**Data Mining and Analysis for Managers**

**(MGMT 635 – Group Assignment #1)**

**German Data Set: Classification problem analysis using Neural Networks**

July 1st, 2018



**Group #4**

**Group Members: Bikalpa Pahari (bp359)**

**Kalpit Khamar (kkk0482)**

**Mona Nandy (mrd2)**

**Sahana Umashankar (su83)**

**Instructor: Dantong Yu and Shibo Yao**

## Table of Contents

**1. Summary..……………………..…………….………………..…………….………………..………..…..3**

**2. Introduction …………………………..……………………..…………….………………..……….…3**

**3. Data Mining Methods…………………………………………………………………………...3**

3.1 Classification…………………………………………………………………………...4

3.2 Clustering………………………………………………………………………………4

3.3 Association Rules………………………………………………………………………4

**4. Information Theory Analysis…………………………………………………………………..5**

**5. Neural Networks…………….…………………………………………………………………..6**

**6. Techniques and Approaches……………………………………………………………………7**

6.1 Business Understanding…………………………………………………………….…..7

6.2 Data Understanding…………………………………………………………………….7

6.3 Data Preparation………………………………………………………………………...8

6.4 Model Building based on Neural Network………………………………………….….9

6.5 Testing and Evaluations………………………………………………………………...10

6.6 Deployment ………………………………………………………………………….….12

**7. Results and Description of Results……………………………………………………………...12**

**8. Conclusions…………………………………………………………………………………….…16**

**9. References………………………………………………………………………………………...17**

**10. Appendix………………………………………………………………………………………...17**

1. **Summary**

This paper presents a data mining project of finding a better accuracy of the credit rating predictions of loan applicants by building a predictive model using neural network algorithm. The objective of the construction of credit scoring models is to prepare and calculate credit risk score to find if customer will likely to default or customer should payback banks to keep their credit worthiness. Using CRISP-DM approach, data is reviewed, preprocessed, before building neural network model. For data preparation, techniques from information theory are used in the process of selecting important attributes, which can be used for neural network algorithm. Data is trained using neural network algorithm in MATLAB. Testing and evaluation are completed for results from 15 neural network trials, using calculated performance indicators such as cost point, accuracy %, error rate, precision and recall rate for each one. In this paper, we also discussed the common data mining methods, techniques and approaches. At the end, we also presented our results, showing which neural network has the best accuracy against the “German dataset” provided.

1. **Introduction**

In today’s digital world, we are overflowing with data. There is a lot of new knowledge to be discovered. This is where data mining plays a key role. Based on Dursun Delen, “Data mining is the process of discovering new knowledge in the forms of patterns and relationships in large data sets.” Discovery of the new knowledge is the first step. But how do we use the new knowledge in problem solving and decision making, are the main objectives. Data mining and analytics has become buzzword in corporate world today. The culture of managerial decision making is changing from intuition-driven to data or evidence based practices. Analytics is not only being incorporated at management level, but it’s being useful at every level. There are already many industries, where data mining techniques and methods are being used extensively. Some of the application areas are customer relationship management, fraud detection, risk mitigation, actuarial estimation, medical research etc.

In this paper, we have discussed data mining methods used in the process, review approaches and techniques and outline our neural network algorithm to get the evaluation score of the last 20 holdout data. A neural network is trained, validated and tested using these selected helpful attributes with data observations to partition the data into training (980 data points) and holdout datasets (last 20 data points).The dataset is pre-processed before building the model to get the efficient predictions on the data. The final model compares the predicted values with the actual test data output to calculate the efficiency of the built model.

1. **Data Mining Methods**

Data mining techniques are set of algorithms intended to find the hidden data knowledge from the dataset. Data mining is the process of extracting specific information from dataset and represent relevant and usable information that can be used to solve problems for analyzation. There are various data mining techniques available today, which are being used for knowledge discoveries, problem solving and decision making. Some of the main techniques are classification, clustering, regression and association rules. Some of these are briefly described in following sections.

* 1. **Classification**

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. Classifications are discrete and do not imply order. Continuous, floating-point values would indicate a numerical, rather than a categorical, target.

To meet the challenges of Big data, which is about volume, velocity, variety and veracity of data, a range of automatic methods for extracting useful information has been developed, among them classification.

