Neural Network Analysis Utilizing German Credit Dataset

MGMT 635 SPRING 2017

22 FEB 2017

Team Alpha

Group Project 1

Matthew Eustice
Mujahidul Islam
Sharlene Mak
Gabrielle Mack

Neural Network Analysis Utilizing German Credit Dataset

Team Alpha – Group Project 1 - MGMT-635 – 22 FEB 2017

Matthew Eustice
Mujahidul Islam
Sharlene Mak
Gabrielle Mack

***Abstract**: The objective of this paper is to choose an appropriate data mining methodology for analysis of a credit worthiness predictor data set. Our team chose the CRISP-DM methodology utilizing a neural network and then compared its results to an existing research project using the same methodology. The result was an improved score prediction without the use of feature selection. This paper provides a detailed overview of alternative data mining methodologies combined with the results of our neural network in comparison to the results discovered by the prior research team.*

Part I. Overview of Data Mining and Synopsis of Project

In a world that is driven by consumerism, businesses must rely heavily on understanding consumer behavior in order to create valuable and desirable products. The first step in understanding how to appeal to your target audience is to collect various sorts of data that the business believes has any correlation to their revenue. This data can come from industry surveys, research gathered from social media conversations, competitive data, market trend analyses, historical data from the business, and other pertinent sources. After obtaining the data, the next step would be to break down the data accordingly to draw conclusions. However, the process of breaking down multiple data sets into useful information is not an easy step. It is cumbersome, time consuming, and often times extremely complicated to draw patterns from multiple factors. Thus, the process of data mining was created to make predictions that older analysis techniques were not capable of making.

Data mining can be used in many industries to predict and discover trends. It is defined as the process of analyzing large data sets to turn the data into useful information. With a large data set, it is nearly impossible to discern any useful information without breaking down the data using computational methods, statistical analysis, database systems, and other techniques. This technique is a game changer in the world of statistical analysis and business because it creates a powerful tool for businesses in the competitive market. By utilizing data mining techniques, businesses can make informed, more effective and proactive decisions that will help the company grow. This process does not require an experienced statistician, but rather empowers businesses themselves to utilize this quicker and easier process to draw conclusions.

For our project, we have utilized the application of Artificial Neural Networks (ANN) to analyze the German Credit Data Set made up of 1,000 observations, or loan applicants. The credit dataset was comprised of a set of 20 attributes with seventeen discrete attributes and three continuous attributes. The Neural Network has been used in many business applications for pattern recognition, forecasting, prediction, and classification. It represents a brain metaphor for information processing and has exemplified their ability to “learn” from the data and generalize patterns. Artificial Neural Networks use computational methods to replace the biological components of a neuron with artificial components.

In the biological setting, a neuron is comprised of a nucleus, dendrites, axons, and a synapses. In the human brain, a group of neurons is known as a network. These networks are made up of highly interconnected neurons, creating a collection of neural networks which process an extensive and diverse amount of information. This is how the human brain is capable of controlling the central nervous system which is responsible for storing intelligence and controlling thinking. Comparatively, an artificial neural network mimics the function of processing large amounts of data into intelligible pieces of information that can be utilized to make smart, data-driven decisions. In an artificial neural network, there are nodes, inputs, outputs, weight, and fast speed to drive machine learning and information processing.

While biological neural networks can have multiple layers of neurons, Artificial Neural Networks are comprised of input and output layers. Learning can be achieved by modifying the set of weighted edges and nodes. The method that was utilized in our project consisted of partitioning the data set into a training model, otherwise known as the training set, and a testing set. The network was repeatedly tested and modified to reach our desired output. In order to train the network, the errors in the output are fed back into the network, which causes the nodes to rearrange. As the article published by the National University of Ireland describes, the error at the output layer is used to "re-modify" the weights coming out of the output layer. We utilized all twenty attributes for the creation of the neural network and repeatedly tested and modified the inputs to reach our desired output. This process is described in further detail in the later section of this paper.

