot:



The credit amount of applicants be all over the place with high rate of deviation.

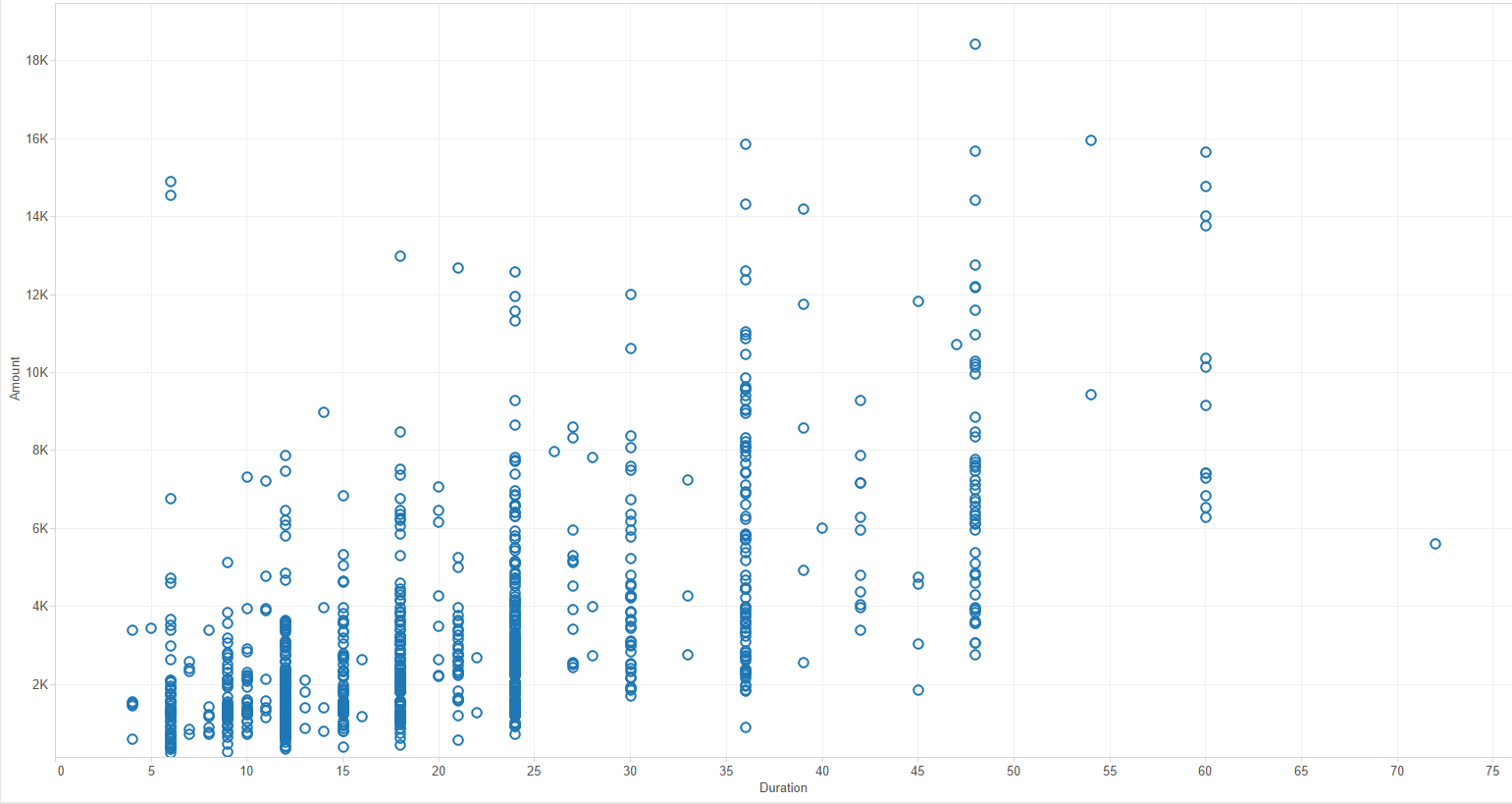
b. Explore scatter plots with the numerical predictor variables. Do you see any relationships?

