2016

A Comparison of Machine Learning Techniques and Logistic Regression Method for the Prediction of Past-Due Amount

Jie Hao
Kennesaw State University

Jennifer L. Priestley
Kennesaw State University, jpriestl@kennesaw.edu

A Comparison of Machine Learning Techniques and Logistic Regression Method for the Prediction of Past-Due Amount

Jie Hao
Department of Statistics and Analytical Sciences
College of Science and Mathematics
Kennesaw State University

Jennifer Lewis Priestley
Department of Statistics and Analytical Sciences
College of Science and Mathematics
Kennesaw State University

***Abstract*—The aim of this paper to predict a past-due amount using traditional and machine learning techniques: Logistic Analysis, k-Nearest Neighbor and Random Forest. The dataset to be analyzed is provided by Equifax, which contains 305 categories of financial information from more than 11,787,287 unique businesses from 2006 to 2014. The big challenge is how to handle with the big and noisy real world datasets. Among the three techniques, the results show that Logistic Regression Method is the best in terms of predictive accuracy and type *I* errors.**

Keywords: Logistic Regression, k-Nearest Neighbor, Random Forest

I. Introduction

The first step of any model building exercise is to define the outcome. Common prediction in the financial services industry is to use binary outcomes, such as "Good" and "Bad". For instance, for a lender, a "good" consumer may have an account that has been no more than 30 days past due while a "bad" consumer is one whose account has been 90 days past due or more. Good and bad outcomes are mutually exclusive events. For our research problem, the most common approach is to reduce past-due amounts into two cases, good and bad. Next, we build a two-stage model using logistic regression method; that is the first predicting likelihood of bad, and the second predicting past-due amount given bad. Logistic analysis as a traditional statistical technique is commonly used for prediction and classification in the financial services industry [7]. However, for analyzing big, noisy or complex datasets, machine learning techniques are typically preferred to detect hard-to-discern patterns[4].

In this paper, using both machine learning techniques and Logistic analysis, we developed models to predict a past-due amount by analyzing datasets provided by Equifax. The next section is a brief review of previous work by other researchers. In III, we describe how to handle real-word big datasets. In IV, we present the methodologies in this paper for k-Nearest Neighbor (kNN), Random Forest (RF), and Logistic Analysis (LA). In the last two parts of the paper are the results and discussion. To compare with all techniques, we use ROC index, sensitivity and specificity as criteria.

II. Literature Review

In recent years, serveral publications in the medical domain discuss machine learning techniques for prediction instead of logistic analysis. Cruz and Wishart (2006) concludes that machine learning methods generally improve the predictive accuracy of most cancer prognoses. Moreover, the use of machine learning classification will become much more commonplace in many clinical and hospital settings [4]. Rana M. et al. (2015) implemented Support Vector Machine, k-Nearest Neighbor, Logistic Regression and Naive Bayes to classify whether the breast cancer is benign or malignant and predict the recurrence and non-recurrence of malignant cases after a certain period. k-Nearest Neighbor technique gave the best result for overall methodology[9].

In the financial domain, Sharma (2012) illustrated an approach to improving credit risk scorecards using Random Forests. It was shown that on data sets where variables have multicollinearity and complex interrelationships, Random Forests provided a more scientific approach to accessing variable importance and achieving optimal predictive accuracy. In addition it was indicated that Random Forests were preferred for econometric and credit risk models as they provide a powerful methodology to assess meaning of variables and thus allow for more robust findings [10]. Babu and Satish (2013) illustrated the advantages of using K-Nearest Neighbor to tackle the credit scoring tasks, such as reducing the cost of credit analysis, enabling faster credit decision than traditional methods and insuring credit collections. They used a standard K-Nearest Neighbor method in pattern recognition and non parametric classification to credit scoring tasks based on learning by similarity [1].

III. Data Discovery

The data for this paper came from Equifax. There are thirty-six datasets in total. Each dataset represents a quarterly report between 2006 and 2014 collected by Equifax, which was named by the archive month. Each dataset contains same 11,787,287 observations representing unique businesses and same 305 variables representing businesses' general information that contain region, zip code etc, account activities

followed by non-financial, telco, industry and service and financial credit information such as reject code, business credit risk score etc.

