Effects of Normalization Techniques on Logistic Regression in Data Science

Adekunle Adeyemo

Hayden Wimmer
hayden.himmer@gmail.com

Georgia Southern University
Statesboro, GA, 30458

Loreen Powell
lpowell@bloomu.edu
Bloomsburg University
Bloomsburg, PA 17815

Abstract

The improvements in the data science profession have allowed the introduction of several mathematical ideas to social patterns of data. This research seeks to investigate how different normalization techniques can affect the performance of logistic regression. The original dataset was modeled using the SQL Server Analysis Services (SSAS) Logistic Regression model. This became the baseline model for the research. The normalization methods used to transform the original dataset were described. Next, different logistic models were built based on the three normalization techniques discussed. This work found that, in terms of accuracy, decimal scaling marginally outperformed min-max and z-score scaling. But when Lift was used to evaluate the performances of the models built, decimal scaling and z-score slightly performed better than min-max method. Future work is recommended to test the regression model on other datasets specifically those whose dependent variable are a 2-category problem or those with varying magnitude independent attributes.

Keywords: Normalization, Logistic Regression, Z-Score, Min-Max, Decimal Scaling

1. INTRODUCTION

Advancements in the field of data science have allowed the application of several mathematical concepts to behavioral patterns of data. Precisely, different normalization techniques have been applied to numerous datasets to solve problems from all walks of life. Data normalization is a preprocessing method used in different data mining systems, particularly, for classifying algorithms such as neural networks, clustering and neighbor classification (Evans, 2016). A lot of works have been published in data normalization and its application to different fields of human endeavors; Statistical Normalization and back Propagation for Classification, Min-Max Normalization based on Data Perturbation method for Privacy Protection, Importance of Data Normalization for the application of Neural Networks to Complex Industrial Problems and the Impact of Normalization Methods on RNA-Seq Data Analysis. In this research, we investigated how different normalization techniques affect the Performance of a Logistic Regression Classifier. Logistic regression is an ideal tool for answering

classification questions. It is a model that can be used to forecast the binomial outcome of a dependent (target) variable using one or more independent (predictor) variables. The independent variables can be binomial, numerical or even categorical. Logistic Regression algorithm is used to classify Red Wine dataset based on its quality, the dataset was then normalized using three different normalization methods and different models were built as a result. These new models were compared with the baseline model and performance effect was then discussed.

The original dataset is modeled using SQL Server Analysis Services (SSAS) Logistic regression tool. This serves as the baseline model, this is followed by describing the normalization methods used in transforming the data and different logistic regression model are then built as a result.
Since the aim of this study is to compare how various normalization techniques affect the performance of logistic regression model, the most commonly used normalization methods; Min-Max, Z-score, and Decimal Scaling are used to transform the original data and the performance of the resulting models are evaluated using the accuracies and model lifts as the major metrics. The remaining format of this paper is the following: literature review, methodology, results, implications, and conclusion.

2. LITERATURE REVIEW

There have been different publications on data normalizations and how different normalization methods are applied in different fields to solve various problems. The publications in this category are described in the subsequent paragraphs.

Min-Max normalization techniques was used to preserve privacy of data as a distorting method by (Jain & Bhandare, 2011). Min-Max normalization technique was applied to the original dataset (M) to get a newly transformed dataset ($\bar{M}$) with same number of rows (records) and columns (attributes). The $\bar{M}$ can appear as a distorted form of M. $\bar{M}$ was then altered further to improve its security by multiplying it with a negative number. This action changed the other and values of $\bar{M}$ as positive numbers become negative. This technique was applied to four different real-life databases obtained from UCI Machine Learning Repository. This data perturbation method with shifting factor SF = -15 was applied on these real-life databases. The experiments conducted showed that the value difference (VD) and accuracy of two of the datasets changed with respect to SF. Another SF was carefully selected to better the result. The publication is important to this research as it described how data normalization technique was being used to change the meaning of dataset to preserve its' privacy.