The following is an observation of a scatter plot between ‘Age’ and ‘Amount’ :



It is observed that a huge chunk of it is between the age of 22-40 and between $1000 to $5000 which is understandable as most people start taking credit from that point of time i.e starting of their careers .

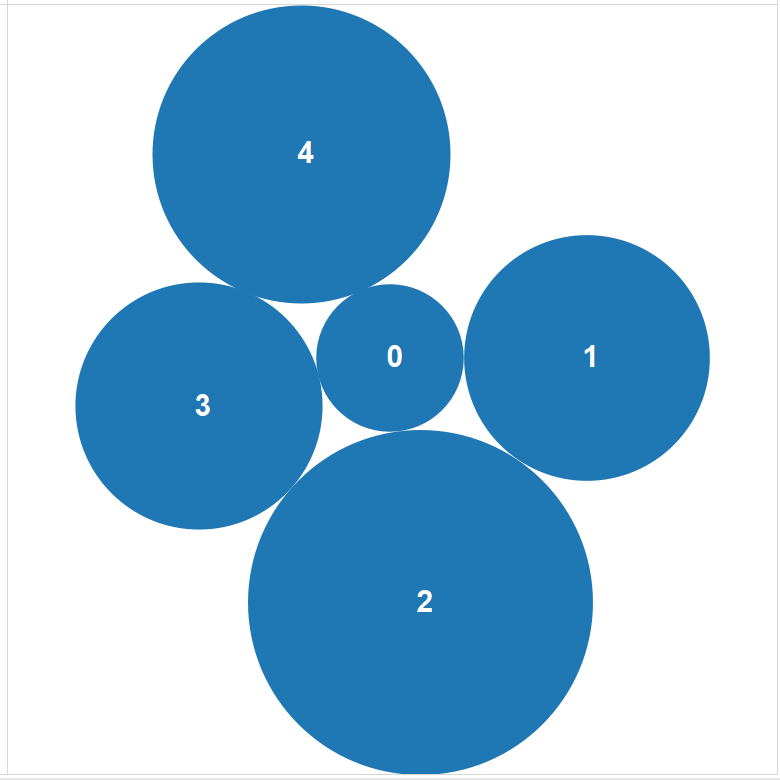
The following is an observation of a scatter plot between ‘Amount’ and ‘Duration’ (in months) :



By observing the graph is it is evident that that usually the duration of the loans(in months) are in factors of 3 i.e. 6 months, 9 months, 12 months etc. Another observation from the graph is that the bigger the amount the more the duration in most of the cases.

c. Choose 3 categorical variables and bucket frequencies of each categorical variable. Use visualization tools to illustrate these variables. Interpret your results

Let us take 3 categorical variables: ‘Employment’, ‘Checking Account’ and ‘History’ and visualize the frequency of each using bubble charts.

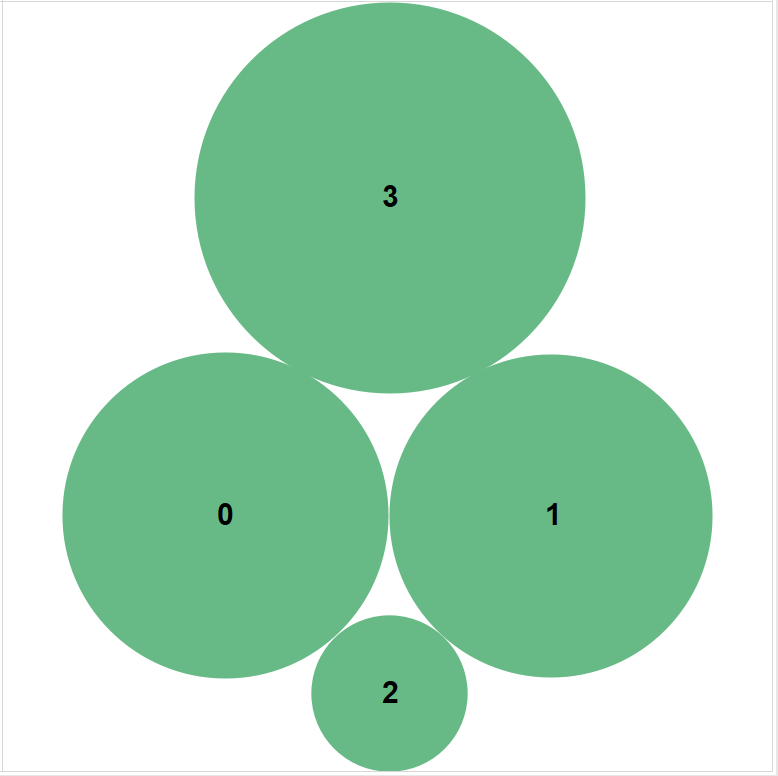
Emloyment (Present employment since):