In the model build (training) process, a classification algorithm finds relationships between the values of the predictors and the values of the target. Different classification algorithms use different techniques for finding relationships. These relationships are summarized in a model, which can then be applied to a different data set in which the class assignments are unknown. Classification has many applications in customer segmentation, business modeling, marketing, credit analysis, and biomedical and drug response modeling.

There are various classification models available today which include neural networks, k-nearest neighbor, decision tree, Naive Bayes etc. For our project we are considering Naive Bayes and neural networks classification models.

* 1. **Clustering**

Clustering is an analysis conducted to find groups of objects such that objects in group would be similar to each other and different from the other objects in the groups. Having a supervised classification and simple segmentation is not regarded as cluster analysis. For instance, dividing the student by last name is not cluster analysis. Clustering is a set of clusters and mainly there are two types of clustering namely

* Partitional Clustering: It is a divisional data objects into non-overlapping subsets such that each object is in exactly one subset.
* Hierarchical Clustering: This is a set of clusters organized as a hierarchical tree.

There are various distinctions between set of clusters. For instance, clusters can be exclusive or non-exclusive, it can be fuzzy vs non-fuzzy, can be partial vs complete and heterogeneous vs homogeneous. Furthermore, there are various kinds of clusters namely; they are well-separated clusters, center-based clusters and contagious clusters or transitive clusters.

* 1. **Association Rules**

Association rule is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.

It is an implication of X->Y where X and Y are disjoint items. It assumes that all data are categorical, and it is not the best algorithm for numeric data. The basic example of this rule would hover around examples like, customer who buys dolls are 80 percent more likely to buy four kinds of candy bars. Association Rules can be automatically generated, and it represents a pattern in the data without a specified target variable. It is a classic example of undirected data mining. It has its application in various industries, namely retail, banking services etc. The main aspect of association rule is that it defines relationship between two disjoint items that is to say that that two item set cannot have common items. So, the general pattern is that, when X occurs, Y also occurs. Association rules do not represent causality or correlation between two item set. i.e. is to say that:

X->Y does not mean that X causes Y, it only suggests that a strong relationship exists between X and Y.

In the process of Association rule mining, we firstly find the given set of transaction to find the rules predicting the occurrence of item based on occurrence of another item. It mainly uses two indicators to measure the strength.

* **Support = Count (X and Y)/N**
* **Confidence= Count (X and Y)/Count(X)**

To conduct an association mining task, data miner’s goal is to find all the rules having

* Support ≥ minimum support threshold
* Confidence ≥ minimum confidence threshold.

1. **Information Theory Analysis**

Information theory is an important concept of data science. It was invented by Claude Shannon to solve the communication problem. In the data world, to get a solution, we have many variables to consider. How do we know which variable is more important over others? How do we know which is more useful and contains better information to solve the problem? How do we quantify or measure the information in random variables to compare? This is where ‘information theory’ comes in play. Information theory is about quantification, measurement and communication of information.

One of the key factors in information theory is ‘Entropy’. Entropy is the measurement of ‘impurity’. Entropy gives the measurement of “uncertainty” in each variable or the information obtained from a random variable. The higher the entropy, higher is the ‘uncertainty’ and vice versa. Mathematically, ‘entropy’ is calculated as below:

[**Entropy** **= -p1\*log (p1) – p2\*log(p2) + ….** ], where p1 is the probability of attribute 1.

Entropy is only part of the story. When we solve a classification problem, we need to determine which variables are more useful than others. This is part of attribute selection measure. This is to determine how well each attribute splits a set of data, with respect to a chosen target variable. This most common splitting criterion is called “information gain”. It is used to rank the attributes in a data mining process. This can be used simply to understand the data better. Or it can be used to reduce the size of the data to be analyzed, by allowing to select a subset of attributes (based on information gains). Mathematically, information gain can be calculated by:

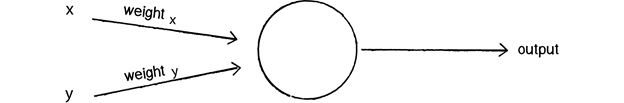
**Information gain = entropy (parent) – [average entropy(children)]**

Information theory has made significant contributions in applications such as cryptology and linguistic fields among many others. For our German dataset classification problem, we have also taken advantage of the information theory to reduce the size of the data set by selecting attributes with higher information gain for our analysis using neural network algorithm.