Part II. Overview of Data Mining Methods

There are several data mining methods that have been developed and proven to be efficient and effective. The following are commonly utilized techniques:

Association: This is one of the most recognized data mining methods. It involves discovering a pattern based on the relationship of items in the same set of items. There is usually a given relation in which the goal is to extract a set of products that are generally associated with each other. This is otherwise known as the *relation technique* and is frequently used in "market basket analysis". This is a popular technique in the retail industry to uncover hidden patterns for recommending new products to others based on prior purchases or products which were purchased simultaneously.

Classification: This method utilizes a technique based on machine learning. This is a systematic process used to define the data into sub classes or groups by obtaining important and relevant information about data and metadata. The data can be classified into different categories by utilizing mathematical techniques which often times includes decision trees, linear programming, Neural Networks, and statistics.

Clustering: This method identifies the similarities in data sets while understanding the differences within the data in order to cluster the data into groups. This automatic process defines the classes and puts objects in each class, which are usually predefined. Clusters have certain traits in common which can be used to improve targeting algorithms. There are a variety of clustering algorithms which varies significantly based on the definition of the cluster in the algorithm.

Prediction: This method identifies the relationship between independent variables and also the relationship between dependent variables. The difference between classification and prediction is that classification is more commonly used to classify existing data while prediction is used to classify new data.

Part III. Data Mining Process Methodologies

Among many other computational paradigms, knowledge gains from large data began as experimental projects. Data mining projects were carried out as artistic experimental endeavors. Practitioners have looked at the problem from the perspective of trying to characterize what works and what doesn't work. To methodically conduct data mining analysis, a standardized process needed to be developed and followed. Data mining researchers and practitioners proposed several processes & workflows for the success in conducting data mining projects. All the efforts led to several standardized data mining processes. [Delen, pg, 67] An overview of each methodology is described below:

KDD – Knowledge Discovery in Databases: The term KDD refers to the broad process of finding knowledge in data, and emphasizes the high level application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization. The unifying goal of the KDD process is to extract knowledge from data in the context of large databases. In the KDD methodology, data mining is a single step where the patterns are extracted from data. Figure [Class mod 2-1, slide 3] shows KDD process flow:

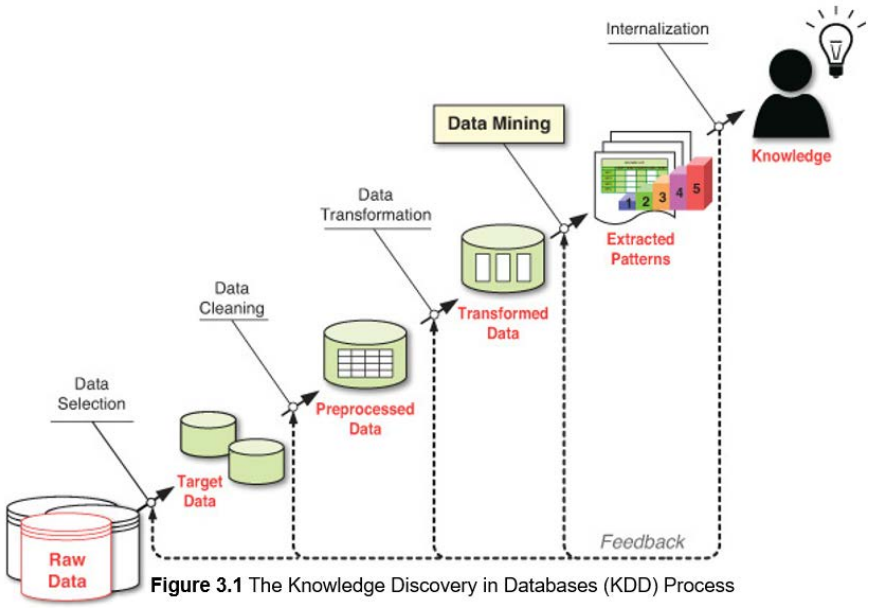


Figure 3.1 The Knowledge Discovery in Databases (KDD) Process

SEMMA – Sample, Explore, Modify, Model, Assess: This methodology focuses on the modeling tasks of data mining projects in order to be applied successfully. SEMMA, developed by the SAS Institute, stands for Sample, Explore, Modify, Model, and Access. The intention is to make it easy while applying exploratory statistical and visualization along with other techniques, and finally confirm the model's accuracy. [Delen, pg,78] A pictorial representation of SEMMA is below: [class mod 2-1, slide-3]