Normalization methods as they relate to sequencing of RNA data and Impact analysis of the results of gene expression were compared by (Zyprych-Walczak et al., 2015). Five Normalization methods were compared using three real-life RNA-seq datasets. Housekeeping Genes (HG) was selected as the analytical criterion for comparing the normalization methods used in processing of RNA-seq data. Since the goal of the study was to find out how normalization techniques impact differential expression results, differential analysis was conducted using edgeR method in the edgeR Biconductor Package after the application of each normalization approach. The results of the experiments conducted were compared using different factors. These results showed that the impact of the normalization technique depends on the data structure and the criteria for comparison. This study opens explores the fact that the influence of data normalization method is dependent on the dataset and the criteria for comparing the performance.

How input data normalization improve the performance of parameter predictors trained to assess the value of several attributes of a nuclear plant was showed by (Sola & Sevilla, 1997). Two different systems were studied, pressurizer pressure and power transferred between the reactor coolant system and the main vapor system. These two networks were studied using neural network simulator SINAPSIS. Three-layered perceptron was used in both systems and training was done through back propagation algorithm. 6 and 8 input variables were used accordingly. The influence of network architecture on the results was studied by evaluating the behavior of a wide range of options. The input variables were normalized using five different normalization techniques. The results showed that a suitable normalization of input variables before network training reduced estimation errors by 10% and the required calculation time during the training process is also reduced. This study proves that normalization of input data can improve the performance of neural network classifier.

Different normalization methods applicable in back propagation neural networks as they enhance the reliability of the trained network was presented by (Jayalakshmi & Santhakumaran, 2011). The reliability of each of the method described was stated and how they affect the attributes of the datasets. The simulations were conducted using MATLAB, different networks were reproduced and experiment with. The network was trained 10 times and the performance was examined at different periods. The results of the experiments clearly revealed that the performance of the dataset used in the classification model relied on the normalization methods and that the Statistical Column normalization produced the most accurate result. The study showed how the performance of back propagation neural nets can be improved upon using some applicable normalization approaches.

Other relevant literatures to the work at hand described data normalization protocols and features specifically constructed to improve the performance of some classifiers, which are important to our research topic. Specifically, a protocol for data exploration that can be used to avoid common statistical problems was proposed by (Zuur, Ieno, & Elphick, 2010). The protocol in question was divided into eight linear flexible stages. The stages include identifying and removing outliers, variance homogeneity, normal distribution of data, lots of zero in the data, the existence of correlation between covariates, considering the relationships between response and predictor variables, considering interactions between output attribute and different type of covariates and independent observation of response variable. The paper discussed a series of drawbacks that can impact the output of an analysis, but these can be avoided using the systematic data exploration procedure described before undertaking any analysis. This paper is such an important one as it teaches how to prepare dataset before applying it in analysis or modeling.

Kaizen Programming (KP) approach was employed to improve Logistic Regression model to find high-quality nonlinear combinations of the original features in a dataset by (de Melo & Banzhaf, 2016). KP together with LR model was used to filter important features of credit scoring dataset and Akaike Information Criterion (AIC) was used as selection model aimed at improving the prediction performance of LR. The performance of KP was implemented using Australian Credit Approval dataset, the continuous variables in the dataset were discretized since the implementation did not work with mixed type attributes. Before models were built, identical or highly connected features were discarded. KP was implemented in Python and experimental analysis was executed on Weka using the LR as the classifier. The new dataset with the best accuracy for each desired feature was selected. The experiment showed that KP results were competitive though some imbalanced because it generated different features compared to other methods in the literature. The study proved that Logistic Regression model can be used together with another problem-solving method such as KP to improve its predictive performance.

A recurrent neural network approach was applied to stock price pattern recognition by (Kamijo & Tanigawa, 1990) . The proposed network was a four-layered architecture with one layer for input, two as hidden and one layer as output. The output layer is used to discriminate nonlinear patterns. Sixteen experiments were conducted with sixteen stock price patterns for recognition. After the experiments, the actual pattern was correctly recognized 15 times out of 16 experiments that were conducted. The results of the experiments showed that normalization by exponential smoothing introduced s bias which is the difference in name and time span. The research work showed that exponential smoothing way of normalizing data introduced some errors to the neural network model for pattern recognition.