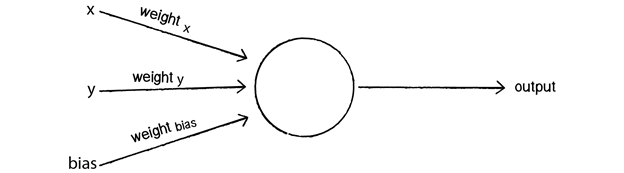
1. **Neural Networks**

Neural Network algorithm is an implementation of the popular and adaptable neural network architecture for machine learning. Input data is passed through input layer and obtain result by activating the output layer. The algorithm works by testing each possible state of the input attribute against each possible state of the predictable attribute, and calculating probabilities for each combination based on the training data. You can use these probabilities for both classification or regression tasks, to predict an outcome based on some input attributes. A neural network can also be used for association analysis. In order to get the accurate result from neural network, the model can be trained and tested several times by changing the number of layers and neurons.

Neural network method learns information from the data by calculating and modifying the weights that are associated with neural networks.



We can see how there are two inputs (x and y), a weight for each input (weight x and weight y), as well as a processing neuron that generates the output.



0 \* weight for x = 0

0 \* weight for y = 0

1 \* weight for bias = weight for bias

**Why we chose Neural Network:**

* A neural network model supports regression, association, and classification analysis, Therefore, the meaning of each prediction might be different.
* We can also query the model itself, to review the correlations that were found and retrieve related statistics.
* Neural Network is nonlinear model that is easy to use and understand compared to statistical methods.
* Neural Network with Back propagation (BP) learning algorithm is widely used in solving various classifications and forecasting problems.

1. **Techniques And Approaches**

For data mining projects, one of the standard methodologies used in the industry is CRISP-DM. CRISP-DM stands for Cross-Industry Standard Process for Data Mining. This methodology provides a structured approach to planning a data mining project in six steps. The process starts with a good understanding of business problem and ends with the deployment of the solution that fits the specific business needs or problems. Since data mining is driven by experience or experimentation, depending on the problem, the steps can be very iterative. Sometimes it can be very time intensive.

For our classification data mining project, our objective was to evaluate the performance (in terms of accuracy) and try to improve it by using different parameter set in Neural Network algorithm. We followed the CRISP-DM methodology, step by step, to achieve the solution.

* 1. **Business Understanding:**

Business understanding is the first step in CRISP-DM. With any problem solving project, the first step is to understand the problem. Same for data mining projects, the success depends on understanding what we are trying to solve. In this step, managers specify the business objectives and a project plan is developed on who and how these information / goals would be attainted. The high level budget is also established during this step.

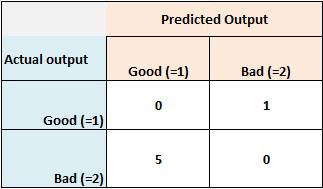
For our data mining project, the main objective of the project is to find a better accuracy of the Credit rating predictions of loan applicants by building a predictive model based on supervised learning. The project also requires us to use a neural network algorithm to find the better accuracy using the given dataset.

* 1. **Data Understanding**

Data understanding, the next step in the CRISP-DM process, is a critical step. In real-world data mining project, analyst will gather or identify the relevant data set from different data sources, such as demographic, sociographic, transactional etc. among many other sources. They use different variety of statistical and graphical techniques to measure the variables which can be used in data mining algorithms to find useful knowledge patterns.

For our project, a German data set is provided to us with 1000 data records. Each record or row has information about one customer’s loan application. Each applicant is described by 20 attributes. The target column was included in the dataset, which has information such as which applicant is a “good” credit risk versus “bad” credit risk. The bad applicants are assigned a value of “2” and good applicants are assigned a value of “1”. In total, there are 700 “good” applicants and 300 “bad” applicants in the dataset. Among all the attributes, there are 17 categorical (qualitative and quantitative) and 3 continuous attributes. .

As part of the data understanding, we also learnt about the cost matrix guidelines provided in determining if our prediction is correct or incorrect, then how much weight can be assigned to the cost. In summary, if the predictions are correct (good or bad), the count is ‘0’. If predictions for good applicants are ‘bad’, then a count of 1 is assigned to each prediction. And the most risky scenarios, where predictions for ‘bad’ applications are ‘good’, then a count of 5 is assigned to that prediction. Following is the summary of the cost matrix for our project:



The total count of the cost points is calculated for each neural network trials. The lower the total cost points, the better our neural network would be.