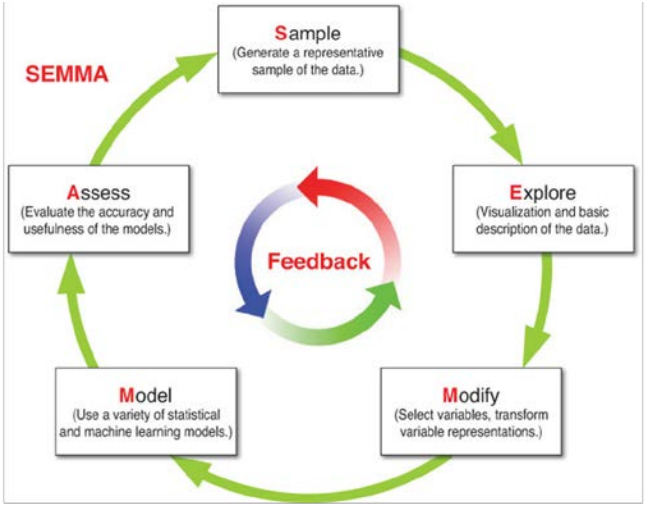


Figure 3.3 Schematic of SAS Institute's SEMMA

There are five steps for SEMMA process:

1. *Sample*: During this step, large amounts of data are gathered into samples and a strategy is devised for manipulation. The sample then undergoes additional training, validation, and testing to detect outliers in the data that could potentially skew the end results of the model.

2. *Explore*: During this step, the analyst searches for unanticipated trends within the data set. Anomalies are detected and processed accordingly and trend analysis allows clear patterns to be distinguished.

3. *Modify*: In this step, an analyst would perform grouping of data into significant subgroups and may also modify when necessary. This step is iterative and involves refining the data set and any attributes that may or may not be relevant.

4. *Model*: After the dataset has been thoroughly analyzed, an appropriate model is selected to utilize in order to find the desired outcome. Choices between the use of Artificial Neural Networks, Decision Trees, or Statistical analysis, and various other modeling options are completed in this step.

5. *Assess*: Finally, the results of the modeling phase are analyzed. Assessment involves a portion of data that was selected during sampling, and reviewed for use in modeling. The same model can be tested against other data sets for accuracy.

SIX SIGMA: Six Sigma is a set of techniques and tools for process improvement. It seeks to improve the quality of the output of a process by identifying and removing the causes of defects and minimizing variability in manufacturing and business processes. Six Sigma methodologies have manifested itself in the business world with DMAIC, expressed as Define, Measure, Analyze, Improve, and Control. [Delen, pg., 83]

The figure below illustrates the DMAIC methodology as a simplified flow diagram. [Class mod 2-1 Slide- 7]

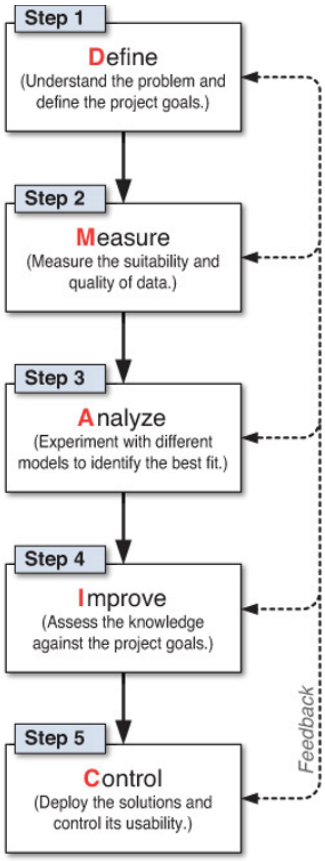


Figure 3.4 Six Sigma DMAIC Methodology

1. *Define:* The first step involves defining the scope of the project and gathering business requirements. During this step, goals and objectives for the project are established and a detail project plan is developed.

2. *Measure:* The second step involves assessing the mapping between organizational data repositories and the business problem. This step requires that data from multiple sources is consolidated into one format that is apparent to machine learning.