A. Dependent Variable

In this research, we try to examine the prediction of past-due amount. Among our data, there are 23 variables related to "past-due" as potential dependent variables. However, there exists a large ratio of coded values which do not carry meaningful information and missing values. For example, total service past due amount reported in last 3 months (totSPDAmt3mon) is one of potential dependent variables, and Fig. 1 shows that 50% values of totSPDAmt3mon are coded in the record of last quarter of 2014. Hence, one of the big issues in the dataset is how to handle with coded values.

Fig. 1: Distribution of totSPDAmt3mon

totSPDAmt3mon	Frequency	Percent
Missing	4902621	41.59
Valid Values	953466	8.09
Coded Values	5931200	50.32

Considering the large proportion of coded values, total number of past-due days in non-financial accounts (totNFPD) is taken as the target response. We have two conditions in response variable selection: one is that there are almost one third of total variables related to non-financial accounts in datasets, which guarantees a large scale for us to filter variables; the other is that the percent of coded values is below 50%. Fig. 2 shows that totNFPD meets the above conditions. Filtering missing and coded values in totNFPD, we merged all 36 datasets to be a new dataset which contains 47,131,479 observations. The size of the new dataset is still large enough.

Fig. 2: Distribution of totNFPD

totNFPD	Frequency	Percent
Missing	4902621	41.59
Valid Values	1490346	12.64
Coded Values	5394320	45.76

Fig. 3 illustrates that it is necessary to transform the values of totNFPD into 0 and 1, where 0 denotes no past-due and 1 denotes at least 1 day passing the deadline ever in account. This is because at least 75% values are recorded as 0. Fig. 4 shows that we create the binary dependent variable named as pastdue, which is the response being predicted in the following three models.

Fig. 3: Distribution of totNFPD in Merged Dataset

Analysis Variable : totNFPD							
Minimum	25th Pctl	50th Pctl	75th Pctl	90th Pctl	95th Pctl	99th Pctl	Maximum
0	0	0	0	391.0000000	1461.00	12740.00	999999992

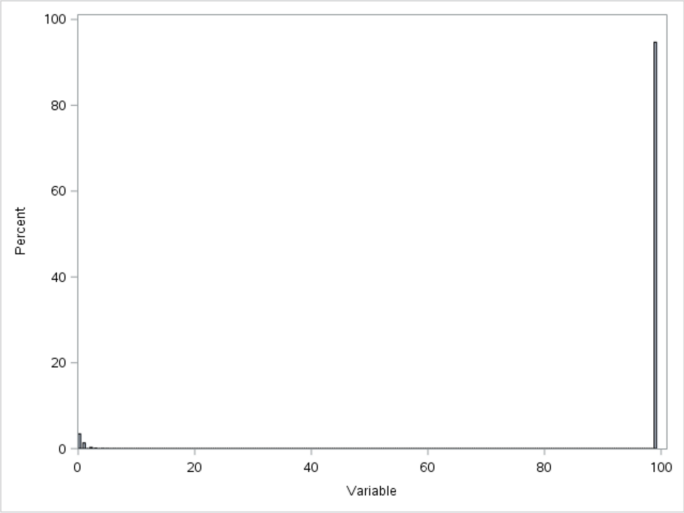
Fig. 4: Distribution of Binary Dependent Variable (pastdue) in Merged Dataset

pastdue	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	37393210	79.34	37393210	79.34
1	9738269	20.66	47131479	100.00

B. Independent Variables

1) Simple Dimensionality Reduction: Variables with a high ratio of coded values are not supposed to carry useful information. The variables will be removed where the precent of coded values is greater than 80%. For instance, Fig. 5 presents the variable with 95% coded values, which is obviously removed. 74 variables are removed based on the given threshold.