3. METHODOLOGY

This research work investigates the effect of different Normalization Techniques on the prediction accuracy of Logistic Regression model. The SQL Server Analysis Services (SSAS) is the major tool used for the work. SSAS is a tool from the Microsoft Business intelligence team, for developing Online Analytical Processing (OLAP) solutions. A typical workflow consists of authoring a multidimensional or data model in tabular format, deploying the model as a database to an SSAS or Azure Analysis Services. SSAS environment is a collection of machine learning algorithms such as, Neural Network, Decision Tree, Naïve Bayes, Logistic Regression and so on.

Dataset

The problem that was selected for this research is to predict the quality of red wine data using Logistic Regression model of SQL Server Analysis Services (SSAS). To investigate the

performance of this classifier, the model was applied to Red Wine Quality dataset of the Portuguese "Vinho Verde" wine collected from UCI Machine Learning Repository. The output attribute is a 11-class problem between 0(very bad) and 10(very excellent) for red wine quality. The dataset consists of 1599 instances. Each record consists of 11 input attributes. The relevant independent attributes determined by the dependency capacity of Logistic Regression Model are:

1. Alcohol
2. Sulphates
3. Fixed Acidity
4. Citric Acid
5. pH Values (objective test)
6. Residual Sugar
7. Free Sulfur Dioxide

There are 10 instances of the dataset with quality of 3, 53 with quality of 4, 681 with quality of 5, 638 with quality of 6 and 18 with quality of.

Structural Explanation of a Logistic Regression Model

The SSAS Logistic Regression model is formed using Neural Network algorithm with the elimination of hidden node. Hence, the general model for a logistic regression is almost the same as that of neural network; each model has a single root node representing the model and the details about it, and a distinct marginal statistics that gives the details about the independent attributes used in the model.

Furthermore, the model consists of a subnetworks for each dependent attribute. Each subnetwork contains two branches; one for the input layer and the other contains the hidden layer and the output layer. However, in this model, the hidden layer is empty as it has no children. So, the model consists of nodes that stand for individual outputs and inputs with empty hidden nodes. As it is shown in Figure 1, the logistic regression model is presented using the Neural Network Viewer. The neural network viewer allows the filtering of input attributes and their values and graphical view as these affect the outputs. There are various tabs in the viewer that show the probability and lift association as regards to the input and output values.

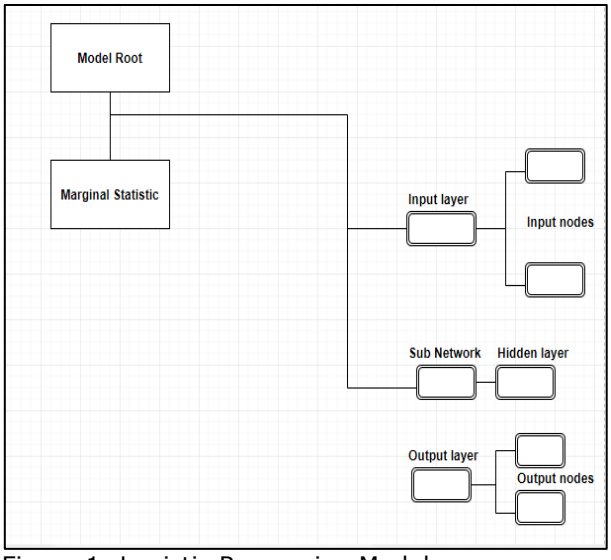


Figure 1: Logistic Regression Model

Normalization Techniques

The three types of Normalization techniques applied to the dataset are described in the subsequent paragraphs. There are numerous types of Normalization methods but these three are chosen based on their popularity in the reviewed literatures. These techniques include: Min-Max, Z-score and Decimal Scaling.

Min-Max Normalization

This technique is a strategy that linearly transform the attributes or outputs from one range of values to a new range of values. Mostly, the variables are transformed to lie between 0 and 1 or -1 and 1. The rescaling is usually achieved using the linear transformation given as:

$$y = (x - \min(x))/(\max(x) - \min(x))$$

Where min and max are the minimum and maximum values in X, where X is the set of observed values of x. In other words, max(x) – min(x), is the range of X. The advantage of this normalization method is derived from the fact that all relationships in the data are exactly preserved.