3. *Analyze:* The third step involves modeling against the data that has been gathered. Various techniques are utilized in this step to determine which model satisfies the business need and project scope.

4. *Improve:* The fourth step involves improvement against the modeling techniques used. Results are analyzed and modifications to the model are made to fine tune performance.

5. *Control:* The final step of DMAIC process involves controlling and disseminating the results to appropriate stakeholders and evaluating the results.

Part IV. CRISP-DM

The Cross Industry Standard Process for Data Mining [CRISP-DM] is arguably the most one of the popular processes for data mining. According to a 2007 KDNuggets poll, 43% of respondents utilized CRISP-DM as their preferred methodology. Our team chose this methodology because of the processes robust and effective implementation.

CRISP-DM is a comprehensive data mining methodology and process model that provides anyone from novices to data mining experts with a complete blueprint for conducting a data mining project. CRISP-DM breaks down the life cycle of a data mining project into six phases. Figure 3.2 below illustrates six-step standardized process.

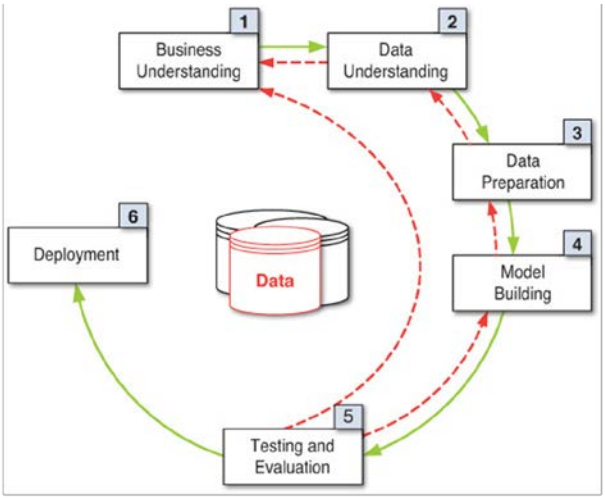


Figure 3.2 The CRISP-DM Data Mining Process

Each step in CRISP-DM is built upon its predecessor allowing the methodology to follow an iterative implementation when necessary. This allows the model to have a certain amount of flexibility which made it an ideal choice for our project in analyzing the German Credit Data Set.

Part V: Steps of CRISP-DM

Business Understanding: Fully comprehending the scope of the project and the need for data mining within an organization is crucial. As such, the first step in the CRISP-DM is business understanding. This step includes a thorough analysis of the business case for which the model will be used. Information such as current business processes, specific goals for the model, and feasibility for implementation are completed at this step.

In the case of the German Credit dataset, the data set was used to determine the fiscal viability of candidates applying for credit. The model was created to establish a predictor score for consumers who may or may not qualify for credit based on an institutions' standards. The intent is to use the predictor score to better assess candidates for which credit can be extended with minimal chance of default.

Data Understanding: Once the business case for the data mining project is established, sources for the data which will be used must be identified and analyzed. This step is known as data understanding. Data can be generated from multiple sources in various formats. Identifying key stakeholders and processes for which the data is generated must also be understood. Attribute and variable selection for the model occurs at this step as well as verifying data integrity.

The German Credit Dataset provided a list of 20 attributes to be used for analysis in generating a prediction score. From this data set, key attributes can be selected for use in the data model.

Attribute ID	Description
1	Status of existing checking account
2	Duration in month
3	Credit history
4	Purpose
5	Credit amount
6	Savings account/bonds
7	Present employment since
8	Installment rate in percentage of disposable income
9	Personal status and sex
10	Other debtors / guarantors
11	Present residence since
12	Property
13	Age in years
14	Other installment plans
15	Housing
16	Number of existing credits at this bank
17	Job
18	Number of people being liable to provide maintenance for
19	Telephone
20	Foreign worker

Data Preparation: Data obtained in various formats needs to be standardized prior to being integrated into the model. This step is known as data preparation. Depending on the source, raw data will need to be aggregated and converted to a format which is understandable to the model. For example, null values will need to be eliminated. Character fields will need to be converted to numerical values. Any extreme outliers will need to be analyzed and also eliminated. The end result of this step is to provide the model with the cohesive format of data that is translatable and capable of obtaining a result.