* 1. **Data Preparation**

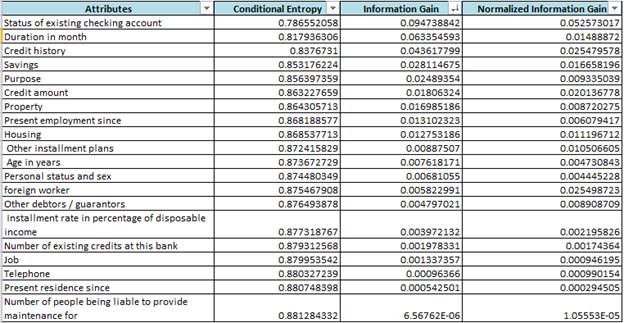
The purpose of data preparation or processing is to prepare the data which we identified in previous step for analysis using different data mining technique. This step is the most effort and time consuming. The real-world data is mostly incomplete and noisy and needs cleaning, reformatting, elimination etc before it can be used for analysis purposes. Data preparation also includes the selection process of attributes to be used for the mining algorithm.

In our project dataset, the data is clean. There isn’t any missing value or outlier. But we did performed few steps to make our data more informative and we went through the process of attribute selection. Attribute selection is done using the domain knowledge and Information gain calculation. First, we identified, which attributes are categorical and which ones are continuous. We calculated entropy, information gain and normalized information gain to identify the best set of attributes to be used in our analysis. For detail calculations and summary, we have included the workbook below.



\*Appendix has detail about this excel file “NeuralNetwork\_Trials\_All Results Final”

We calculated the entropy, information gain and normalized information gain for categorical attributes first. To calculate these measurements for continuous attributes, we used binary search method. As mentioned in previous section, information gain is an important factor in information theory, which helps to quantify the information each attribute carries. We summarized all 20 attributes’ conditional entropy, information gain and normalized information gain, ordered by “information gain” values, highest to lowest. This summary helped us in identifying which attributes are important and preferred in algorithm.



Next, we prepared our ‘Target’ data. We converted the values to binary values. We assigned ‘good’ applicants as ‘0’ (previously, it was ‘1’) and ‘bad’ applicants as ‘1’ (previously, it was ‘2’).

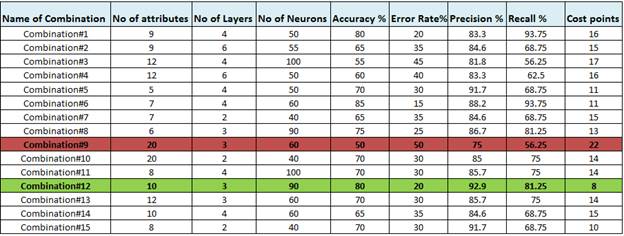
Last part of our data preprocessing was to split the 1000 dataset into training set and testing set. Based on our project, we kept 980 dataset for training and 20 dataset for testing and validation.

* 1. **Model Building Based On Neural Network**

In fourth step of CRISP-DM methodology, techniques and algorithms are selected and applied to a preprocessed dataset in order to extract the information which business need. Modeling step includes the validation and performance analysis. For this phase, various tools are used. Some of these are MATLAB, Python, R, and SAS among many others. There can be many different techniques or algorithm used to solve problems.

Based on our project, the model we used is Neural Networks algorithm in MATLAB. MATLAB provides already coded algorithm to create and train the data. It also simulates and validates the output. The purpose of this data mining project is to try different parameters such as number of attributes, number of layers and number of neurons to analyze if we can improve the accuracy of the predictions. We used 980 dataset for training data. We separated “TrainingInput” and “TrainingOutput” to train the data. We kept 20 “HoldoutInput”(Testing Input) and “HoldoutOutput”(Testing Output) for validation purposes, which we will discuss in next section.

In total, we tried 15 different combinations with various different parameters as shown in the table below. The attribute selection was mostly based on the value of information gain (the higher the value, the higher the priority of that attribute). But in few trials, we also used our domain knowledge and thought about what attribute would have higher importance in this model regardless of information gain values. Besides number of attributes, we experimented with different number of layers and number of neurons to see which combination generates the output with best performance.



This table contains the combinations of each trial as well as the performance and calculated results which we will discuss in next section.

* 1. **Testing And Evaluations**

In the fifth phase of CRISP-DM, we evaluate the model for accuracy and generality. This is a critical and challenging step. This is to test how good is our model and the parameters used. Data analysts need to understand the data mining objectives, and how it will all fit in with the business objectives. In order to properly analyze the results, confusion matrix is used.