Z-Score Normalization

This method is the most popular normalization method which converts all input values to a common measure with an average of zero and standard deviation of one. The mean and standard deviation are calculated for each attribute. Each value of an attribute X is normalized using the computed mean and standard deviation. The transformation equation is given as:

$$y = (x - \text{mean}(X))/\text{std}(X)$$

Where mean(X) = mean of attribute X and std(X) = standard deviation of attribute X. The advantage of this method is deduced from the fact that it reduces outliers' effect on the data.

Decimal Scaling

This normalization technique works by moving the decimal point of values of attribute X. The number of points moved is determined by the maximum absolute value of X. The value x of attribute X is normalized to y by using the formula:

$y = x/10i$

Where i is the smallest integer that satisfy the condition $Max(|y|) < 1$.

Structural Representation of Activities

Figure 2 shows the structural representation of all the activities involved in this research. The dataset is presented into Logistic Regression (LR) model in four different views and structures. The original dataset and the resulting datasets after normalization using three different techniques as depicted in the model above served as inputs to the LR model. Four different prediction models were generated as outputs. The three outputs from the normalization techniques were compared with the output of the original dataset based on the prediction accuracies of the model.

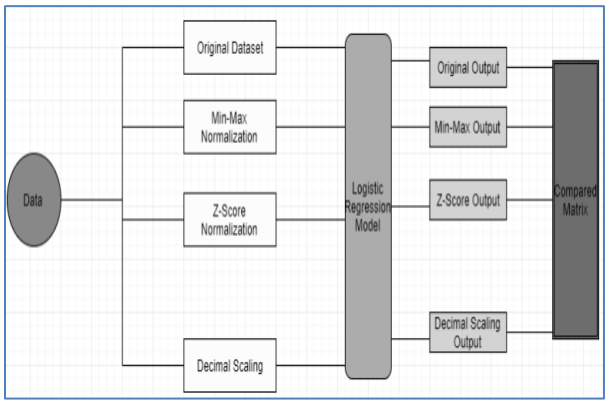


Figure 2: Structural Activities Model

4. RESULTS

The tests to evaluate how different normalization techniques affect the performance of Logistic Regression Model (LRM) have been conducted. The Red Wine Quality (RWQ) dataset which consists of 1599 records was divided into two sets using SSAS LRM. The training set consists of 70% of the original dataset (1119 records) and the testing set consists of the remaining 30% of the dataset (480 records). The RWQ dataset was normalized using the Min-Max, Z-Score and Decimal Scaling normalization methods as described under the method section. These same percentages of training and testing sets were used for each of the normalization techniques.

Tables 1, 2, 3 and 4 show the classification matrices for the four models, that is, one for the original dataset and the remaining three for each normalization techniques. These matrices are also called Confusion Matrices and are used for summarizing the performance of a classification algorithm or classifier. The columns of the classification matrices correspond to actual values, rows correspond to predicted values.
Accuracy and lift were the two major metrics used for evaluating the performances of the LR models.

		Actual					
		6	3	8	7	4	5
Predicted	6	92	2	4	41	7	41
	3	0	0	0	0	0	0
	8	0	0	0	0	0	0
	7	9	0	2	14	1	3
	4	0	0	0	0	0	0
	5	68	2	0	11	10	172

Table 1: Classification Matrix for Baseline-Model on Quality

		Actual					
		6	3	8	7	4	5
Predicted	6	106	1	2	40	3	54
	3	0	0	0	0	0	0
	8	0	0	0	0	0	0
	7	11	0	4	14	1	4
	4	0	0	0	0	0	0
	5	77	2	0	2	7	151

Table 2: Classification Matrix for Min-Max on Quality

		Actual					
		6	3	8	7	4	5
Predicted	6	112	1	3	33	8	53
	3	0	0	0	0	0	0
	8	0	0	0	0	0	0
	7	9	0	1	11	1	2
	4	0	0	0	0	0	0
	5	79	2	0	7	9	148

Table 3: Classification Matrix for Z-Score-Model on Quality

		Actual					
		6	3	8	7	4	5
Predicted	6	113	1	3	37	7	43
	3	0	0	0	0	0	0
	8	0	0	0	0	0	0
	7	13	0	3	12	1	3
	4	0	0	0	0	0	0
	5	76	2	0	8	9	148

Table 4: Classification Matrix for Decimal-Scaling on Quality

Accuracy

The accuracy of a model is defined as the percentage of the test dataset correctly specified. This is given as:

$$\text{Accuracy} = \frac{No\ of\ correctly\ classified\ test\ samples}{\text{Total no of Test Samples}}$$

Number of correctly classified test samples is the summation of all the diagonal values in a matrix. Table 5 contains the accuracies as reported by the four models. Figure 3 shows the graphical representation of the models accuracies.