Legend

|  |
| --- |
| **0 : unemployed** |
| **1: < 1 year** |
| **2 : 1 <= ... < 4 years** |
| **3 : 4 <=... < 7 years** |
| **4 : >= 7 years** |

The top two categories of applicants are observed to be working between 1-4 years and greater than 7 years.

Checking Account

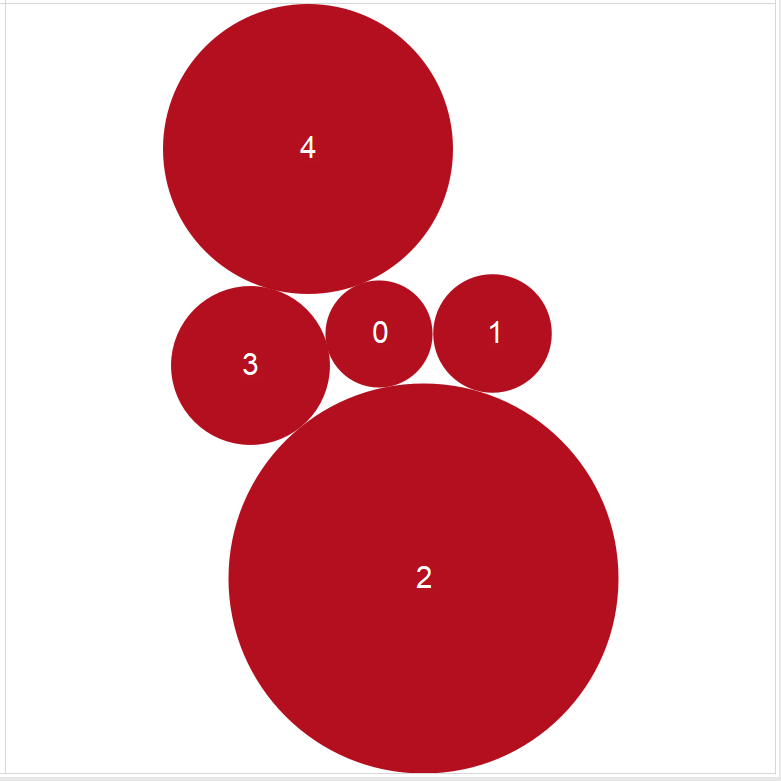


Legend

|  |  |
| --- | --- |
| **0 : < 0 DM** | **0 : unemployed** |
| **1: 0 < ...< 200 DM** | **1: < 1 year** |
| **2 : => 200 DM** | **2 : 1 <= ... < 4 years** |
| **3: no checking account** | **3 : 4 <=... < 7 years** |
|  | **4 : >= 7 years** |

Most of the applicants for credit don’t seem to have a checking account. Even in the presence of a checking account a most of it is empty or less than 200 DM.

History (Credit history)



Legend

|  |
| --- |
| 0: no credits taken |
| 1: all credits at this bank paid back duly |
| 2: existing credits paid back duly till now |
| 3: delay in paying off in the past |
| 4: critical account |