Fig. 5: Distribution of the Variable with High Coded Values Ratio

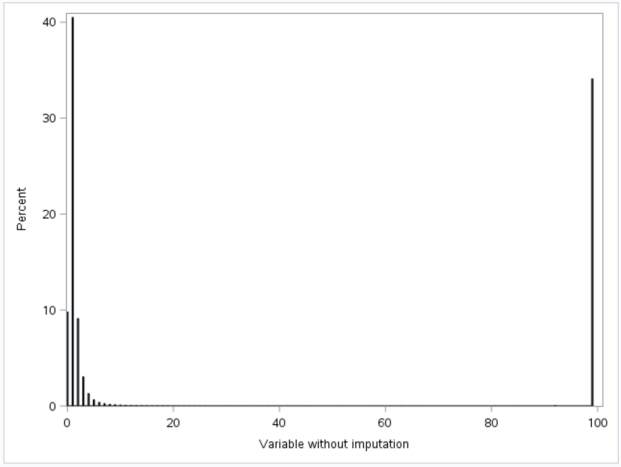


2) Median Imputation: The other big issue in our data is large ratio of missing values. Moreover, all coded values will be treated as missing values. Mean or median imputation is the most common missing values treatment. Since the distributions of variables are right-skewed, median imputation is more robust than mean imputation. Generally, the mean is affected by the presence of extreme values or outliers. In this step, the missing values of a variable are replaced by the median calculated by all known valid values of that variable. Taking one service variable as an example, Fig. 6 shows the effect of median imputation.

3) Dimensionality Reduction Using Variable Clustering: There are four types of accounts based on the design of raw data, which are non-financial, telco, industry and service. And 90 variables are related to non-financial accounts. To reduce the likelihood of multicollinearity, variable clustering is performed on 90 non-financial variables, 41 telco variables, 42 industry variables and 10 service variables, respectively. Chosen 90% as the threshold of total proportion of variation

Fig. 6: Before and After Median Imputation

(a) Before



(b) After

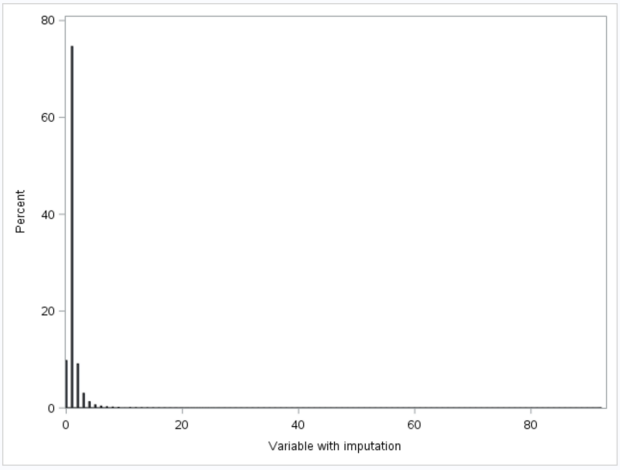
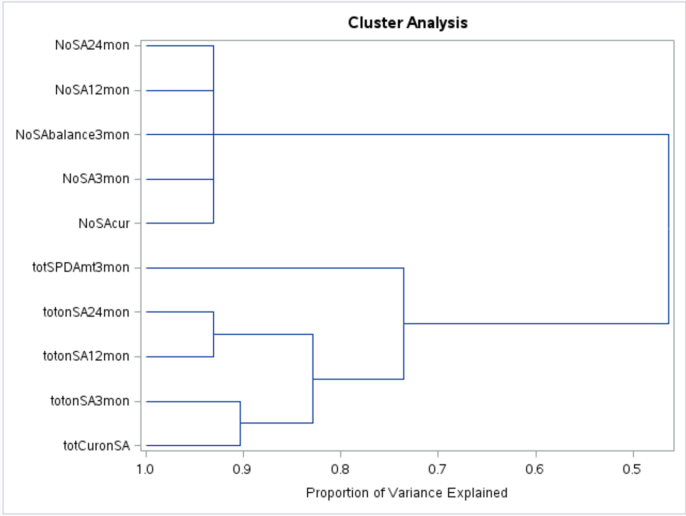


Fig. 7: Performance of Variable Clustering among 10 service variables



explained, 4 is the optimal clustering number for 10 service variables shown in Fig. 7. In each cluster, the variable with smallest ratio of $1 - R^2$ will be picked. Finally, we have 19 non-financial variables, 15 telco variables, 11 industry variables and 4 service variables after clustering. The reduction is aggressive since 73% variables has been removed.

4) Normalization: Data normalization is required for kNN classification. Avoiding the discriminative issue, all independent variables are supposed to be in the same scale. Using simple linear normalization approach, values of each variables are to be in range [0,1]. The formula is shown below:

$$z = \frac{x - min}{max - min}$$

Until now, there are still 47 millions observations and 47 variables after data cleansing. In order to build and run models quickly, sampling is necessary to be considered. As we know, larger sample can increase the accuracy of predictive analytics. In this case, 50,000 observations are drawn as the sample data using simple random sampling. Then we divided the sample data into two parts: training set (60%) and testing set (40%).