Model	Accuracy
Data	57.92
Min-Max	56.46
Z-Score	56.46
Decimal Scaling	56.88

Table 5: Model Accuracy in Percentages

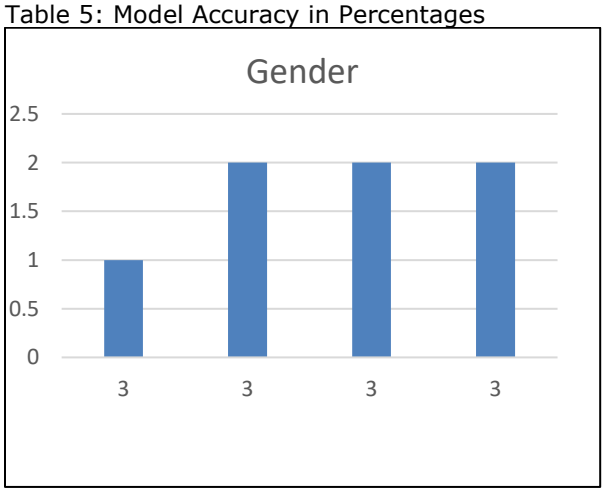


Figure 3: Graphical representation of the models accuracies

Model Lift

A lift measures the proportion of the true positives from the model compared to proportion of positive hits in the dataset overall. The lift of a model can be obtained directly from the SSAS by clicking on the Lift Chart tab under the Mining Accuracy Chart.

From table 5, the LRM performs better on the original dataset as the prediction accuracy was about 58% compared to when the dataset was normalized with accuracies of about 56%, 56% and 57% for Min-Max, Z-Score and Decimal Scaling normalization methods respectively. The LRM behaves similarly even when the training set was increased to 80% for all the normalization techniques.

The figures 4, 5, 6, 7, 8, 9, 10 and 11 showed the Lift Charts and Legends for our Baseline, Min-Max, Z-Score and Decimal-Scaling models respectively. In the figures, the legends show that two lines should be displayed, one for the specific model, such as, baseline model, and one for the ideal model.