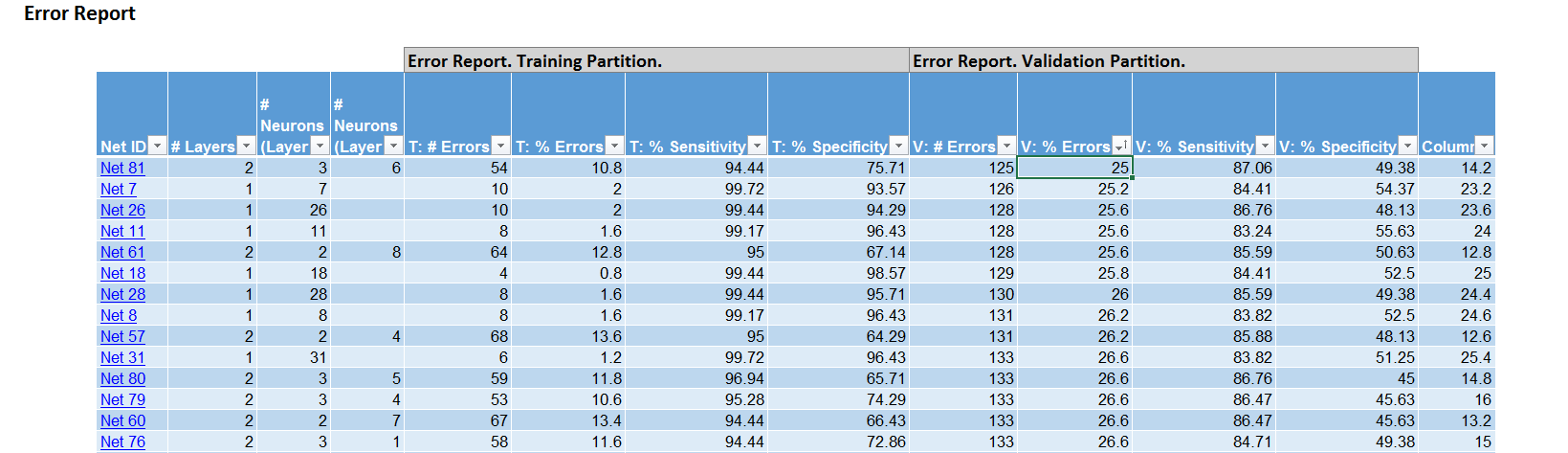
A majority of the applicants seem to have paid their existing credit back duly. Also an alarming number of applicants seem to be in the ‘critical account’ category.

**2. Neural Networks:**

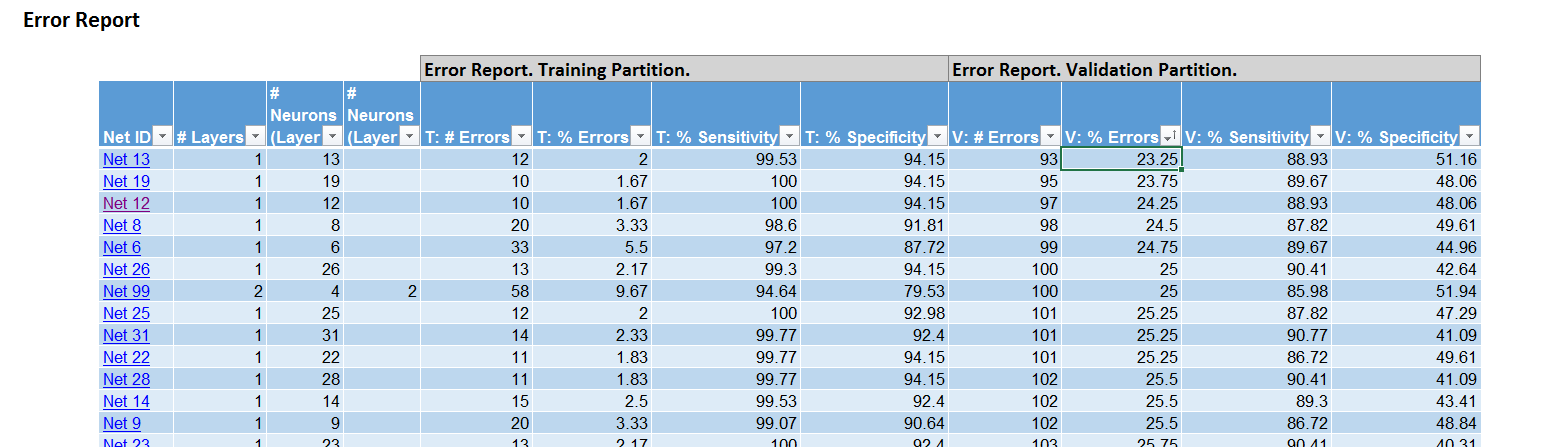
**a. Partition the data into a 50% Training, 50% Validation and 60% Training, 40% Validation data sets and create a neural network. Analyze the results.**

The data was partitioned as 50% Training and 50% Validation dataset and a 60% Training and 40% Validation dataset with default settings for number of epochs, step size, number hidden layers. Both the neural networks were created and analyzed.