IV. Methodology

A. K-Nearest Neighbor (kNN)

K-Nearest Neighbor (kNN) was first demonstrated by Cover and Hart in 1967 [3], which is one of the most fundamental and simple classification methods. The kNN Classification should be one of the first choices for a classification study when we have little knowledge about the data [8]. Firstly, kNN classifier identify the K neighbors in the training data that are closest to the new input to be classified. The proximity of the neighbors or the nearest neighbor to the new input is defined by Euclidean distance. The formula of Euclidian Distance is as below:

$$D(x, y) = \sqrt{\sum_k (x_k - y_k)^2}$$

Then we count the number of nearest neighbors that belong to 0 or 1 in response variable. In the end, we classify the new input to be 0 or 1 where the greater number of nearest neighbor that belong. In order to minimize the error rate, we optimize K, the number of nearest neighbors, by the design from [5]. In SAS, PROC DISCRIM conducts kNN Classification directly using nonparametric method [6].

B. Random Forest (RF)

Random Forest is an advanced method of machine learning, which grows a collection of independent decision trees and each tree casts a unit vote for the most common class at a new input [2]. Each decision tree in a forest is constructed using a bootstrap sample from the data. There are two third of data instances used to construct a tree; the other instances will be into out-of-bag data as a control set. There are m variables out of all the n inputs are randomly selected at each node of the

tree that split based on the selected m variables. The random selection of features at each node decreases the correlation between the trees in the forest. Thus, the RF algorithm can handle many redundant features and avoid model over-fitting [11]. We build RF based model using R.

C. Logistic Analysis (LA)

Binary Logistic regression is a traditional statistical technique that is wellness suitable for examining the relationship between a binary categorical response variable and at least one categorical or continuous independent variables. The model is generally presented in the following format:

$$\ln(\frac{p}{1-p}) = b_0 + b_1x_1 + b_2x_2 + ... + b_kx_k$$

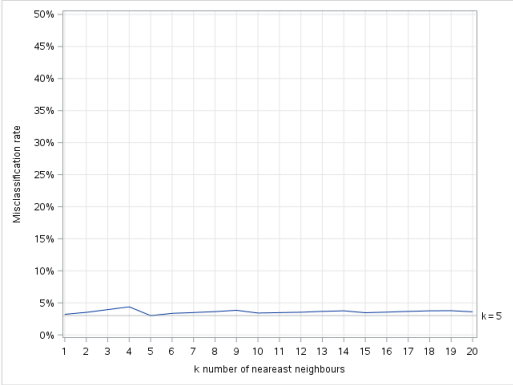
where $\ln(\frac{p}{1-p})$ defines the natural logarithm of the odds ratio, b denotes to the coefficients of parameters and x represents the independent variables.

PROC LOGISTIC conducts logistic analysis in SAS. In our case, all 47 variables will be built into the logistic model for the comparison with other two techniques.

V. Results

In this design, k for kNN classification for the testing set ranges from 1 to 20 in Fig. 8. Most misclassification or error rates stay at pretty low level. And the best k is 5.

Fig. 8: Misclassification rates for the testing dataset (k is from 1 to 20



Conducting 5-NN Algorithm in the training set, we achieved a very low overall error rate is 4.54% shown in Fig. 9 (a). The overall error rate of the testing set is 4.63%, which is as well as the performance in the training set. Then the predictive accuracy using 5-NN classification to predict a past-due amount is 95.37%. In terms of Type I error, false positive rate, is 7.39% shown in Fig. 9 (b).

Fig. 10 displays the performance of RF that the accuracy in the training set is 81.86% and the predictive accuracy is 82.01%. Based on the confusion matrix in Fig. 10 (b), Type I error is 14.4% that is double times than 5-NN Classification's. As we can see Fig. 11, ROC curve is close to the baseline, and Area Under ROC Curve (AUC) is 0.60573 that is considered to be poor.