**Confusion matrix:** The confusion matrix is a two by two table that contains four outcomes produced by a binary classifier. Various measures, such as error-rate, accuracy, sensitivity, and precision, are derived from the confusion matrix.



**Fig: Confusion Matrix**

**Error-rate:** Error rate (ERR) is calculated as the number of all incorrect predictions divided by the total number of the dataset. The best error rate is 0.0, whereas the worst is 1.0.

**Error rate: (FP+FN)/ (TP+FN+FP+TN)**

**Accuracy:** Accuracy (ACC) is calculated as the number of all correct predictions divided by the total number of the dataset. The best accuracy is 1.0, whereas the worst is 0.0. It can also be calculated by 1 – ERR.

**Accuracy: (TP+TN)/ (TP+FN+FP+TN)**

**Precision:** Precision (PREC) is calculated as the number of correct positive predictions divided by the total number of positive predictions. It is also called positive predictive value (PPV). The best precision is 1.0, whereas the worst is 0.0.

**Precision : TP/ (TP+FP)**

**Recall:** Recall or Sensitivity (SN) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called Sensitivity (SN) or true positive rate (TPR). The best sensitivity is 1.0, whereas the worst is 0.0.

**Recall/Sensitivity: TP/ (TP+FN)**

For our project, using Neural Network in MATLAB, we simulated the prediction model on testing “Holdout Input” dataset. With each trial, we repeated this step to capture the predicted output. We rounded the predicted output values (20) into 0’s and 1’s. We added predicted results and actual results (which are from given dataset) side by side to conclude the classification type. Here are the classification guidelines that we used:

|  |  |
| --- | --- |
| **Classification Type** | **(Predicted, Actual)** |
| True Positive (TP) | (0, 0) |
| True Negative (TN) | (1, 1) |
| False Positive (FP) | (0, 1) |
| False Negative (FN) | (1, 0) |

Based on the count of each classification type, we calculated **cost points, accuracy%, precision%, error rate% and recall%**. The worksheet contains the detail information about each trial (combinations) we analyzed. Our goal was to identify the best neural network with respective parameter combinations (attributes, layers and neurons). Here are the mathematical calculations we used to derive these performance indicators:

* + - Accuracy % = (TP + TN) \* 100 / (TP + TN + FN + FP)
    - Precision % = (TP \* 100) / (TP + FP)
    - Error Rate % = (FP + FN) \* 100 / (TP+TN+FN+FP)
    - Recall (sensitivity) % = TP \* 100 / (TP+FN)

where TP, TN, FP FN indicates the count of these classification type output. We summarized the result, which is described in result section.

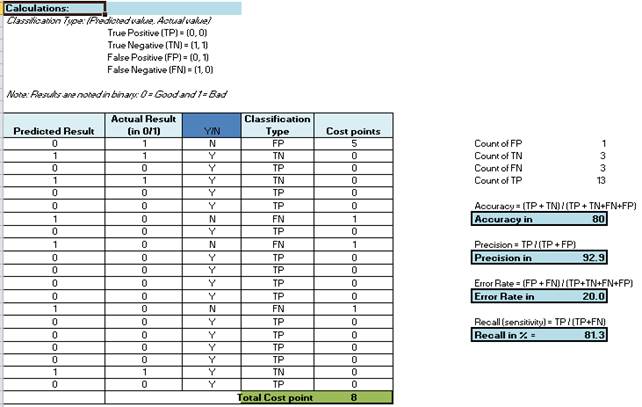
* 1. **Deployment**

After testing and evaluation, comes deployment phase. The knowledge gained from this data mining exploration needs to be organized and presented in a way, so that end user would be able to benefit from it. Sometimes, it can be generating a report or other times, it can be repeating the data mining process on enterprise level.

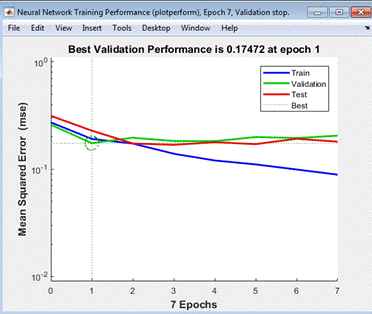
For our project purposes, we have put together this report of our data mining project from business understanding of the problem to concluding the results.