The 50-50 dataset’s best Neural Network model showed 25% of Validation errors.



Whereas, the 60-40 dataset’s best Neural Network model showed 23.25% of Validation errors.



In further comparing training data performance, sensitivity and specificity, the 60-40 partition trumps the 50-50 partition again.

**Thus the 60-40 partition on the dataset is the better option.**

**b. Modify the number of layers, step-size for gradient descent, number of epochs etc and recreate models. Interpret your results. What is the significance of number of layers, step-size for gradient descent, number of epochs in a neural network configuration?**

Taking a 60-40 partition on the data with the best neural network, and as default settings:

Number of layers = 1

Step-size for gradient descent = 0.1

Number of epochs = 30

The following observations were made:

Default:

|  |  |  |
| --- | --- | --- |
|  | Training |  |
|  | **Predicted Class** | |
| **Actual Class** | **1** | **0** |
| **1** | 427 | 2 |
| **0** | 10 | 161 |

|  |  |  |
| --- | --- | --- |
|  | Validation | |
|  | **Predicted Class** | |
| **Actual Class** | **1** | **0** |
| **1** | 241 | 30 |
| **0** | 63 | 66 |

**By increasing the number of epochs 50:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Training |  |  |  |  |  | Validation | |
|  | **Predicted Class** | |  |  |  |  | **Predicted Class** | |
| **Actual Class** | **1** | **0** |  |  |  | **Actual Class** | **1** | **0** |
| **1** | 429 | 0 |  |  |  | **1** | 231 | 40 |
| **0** | 8 | 163 |  |  |  | **0** | 61 | 68 |

In comparison with the default model, the misclassification error is improved a little in training but degraded for unseen data / validation. This may be due to overfitting of the data with increasing number of epochs.

**By increasing the gradient step size to 0.3:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Predicted Class** | |  |  |  |  | **Predicted Class** | |
| **Actual Class** | **1** | **0** |  |  |  | **Actual Class** | **1** | **0** |
| **1** | 427 | 2 |  |  |  | **1** | 237 | 34 |
| **0** | 7 | 164 |  |  |  | **0** | 76 | 53 |

In comparison with the default model, there is not a huge difference in training but its performance has got worse in validation.

**By increasing the number of hidden layers to 2:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Training |  |  |  |  |  | Validation | |
|  | **Predicted Class** | |  |  |  |  | **Predicted Class** | |
| **Actual Class** | **1** | **0** |  |  |  | **Actual Class** | **1** | **0** |
| **1** | 427 | 2 |  |  |  | **1** | 242 | 29 |
| **0** | 27 | 144 |  |  |  | **0** | 78 | 51 |

In comparison with the default model, there is not a huge difference in training but its performance has got worse in validation.

c. Your client is new to neural networks. Summarize a note for your client on what this methodology is, and how to interpret results. Explain whether you would recommend this model to your client.

Neural Networks is an analytic technique based on the process of learning in the cognitive system of the brain. There is a ‘learning/training’ stage which is an iterative process where the model forms a network by ‘learning’ from the training/sample data i.e. by inputting the predictor variables and adjusting weights to the network to correctly predict the sample data. The process continues till a ‘fit’ is produced based on the training data. This network is then used to predict new observations.

Using neural networks is a good idea if you are not aware (or don’t need to know) beforehand about the significance of each of the predictor variables and are only bothered about the result. It is a ‘black-box’ method i.e. one can’t find the significance of each predictor variables unlike other methods like MLR, Logistic regression etc. Also there is a huge tendency to overfit the data.