Fig. 9: Classification Summary Using 5 Nearest Neighbors

(a) Training Set

Number of Observations and Percent Classified into pastdue			
From pastdue	0	1	Total
0	23324 97.88	505 2.12	23829 100.00
1	430 6.97	5741 93.03	6171 100.00
Total	23754 79.18	6246 20.82	30000 100.00
Priors	0.5	0.5	

Error Count Estimates for pastdue			
	0	1	Total
Rate	0.0212	0.0697	0.0454
Priors	0.5000	0.5000	

(b) Testing Set

Observation Profile for Test Data	
Number of Observations Read	20000
Number of Observations Used	20000

Number of Observations and Percent Classified into pastdue			
From pastdue	0	1	Total
0	15561 98.13	297 1.87	15858 100.00
1	306 7.39	3836 92.61	4142 100.00
Total	15867 79.34	4133 20.67	20000 100.00
Priors	0.5	0.5	

Error Count Estimates for pastdue			
	0	1	Total
Rate	0.0187	0.0739	0.0463
Priors	0.5000	0.5000	

Fig. 10: Confusion Matrices and Statistics for RF

(a) Training Set

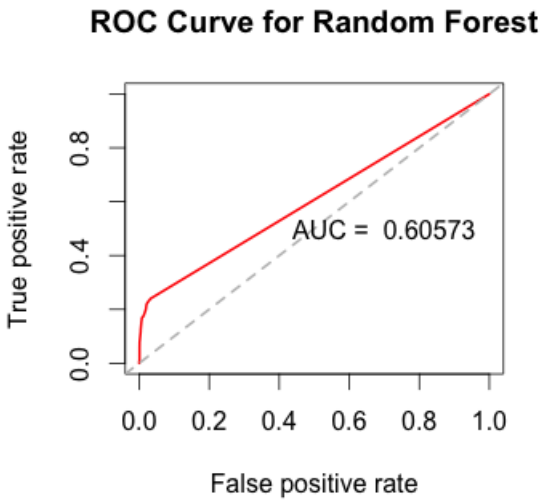
```
          Reference
Prediction     0     1
         0 23643  5256
         1   186   915

               Accuracy : 0.8186
                 95% CI : (0.8142, 0.8229)
    No Information Rate : 0.7943
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2019
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9922
            Specificity : 0.1483
         Pos Pred Value : 0.8181
         Neg Pred Value : 0.8311
             Prevalence : 0.7943
         Detection Rate : 0.7881
   Detection Prevalence : 0.9633
      Balanced Accuracy : 0.5702
```

(b) Testing Set

```
          Predicted
Actual     0     1
     0 15748  3488
     1   110   654

               Accuracy : 0.8201
                 95% CI : (0.8147, 0.8254)
    No Information Rate : 0.7929
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.216
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9931
            Specificity : 0.1579
         Pos Pred Value : 0.8187
         Neg Pred Value : 0.8560
             Prevalence : 0.7929
         Detection Rate : 0.7874
   Detection Prevalence : 0.9618
      Balanced Accuracy : 0.5755
```

On the contrary, Logistic Analysis (LA) performs excellently to predict the response. Fig. 12 shows AUC is 0.9858 being pretty close to 1. In Fig. 13, for instance, With a cutpoint of 0.5, the correct classification rate or the accuracy is 96.3% that is higher than the result of 5-NN. Fig. 14 displays the confusion matrix for Logistic Regression in the testing set

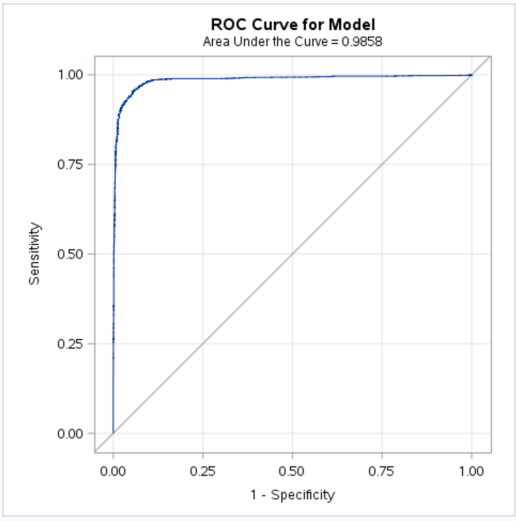
Fig. 11: ROC Index and AUC Index for RF



while Type I error is 1.21% that is lower than 5-NN's.