The German Credit dataset was provided through the University of California – Irvine Machine Learning Repository as submitted by Dr. Hans Hoffman of the Institute for Statistics & Econometrics, University of Hamburg. This data was submitted in both a text and numerical format for use in modeling. As such, there was minimal data preparation necessary for use in our model.

Model Building Based on Neural Networks: After the data has been selected and pre-processed, it can be used in various modeling methods to test a hypothesis. The process of training and testing the model is known as the model building phase. Data is partitioned into two sets. Data used to train the model is known as the training set. Data used to test the model is known as the testing set. Algorithms in this step are generated and modified to prove the hypothesis for the business case. This iterative process continues until the goal for the model is reached.

Multiple models and methodologies exist, however one of the most prevalent approaches is the use of an Artificial Neural Network. This approach uses a set of mathematical weights to map input values to output values thereby establishing or extracting a clear pattern of machine learning. A network is able to learn or establish these patterns through repeated training intervals in which data is supplied to it. The weights supplied to the network can be modified to obtain the preferred result.

The German Credit Dataset contains 20 attributes used to generate a score prediction for fiscal viability. The researchers at the University of Ireland-Galway modified the set of attributes used in their neural network model to demonstrate that a score predictor could be generated with comparable accuracy using fewer attributes. The scientists refer to this as 'feature selection'. Of the 20 attributes available, the team eliminated 13 they deemed irrelevant. The remaining 7 attributes included:

Attribute ID	Description
1	Status of existing checking account
2	Duration in month
3	Credit history
5	Credit amount
6	Savings account/bonds
15	Housing
20	Foreign worker

Testing and Evaluation: When the model is fully developed, testing and evaluation of its results is performed. Although there is a significant amount of training and testing performed during the development phase, this phase differs in the data set for which it is performed. Data sets are usually divided into training and testing data sets. The final model is evaluated using this testing data, or alternatively, with new data sets obtained for testing purposes. The results are evaluated to determine whether or not the model is returning results that are conducive to the original business and project goal.

The University of Ireland Researchers tested their algorithm by creating 20 Neural Networks using the original 20 attributes as well as an additional 20 Neural Networks using their feature selection model. They discovered that although the accuracy score of the training data decreased, the accuracy of the predictor score was comparably higher as denoted by the charts below.

Units	Links	Acc. on train set (%)		Acc. on test set (%)	
		Ave.	Std. Dev.	Ave.	Std. Dev.
1	27	77.83	0.23	75.85	0.35
2	54	77.58	1.02	74.45	0.46
3	81	78.88	1.36	74.45	1.65
4	108	80.38	1.09	73.15	0.46

Table 3. *Results from the german credit problem with 7 selected attributes used as input*

Units	Links	Acc. on train set (%)		Acc. on test set (%)	
		Ave.	Std. Dev.	Ave.	Std. Dev.
1	74	85.99	0.31	72.66	1.13
2	148	86.49	1.48	72.36	2.21
3	222	88.19	2.34	72.36	0.17
4	296	92.69	0.75	71.46	1.75

Table 4. *Results from the german credit problem with all 20 attributes used as input*

Our team also attempted to implement this method of feature selection by choosing various combinations of attributes in order to achieve a score lower than those of the University of Ireland. However, the lowest score we were able to obtain was a 15 using the following 10 columns:

Attribute	**Description**
1	Status of existing checking account
2	Duration in month
3	Credit history
5	Credit amount
7	Present employment since
8	Installment rate in percentage of disposable income
11	Present residence since
15	Housing
16	Number of existing credits at this bank
20	Foreign worker

Deployment: The final step in the CRISP-DM is to deploy the model against real world application. The model can be used to supply metrics and over time, may require maintenance. Modifications might be made in an effort to keep the model scalable with changing business processes.