**3. Analysis:**

**a. Your client currently doesn’t have a model and gives credit to whoever approaches him. If he approves credit to everyone, what is the profit for your current data set(1000 records)**

**b. Choose the best neural network you have modeled and compute the profit.**

**c. Which methodology yields the most profit?**

**d. You want to improve your model. What will you do from a modeling perspective ? What additional information can your client provide to help you build a better model?**

Net profit if client gives credit to whoever approaches him = -80000

i.e.: **a loss of 8000 DM**

Find attached excel sheet below for the calculation (Double – click to open the sheet):



**The best neural network** modeled has the following configurations:

Data Partition: 60-40

Step size: 0.1

Number of epochs: 40

Cut off probability: 0.9629

Find below the profit computed for the training and validation datasets:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Training** |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  | **Predicted Class** | | **Profit Calculation** | **Classified Correctly** | **Misclassified** |
| **Actual Class** | **1** | **0** |  | 388 | 212 |
| **1** | 224 | 205 | 22400 |  |  |
| **0** | 7 | 164 | -3500 |  |  |
|  |  | Profit= | **18900** |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Validation** | |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  | **Predicted Class** |  | **Profit Calculation** | **Classified Correctly** | **Misclassified** |
| **Actual Class** | **1** | **0** |  | 245 | 155 |
| **1** | 139 | 132 | 13900 |  |  |
| **0** | 23 | 106 | -11500 |  |  |
|  |  | Profit= | **2400** |  |  |

From a modelling perspective, one would have to keep tweaking and analyzing each of the factors in the neural network in search of the best ‘fit’ or in this case for example-- to get high profit, the rate of misclassification (false positive) should be low.

Additional data that the client can provide to improve models is the critical factors(predictor variables) that they consider in accepting a credit option based on their company policy

**4.**

**Let us try and improve our performance. Rather than accept XLMiner's initial classification of all applicants' credit**

**status, use the "predicted probability of success" in logistic regression (where *success* means 1) as a basis for**

**selecting the best credit risks first, followed by poorer risk applicants.**

**a. Sort the validation on "predicted probability of success."**

**b. For each case, calculate the net profit of extending credit.**

**c. Add another column for cumulative net profit.**

**d. How far into the validation data do you go to get maximum net profit? (Often, this is specified as a percentile or**

**rounded to deciles.)**

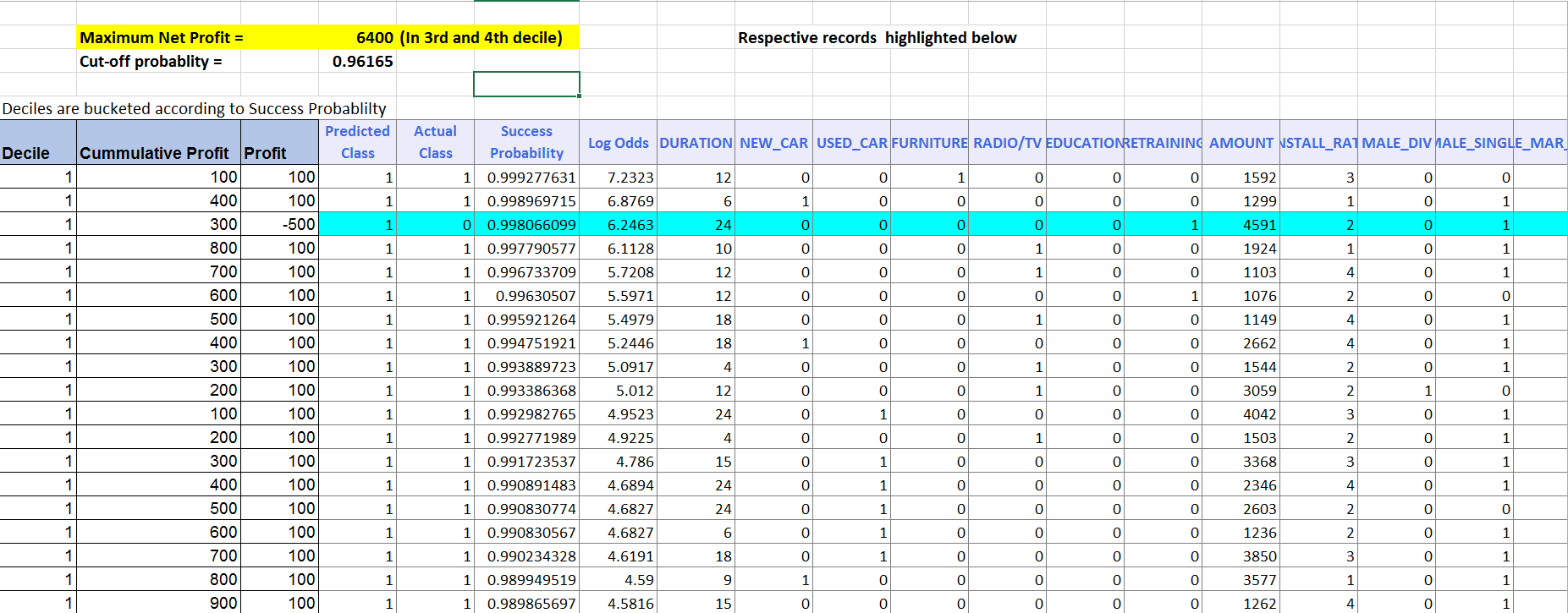
**e. If this logistic regression model is scored to future applicants, what "probability of success" cutoff should be**