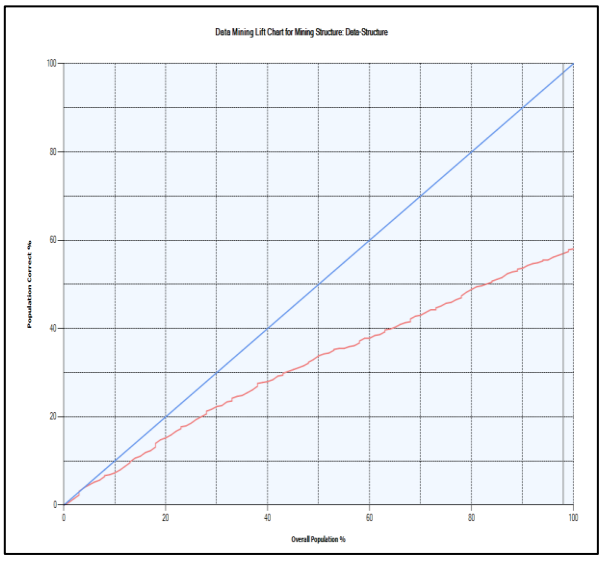


Figure 4: Baseline Model Lift Chart

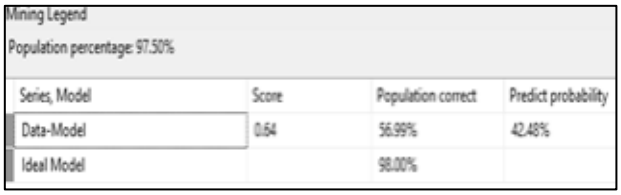


Figure 5: Baseline Model Lift Legend

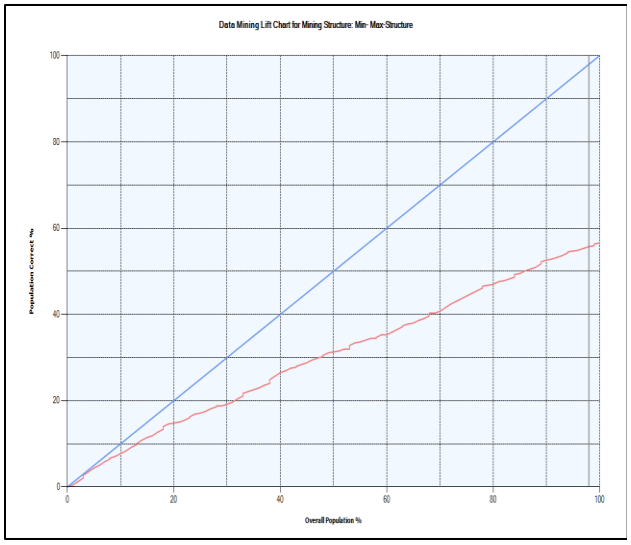
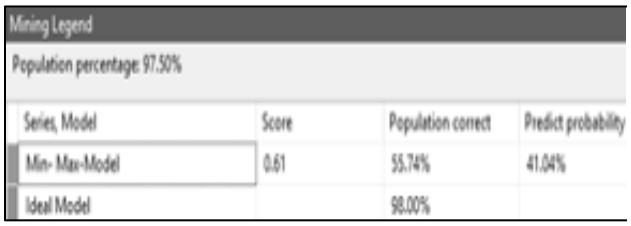


Figure 6: Min-Max Model Lift Chart



Mining Legend

Population percentage: 97.50%

Series, Model	Score	Population correct	Predict probability
Min- Max-Model	0.61	55.74%	41.04%
Ideal Model		98.00%	

Figure 7: Min-Max Model Lift Legend

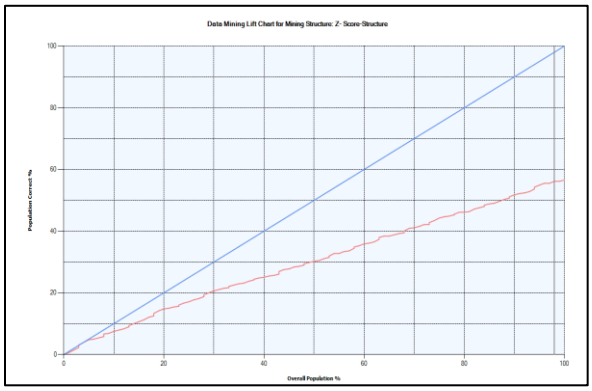
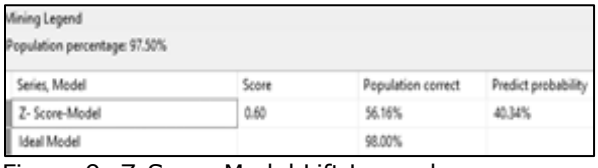


Figure 8: Z-Score Model Lift Chart



Mining Legend

Population percentage: 97.50%

Series, Model	Score	Population correct	Predict probability
Z- Score-Model	0.60	56.16%	40.34%
Ideal Model		98.00%	