1. **Results And Description Of Results**

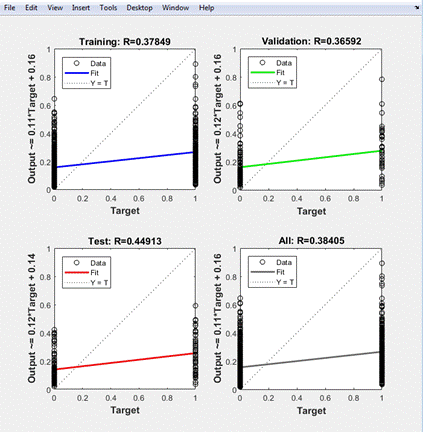
Based on the table compare for each trial and our analysis, we found that “Combination#12” is the best neural network we could come up with. For this combination, we used 10 attributes with highest information gain (Status, Duration, Credit history, Savings, Purpose, Credit Amount, Property, Present employment, Housing, Other installation plan) 3 layers and 90 neurons. Even though “Accuracy%” was best 85% for “Combination#6”, we took other performance indicators and factors into consideration. We looked for the neural network and parameter sets with lowest cost point. Based on the cost matrix described in previous section, the “False Positive” outcomes are more risky and it is assigned a cost point of 5. Thus, the “Combination #12” neural network with cost point of 8 seems to be the best one and also a lot less risky than other trials. With lowest cost point, the accuracy was also really good at 80% of correct predictions.



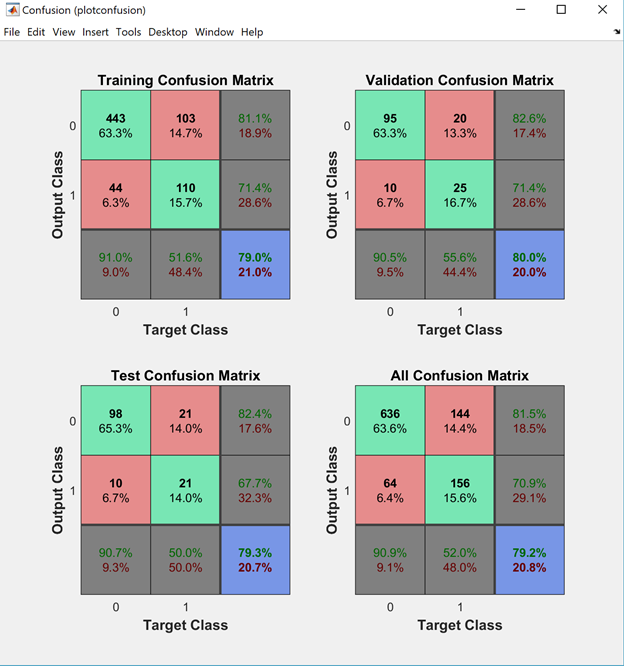
For our best neural network, we looked at the performance and validation graph as well. This is our best performance from epoch with the lowest validation error.



The second graph is the regression model created after training the data. The blue line shows the declining error on training data. The green line indicates the validation error and training stops when the validation error stops decreasing. The red line indicates the error on the test data. It basically says how well this neural network will generalize to new data outside of this given dataset.

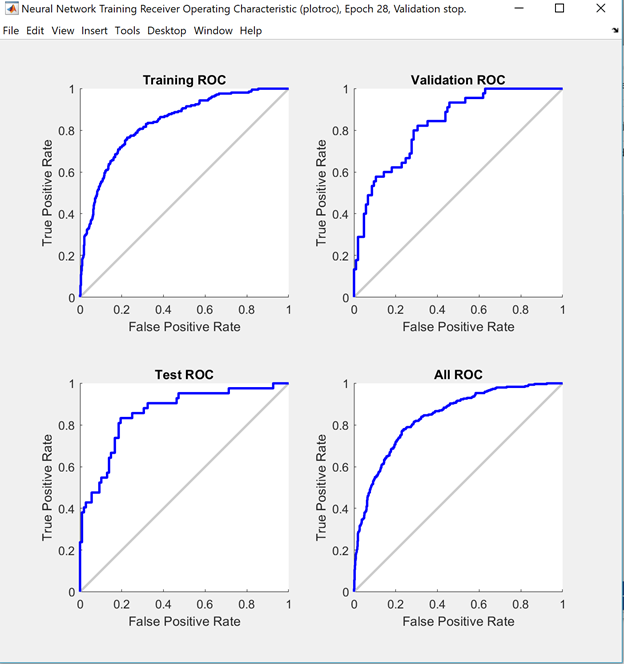


Classification of good and bad customer of German Credit dataset is shown in confusion matrix using the neural network pattern recognition tool (nprtool).