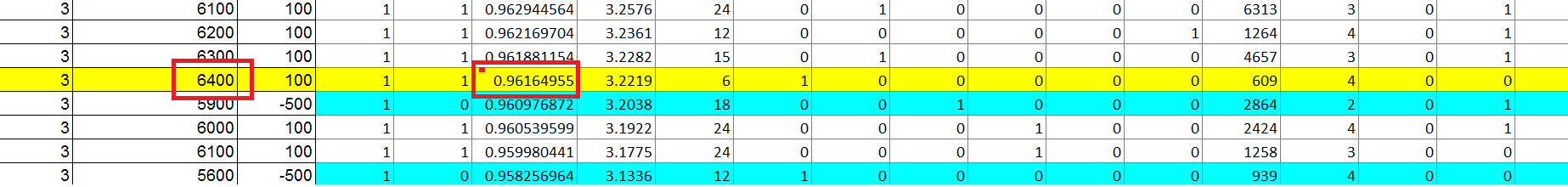
**used in extending credit?**

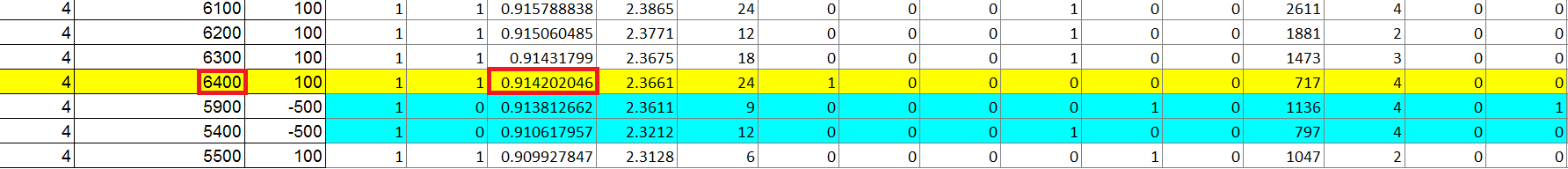
Using logistic regression and steps mentioned above,

The maximum net profit was calculated as 6400 DM found in the 3rd and 4th deciles.

The ‘probability of success’ cutoff used in extending credit is found to be 0.96165 for getting maximum net profit.



****

****

The detailed analysis is found in the sheet embedded below:

(Double – Click to see excel the document)



**5. Executive summary:**

This report consists of an analysis on the German Credit data set. Data exploration and visualizations were performed on the dataset using box plots, scatter plots and bubble charts to get a better understanding of the data.

An analysis of classification of data (Credit worthy or not) using Neural Networks was performed. Different neural network models were tried out to get the best model (like maximizing profit). Various factors considered were the data partition, number of hidden layers, gradient step size and number of epochs.

An alternative analysis using logistic regression was also done to improve the performance by using the ‘predicted probability of success’ and obtaining the ‘probability of success’ cutoff used in extending credit.