Overall, Logistic Regression performs the highest accuracy and lowest Type I error. 5-NN Classification is better than Random Forest.

Fig. 12: ROC Index and AUC Index for LA



VI. Discussion

Before the comparison among three techniques, we thought machine learning technique would beat traditional statistical technique that should have been correct since the dataset to be analyzed was big and complex. kNN and Random Forest are non-parametric while both are automatically cross-validated. However, Logistic Regression technique is still the best after data cleansing in our case. As a general rule of thumb, we recommend to apply Logistic Models at the beginning. Then a nice probabilistic interpretation is obtained.

Fig. 13: Classification Table for LA

Classification Table

Prob Level	Correct		Incorrect		Percentages				
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	6171	0	23829	0	20.6	100.0	0.0	79.4	.
0.100	5908	22358	1471	263	94.2	95.7	93.8	19.9	1.2
0.200	5719	22958	871	452	95.6	92.7	96.3	13.2	1.9
0.300	5611	23256	573	560	96.2	90.9	97.6	9.3	2.4
0.400	5533	23382	447	638	96.4	89.7	98.1	7.5	2.7
0.500	5454	23448	381	717	96.3	88.4	98.4	6.5	3.0
0.600	5337	23510	319	834	96.2	86.5	98.7	5.6	3.4
0.700	5161	23550	279	1010	95.7	83.6	98.8	5.1	4.1
0.800	4928	23639	190	1243	95.2	79.9	99.2	3.7	5.0
0.900	4568	23685	144	1603	94.2	74.0	99.4	3.1	6.3
1.000	0	23829	0	6171	79.4	0.0	100.0	.	20.6

Fig. 14: Confusion Matrix for LA

Frequency
Percent

Table of I_pastdue by F_pastdue

I_pastdue(Into: pastdue)	F_pastdue(From: pastdue)		
	0	1	Total
0	15616 78.08	493 2.47	16109 80.55
1	242 1.21	3649 18.25	3891 19.46
Total	15858 79.29	4142 20.71	20000 100.00

Honestly, choosing a model is always hard. If we would like to predict the response in a very high accurate, different classifiers are supposed to be applied. In the fact, data is the more important than model. This is one of the reasons why we achieve the best result performing Logistic Regression Model is that the raw dataset has been transformed enough. The other reason why Radon Forest is much worse then k Nearest Neighbors and Logistic Analysis is also from the data. we have different proportion of observations in the response that almost 80% values of response variable are 0 term. Based on the algorithm of Random Forest, the new input tend to be predicted as 0 term that increases the error rate.

In the future, we should continue to compare other machine learning technique, such that Support Vector Machine and Deep Belief Network, with Logistic Regression Models.

VII. Acknowledgement

We would like to express our sincere gratitude to Equifax for providing the real word datasets.

References

[1] R. Babu and A. R. Satish. Improved of k-nearest neighbor techniques in credit scoring. *International Journal For Development of Computer Science and Technology*, 1(2), 2013.

[2] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[3] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, IT-13(1):21–27, 1967.

[4] J. A. Cruz and D. S. Wishart. Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*, 2:59–77, 2006.

[5] C. Huang. Using sas to find the best k for k-nearest-neighbor classification. *SAS Programming for Data Mining Applications*, 2011.

[6] X. Liang. K-nearest neighbor in sas. *SAS Programming for Data Mining Applications*, 2010.

[7] D. Memi. Assessing credit default using logistic regression and multiple discriminant analysis: Empirical evidence from bosnia and herzegovina. *Interdisciplinary Description of Complex Systems*, 13(1):128–153, 2015.

[8] L. E. Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

[9] Chandorkar P. Dsouza A. Rana, M. and N. Kazi. Breast cancer diagnosis and recurrence prediction using machine learning techniques. *International Journal of Research in Engineering and Technology*, 4(4): 372–376, 2015.

[10] D. Sharma. Improving the art, craft and science of economic credit risk scorecards using random forests: Why credit scorers and economists should use random forests. *Academy of Banking Studies Journal*, 11(1): 93–116, 2012.

[11] Chi D. Yeh, C. and Y. Lin. Going-concern prediction using hybrid random forests and rough set approach. *Information Sciences*, 254:98–110, 2014.