Proper classifications of customers are shown in green squares and improper classifications of customers are shown in red squares.  Total percentages of proper and improper classifications are shown in blue square.

The Receiver operating characteristic plot shows the percentage of True positive rate we get as a function of how many false positive rates we are willing to accept.



1. **Conclusion:**

Application of data mining algorithms on full scale implementation is cost effective, less error-prone and less time consuming. Application of data mining algorithms in banking sector is gaining momentum. It requires more research and exploration. It has high potential to cater to diversifying needs in the banking sector.

In this project, we were presented with a classification problem of credit rating predictions for loan applicants. We were tasked to train and test Supervised learning models for Credit Risk prediction. The dataset consisted of all the customers transactional and demographic information (Independent features) and defaulter column (target/dependent feature). We followed CRISP-DM methodology to complete this data mining project. We started with business understanding of the problem until the deployment (summarizing this report). Our model was built in neural network in MATLAB. We did 15 trial and error based analysis and compared the performance. We selected our best neural network based on the lowest cost point with optimal accuracy and precision rate. Thus, Neural Networks classification algorithms have classified the dataset to good accuracy thus achieving the objective of this study.

1. **References:**

[1] Dr Rajni Jain, “Introduction to Data Mining Techniques”

<http://www.iasri.res.in/ebook/expertsystem/datamining.pdf>

[2] Delen, Dursun, “Real-World DATA MINING”

[3] Information theory

<https://en.wikipedia.org/wiki/Information_theory>

[4] Basic evaluation measures from the confusion matrix <https://classeval.wordpress.com/introduction/basic-evaluation-measures/>

1. **Appendix:**

All tab explanation for excel file “NeuralNetwork\_Trials\_All Results Final”

* Table of Contents - German Data Set: Neural Network Training Data and Valuation
* German Data Set - 1000 ( This data set contains all data attributes and one target attribute column )
* Training Data - 980 ( This data set contains all training attributes and one target attribute column (in binary values).)
* Holdout Data - 20 ( This data set contains all testing attributes and one target attribute column (in binary values).)
* Entropy and IG - catego attr (Entropy, Information Gain and Normalized Information gain - Categorical Attributes)
* Continuous - Finding Split (Finding best splits for Continuous Attributes)
* Continuous - Entropy & IG (Entropy, Information Gain and Normalized Information gain - Continuous Attributes)
* Summary of Entropys, IGs, NIGs (Entropy, Information Gain and Normalized Information gain - All Attributes)
* Summ of Neural N - 15 Trials (Neural Network Classification problem solving - combinations, training, validation and performance for 15 trials)
* Combination#1 ( Neural Network "NN1": Paramter combinations, accuracy, precision, cost point and graphs)
* Combination#2 (Neural Network "network1": Paramter combinations, accuracy, precision, cost point and graphs)
* Combination#3 (Neural Network "NN3": Paramter combinations, accuracy, precision, cost point and graphs)
* Combination#4 (Neural Network "NN4": Paramter combinations, accuracy, precision, cost point and graphs)
* Combination#5 (Neural Network "NN5": Paramter combinations, accuracy, precision, cost point and graphs)
* Combination#6 (Neural Network "Case2": Paramter combinations, accuracy, precision, cost point and graphs)
* Combination#7 (Neural Network "Case4": Paramter combinations, accuracy, precision, cost point and graphs)
* Combination#8 (Neural Network "Case5": Paramter combinations, accuracy, precision, cost point and graphs)
* Combination#9 (Neural Network "Case1": Paramter combinations, accuracy, precision, cost point and graphs)
* Combination#10 (Neural Network "Case3": Paramter combinations, accuracy, precision, cost point and graphs)
* Combination#11 (Neural Network "GC84100": Paramter combinations, accuracy, precision, cost point and graphs)
* Combination#12 (Neural Network "GC10390": Paramter combinations, accuracy, precision, cost point and graphs)
* Combination#13 (Neural Network "GC12360": Paramter combinations, accuracy, precision, cost point and graphs)
* Combination#14 (Neural Network "GC10440": Paramter combinations, accuracy, precision, cost point and graphs)
* Combination#15 (Neural Network "GC8240": Paramter combinations, accuracy, precision, cost point and graphs)