Figure 9: Z-Score Model Lift Legend

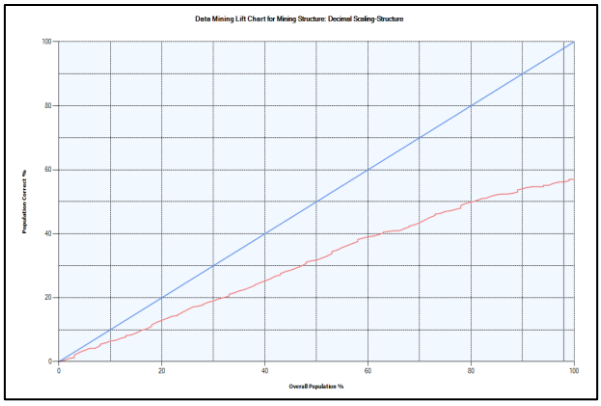
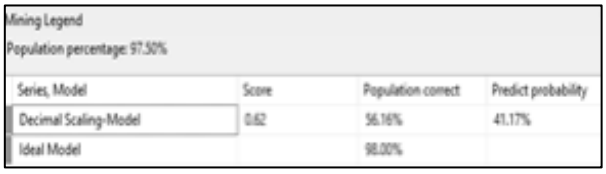


Figure 10: Decimal-Scaling Model Lift Chart



Mining Legend

Population percentage: 97.50%

Series, Model	Score	Population correct	Predict probability
Decimal Scaling-Model	0.62	56.16%	41.17%
Ideal Model		98.00%	

Figure 11: Decimal-Scaling Model Lift Ledged

For a perfect classification model, as we have in figures 4, 6, 8, and 10 the Ideal Model is on top of the specific model chart line. The solid gray vertical bar can be clicked and dragged horizontally to examine different values along the plotted line in the Legend windows. In our case, the Legend windows show that for 98 percent of the overall population 56.99%, 55.74%, 56.16% and 56.16% for Baseline, Min-Max, Z-Score and Decimal Scaling models respectively were correctly predicted.

Although the accuracy of a normalized dataset is expected to improve with classifiers such as Neural Networks, LRM performs poorly to normalized dataset. This LRM's behavior might be because of the magnitude of the independent variables in the dataset that are already close to one another before normalization or the output attribute that is an 11-class problem.

5. IMPLICATIONS FOR PRACTICE

Data Normalization means transformation of all attributes in the dataset to a specific scale. We do data normalization when seeking for relationship between the variables in the dataset. Several works have been published on data normalization and how important these techniques have become as a data preprocessing strategy but little effort has been geared towards how these methods affect the performance of machine learning algorithms especially the Logistic regression. This research work is very important in that it will serve as the

foundation for researchers to build on and for data scientists to see that Logistic regression performs poorly under the influence of normalized data.

6. CONCLUSIONS AND FUTURE DIRECTIONS

The performance of Logistic Regression Model was evaluated with respects to three different normalization techniques. As usual, normalization of dataset is expected to improve the predictive accuracy of a machine learning algorithm but LR behaves poorly to the three normalization techniques tested. Although two different sizes of training datasets were used, the accuracies of both models based on the normalization methods were similar. One of the future works will be to test the performance of LR algorithm on other datasets whose independent variables are vary in magnitude or those whose target variables are a 2-class problem. We will also try to see how Linear Regression algorithm performs under different normalization methods.

7. REFERENCES

de Melo, V. V., & Banzhaf, W. (2016). *Improving Logistic Regression Classification of Credit Approval with Features Constructed by Kaizen Programming.* Paper presented at the Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion.

Evans, J. R. (2016). *Business Analytics, 2e*. Boston, MA: Pearson.

Jain, Y. K., & Bhandare, S. K. (2011). Min max normalization based data perturbation method for privacy protection. *International Journal of Computer & Communication Technology, 2*(8), 45-50.

Jayalakshmi, T., & Santhakumaran, A. (2011). Statistical Normalization and Back Propagationfor Classification. *International Journal of Computer Theory and Engineering, 3*(1), 89.

Kamijo, K.-i., & Tanigawa, T. (1990). *Stock price pattern recognition-a recurrent neural network approach.* Paper presented at the Neural Networks, 1990., 1990 IJCNN International Joint Conference.

Sola, J., & Sevilla, J. (1997). Importance of input data normalization for the application of neural networks to complex industrial problems. *IEEE Transactions on Nuclear Science, 44*(3), 1464-1468.

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution, 1*(1), 3-14.

Zyprych-Walczak, J., Szabelska, A., Handschuh, L., Górczak, K., Klamecka, K., Figlerowicz, M., & Siatkowski, I. (2015). The impact of normalization methods on RNA-Seq data analysis. *BioMed research international, 2015*.