Ethics concerns can arise at this phase. Data mining is often a process of de-individualizing data. Our particular network model utilized all twenty attributes for the creation of the neural network. However, modification of fields might lead to ethics concerns. For example, use of additional columns such as 'Present Employment Since' might eliminate those customers who have a good credit history but a new job. Alternatively use of columns like "Age in Years" and "Personal Status and Sex" can raise ethical concerns and might violate current fair credit laws in the United States.

Part V: Results

The German credit data set is a list of 1000 applicants wishing for a bank loan. The purpose of the analysis is to predict, based on the dataset, those that will be classified as a good credit candidate or bad candidate using various inputs. The 1,000 applicants within the data set are composed of 20 attributes, with descriptions ranging from status of checking account to credit amount and current housing situation. Team Alpha separated the 1000 applicant data into 980 training data sets and 20 holdout data sets for validating our prediction model.

The classification of good or bad credit is assigned to every applicant and helps to validate the simulation output data we've composed. We tested running the simulation using different sets of holdout data, rather than reducing the number of attributes as followed by University of Ireland paper. In varying the holdout and training input, we noticed a dramatic difference in simulation accuracy with the output generated classification of good or bad credit.

Varying holdout data

In order to simulate the holdout input data and produce the most accurate predictive results, we decided to try and vary our holdout input to different sets to determine if there was any difference. We discovered when using applicants 981-1,000 as holdout input, the predictive capability was poor. On average, we were seeing a score of 20-25 using the formula provided by Professor. However, if we simply adjusted the holdout input to applicants 1-20, our score lowered to an average below 10, with our best score of 8. Further, for both training data sets, we trained the data at least 25 times before calculating our score.

Description of results

For our results, we followed the same strategy for both sets of holdout inputs. The only difference in the analysis was in the holdout input data set. The training input and output data was trained 25 times over again with default settings for layers and neurons (2 layers and 10 neurons). Screenshot 1 and Screenshot 2 show the results for each our holdout input variations.

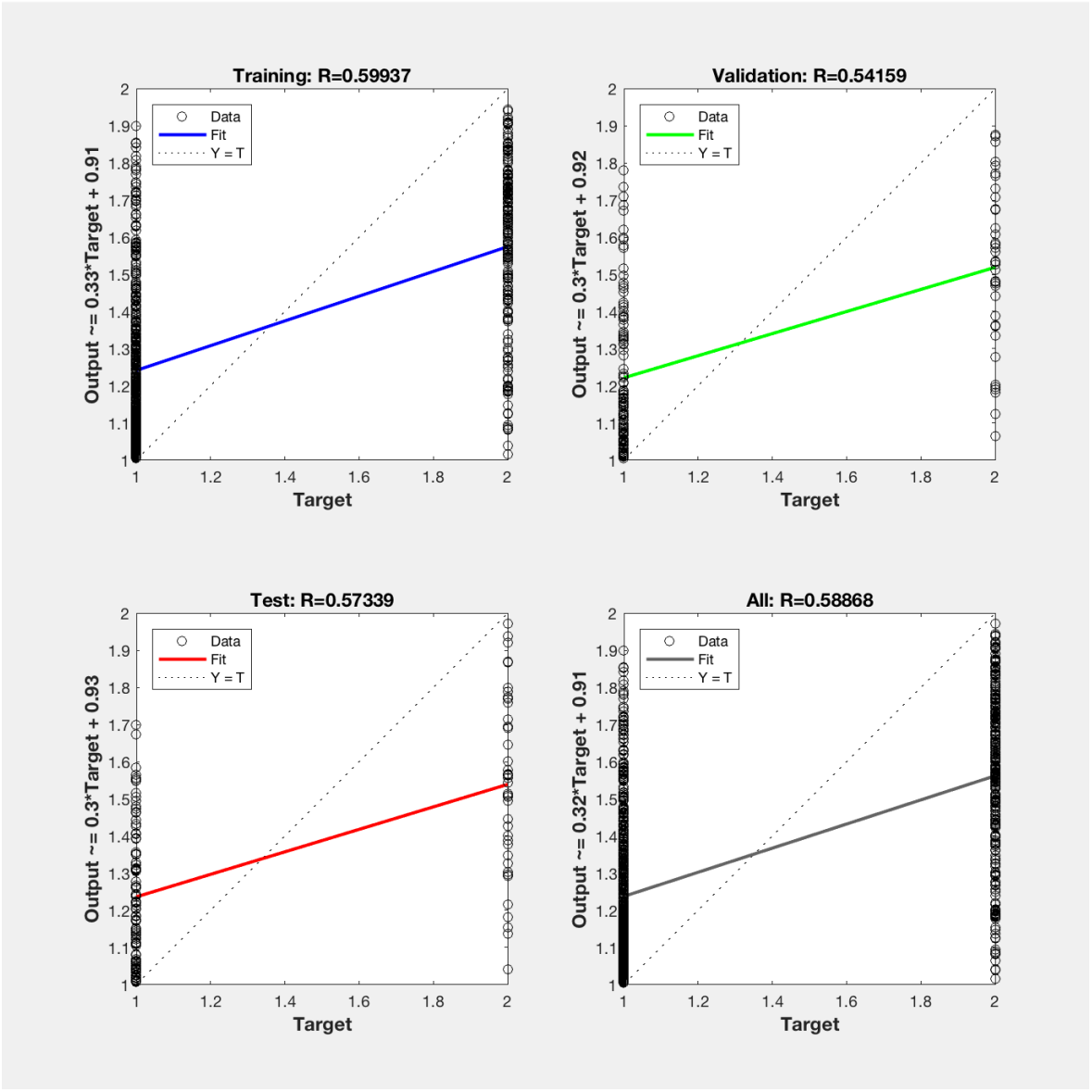
1	2	1	5
1	2	1	5
1	1	1	0
1	2	1	5
1	1	1	0
1	1	2	1
1	1	2	1
1	1	1	0
1	1	2	1
1	1	1	0
1	1	1	0
1	1	2	1
1	1	2	1
1	1	2	1
1	1	1	0
1	1	1	0
1	1	2	1
1	1	1	0
1	2	2	0
1	1	2	1
			23

Screenshot 2 – Using Rows 981-1000 as Holdout Data

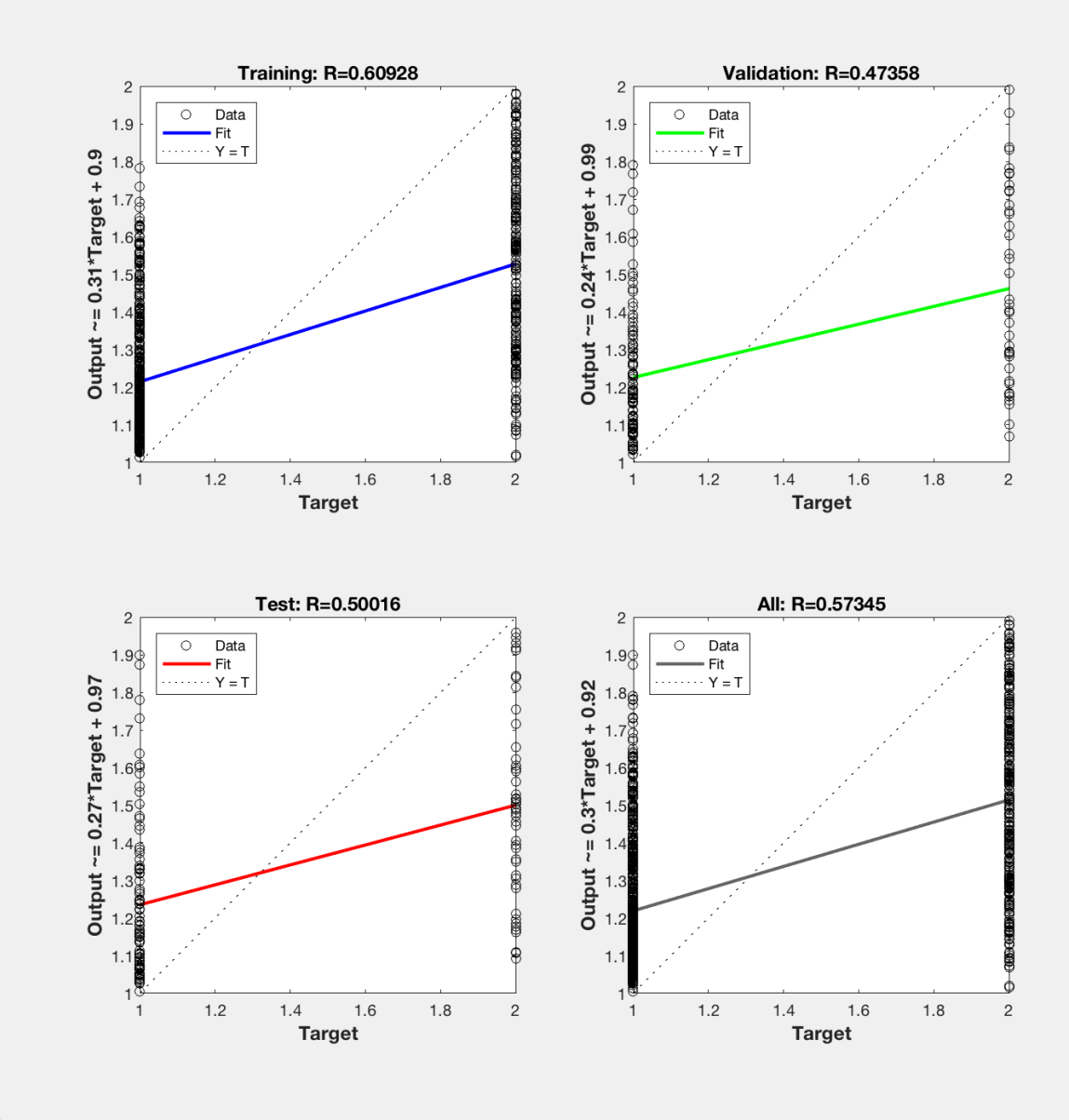
T	U	V	W	X
201	1	1		
201	2	2		
201	1	1		
201	1	1		
201	2	2		
201	1	1		
201	1	1		
201	1	2	1	
201	1	1		
201	2	2		
201	2	2		
201	2	2		
201	1	1		
201	2	1	5	
201	1	2	1	
201	2	2		
201	1	1		
201	1	2	1	
201	2	2		
201	1	1		
201	1			
201	1		**8**	
202	1			

Screenshot 1 – Using Rows 1-20 as Holdout Data

From these results, we can see that simply adjusting the training data has a dramatic difference on the accuracy of the simulation output. Using the first 20 applicants in the German credit set, improved the accuracy of the simulation over 100% versus using the last 20 applicants for holdout input. In both results, however, the R regression remained very similar.



Screenshot 3 – Using Rows 1-20 as Holdout Data



Screenshot 4 – Using Rows 981-1000 as Holdout Data

Part VI: Conclusion

We showed that the credit analysis simulation accuracy is determined largely by variations on the training dataset. Based on these findings, we believe further experimentation should be conducted in order to identify the optimal holdout input data set to produce the most accurate results. Team Alpha's system has been successful in producing superior results to the University of Ireland paper based on our simulation tests using their procedure.

Bibliography

1. Delen, Dursun. *Real World Data Mining*. Upper Saddle River, NJ: Pearson Education Inc., 2015. Print.

2. O'Dea, et al. "*Combining Feature Selection and Neural Networks for Solving Classification Problems*", 18 February 2017.< http://www3.it.nuigalway.ie/cirg/localpubs/aics01.pdf>

3. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, "*An Introduction to Statistical Learning with Applications in R":* http://www-bcf.usc.edu/~gareth/ISL/

4. Jure Leskovec, Anand Rajaraman, Jeff Ullman, "*Mining of Massive Datasets*". http://www.mmds.org/

5. Turban, Efraim, et al. "Neural Networks in Data Mining". *Business Intelligence: A Managerial Approach*. 2008. Web. 18 February 2017. http://www70.homepage.villanova.edu/matthew.liberatore/Mgt2206/turban_online_ch06.pdf

6. Zazai, Said. "Ethics in Data Mining." *Said Zazai, A Tech Enthusiast*. 10 Dec. 2016. Web. 17 Feb. 2017. <http://www.zazai.ca/?p=125>