Regression classification for Unbalanced Data Sets: Some Performance Accuracy Methods

Matthew Yeseta

Department of Luddy School of Informatics, Computing, & Engineering: Indiana University

FA 590 Applied Data Science

Olga Scrivner

Dec 12, 2020

Summary of Study

The research paper by Hong Wang, et., al (Wang, 2015) addressed research to ask the question and test their hypothesis for a test model method for investigating and determining an improved model accuracy measure that would benefit loan regression classification models for scoring loan applicants, and to offer this findings to the public through their research evaluation outcomes.

Logistic regression prediction models in the banking and loan industry focus on credit scoring classification models for accurately scoring default / good customers from large sized big data. The challenge is to find approaches that improve performance of the prediction accuracy of these prediction models. Such a solution can assist banks and lending organizations to improve hedging free from potential high-risk financial loss stemming from default loan applications. Lending organizations' data teams have for years worked to improve algorithms and methods to improve prediction accuracy to minimize financial risk loss from analyzing the data, even un-balanced data, in credit scoring analysis in order to identify high probability, high-risk, default loan applicants (Wang, 2015, p. 1,2,3).

This study attempts to find and test a linear model for unbalanced data. The research paper by Wang, et, al. (Wang, 2015) used a credit scoring challenge, namely an "ensemble learning and Lasso logistic regression" algorithm model for handling un-balanced data, e.g., "majority subgroups" and "minority data", in order to deliver a subtle but improved accuracy performance than what has been more commonly found in many cross validation methods for credit loan prediction models (Wang, 2015, p. 3). Wang's Lasso regression approach uses AdaBoost, bagging, random forests, and gradient boosted machines in order to formulate and build a classification model that points on "how to balance the data and aggregate the diversified base learners (data) within the ensemble (model)" (Wang, 2015, p. 3, 4). Wang's Lasso logistic (LLRE) regression algorithm helped to resolve issues in classification in respect to loan scoring. As pointed out in the study by Wang, et. al., the Lasso-logistic regression ensemble (LLRE) learning algorithm permits practitioners of data science to investigate and glean insight from this clustering approach which "balancing the data" and "classify the diversity and aggregate the base class learners" in order to bring forth improved performance in credit scoring problems", based on a "balanced training data [in order to] maintain data diversity" (Wang, 2015, p. 17). Hong Wang's research outlines how the "Lasso-logistic regression ensemble (LLRE) learning algorithm" can solve performance issues of "unbalanced data classification problem in credit scoring" (Wang, 2015, p. 2).

This study shall apply a similar hypothesis to an unbalanced data classification dataset and try to utilize a similar publicly available algorithm approaches in order to ascertain if it is possible to show any improved model accuracy from publicly available algorithms. This studies effort ws at a disadvantage to equally compare by using the Wang Lasso logistic (LLRE) regression algorithm; but it was not found to be available as a library code, so this study was not able to equally compare.

The hypothesis question is stated to test models for accuracy performance and find if any can simulate similar behavior can show improved model accuracy from the base un-balanced data set. The hypothesis question is:

H0: Does a similar algorithm approach show improved model accuracy

HA: Does a similar algorithm approach fail to show improved model accuracy

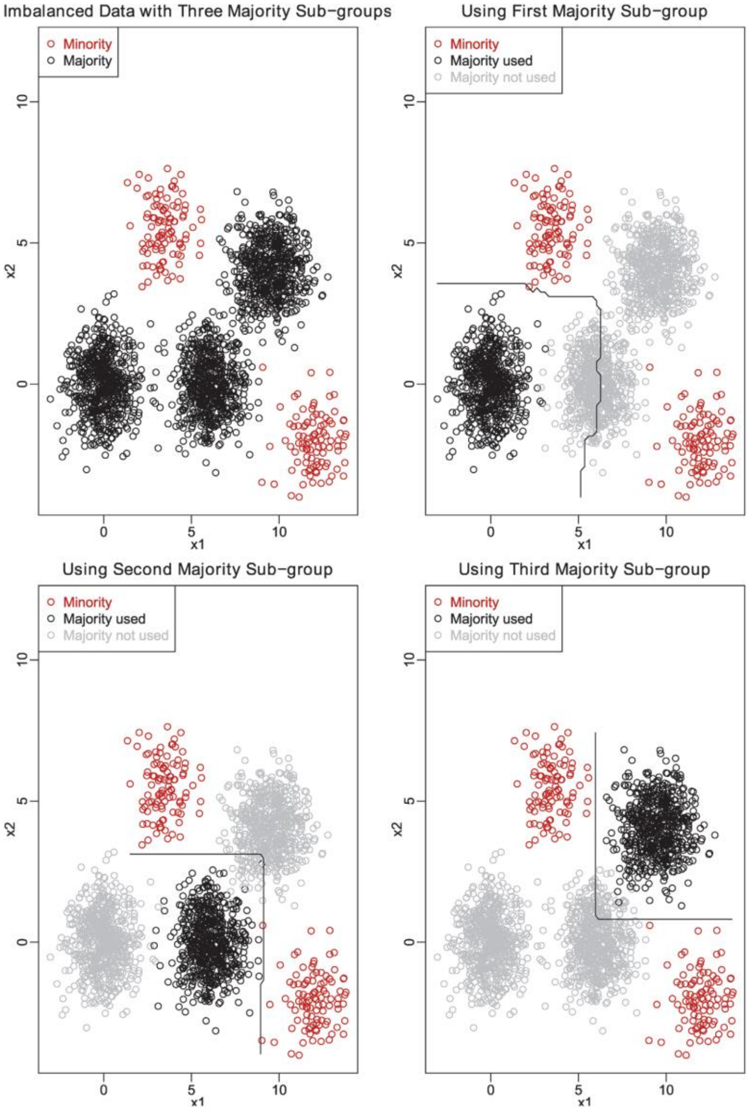
Methods (Both Research Paper and this Accuracy Study)

In Hong's research paper, the authors "demonstrated the effeteness of ensemble learning and Lasso-logistic regression", a Lasso-logistic regression ensemble, to "model large unbalanced data classification problem in credit scoring" (Wang, 2015, p. 2).

In Wang Hong's research paper, the authors proposed to use the AUC performance measure approach, indicating that "Area Under the Curve metric is independent of class distribution and is more suitable than accuracy in the scenario of unbalanced learning" (Wang, 2015, p. 2-3).

For building an appropriate large enough sized balanced data set, it is important to build a large number of “majority class cases” into the classification model, since these would represent non-default majority classes (Wang, 2015, p. 5). Hong Wang proposes that for an appropriate large data set, his method was to divide the “available majority examples into a number (k) of subgroups. Then select one of these subgroups, majority class cases, and include the full set of cases for the minority class cases to this data set, thus creating “a roughly balanced sub-training set”. “Base classifiers built upon these balanced sub-training sets should have better performance on both majority and the minority class example than classifiers trained with the original highly unbalanced dataset” (Wang, 2015, p. 5).

“Classification when majority are from different populations”, where the “balanced sub-training set, and all the training data excluding the currently used sub-training set are called Balanced Sub-training set (BS), and Out of Balanced Sub-training set (OBS) data” (Wang, 2015, p. 5).



Wang proposed a new ‘extension’ algorithm for Lasso-logistic regression ensemble that would “make full use of the available parallel computing facilities” (Wang, 2015, p. 8). This Lasso-logistic regression ensemble algorithm is outlined in Table 2 below:

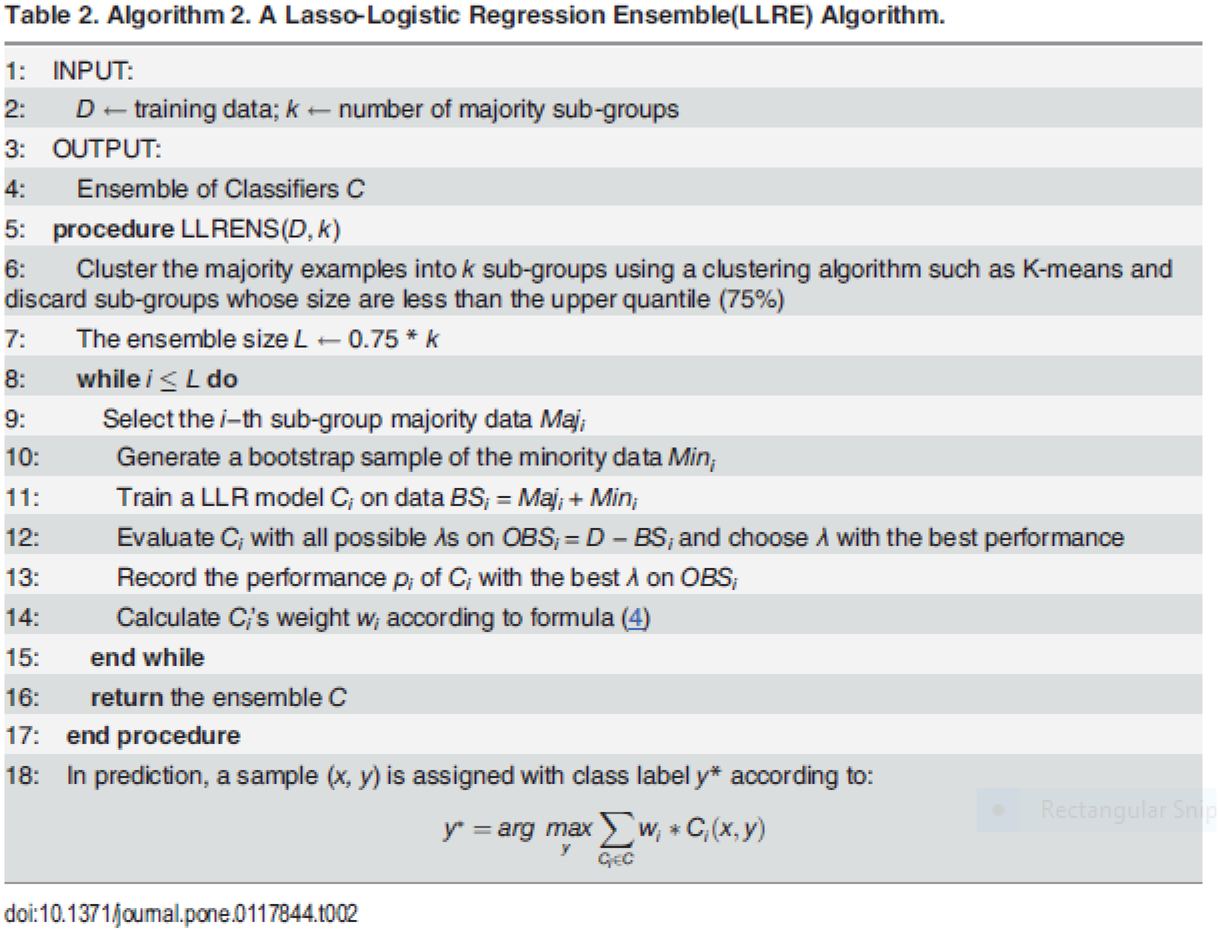


Table 2. Algorithm 2. A Lasso-Logistic Regression Ensemble(LLRE) Algorithm.

```
INPUT:
   $D \leftarrow$ training data; $k \leftarrow$ number of majority sub-groups
OUTPUT:
   Ensemble of Classifiers $C$
procedure LLRENS($D, k$)
   Cluster the majority examples into $k$ sub-groups using a clustering algorithm such as K-means and discard sub-groups whose size are less than the upper quantile (75%)
   The ensemble size $L \leftarrow 0.75 * k$
   while $i \leq L$ do
      Select the $i$–th sub-group majority data $Maj_i$
      Generate a bootstrap sample of the minority data $Min_i$
      Train a LLR model $C_i$ on data $BS_i = Maj_i + Min_i$
      Evaluate $C_i$ with all possible $\lambda$s on $OBS_i = D - BS_i$ and choose $\lambda$ with the best performance
      Record the performance $p_i$ of $C_i$ with the best $\lambda$ on $OBS_i$
      Calculate $C_i$'s weight $w_i$ according to formula (4)
   end while
   return the ensemble $C$
end procedure
In prediction, a sample $(x, y)$ is assigned with class label $y^*$ according to:
   $y^* = arg\ \max_y \sum_{C_i \in C} w_i * C_i(x, y)$
```

doi:10.1371/journal.pone.0117844.t002

Wang’s algorithm was designed to compute the ensemble of classifiers and cluster the majority class cases into k-sub-groups using cluster algorithm K-means. Then loop through each sub-group matrix of majority class cases and train on a LLR model, after which this algorithm was to record the performance of the trained LLR model and forward the ensemble LLR model for the next step prediction (Wang, 2015, p. 8).

For this study, a similar hypothesis pattern question is proposed with similar algorithm approaches to review and ascertain if we can find any improved model accuracy performance resulting from this study in its effort to imitate Wang’s research models.

For this study, the model algorithm approaches selected to test unbalanced classifications, all do handle un-balanced data sets. However, none of these model algorithm approaches were able to imitate as a 1-for-1 matching implementation with the Wang’s Lasso-logistic regression ensemble (LLRE) algorithm.

The key concept for classifications for un-balanced data is to test and deploy some appropriate model approach that can be effectually to generate a new data sets using either over or under balanced data sets in order to add proportion to this “minority class” of data instances in order to implement the creating of a more effective balanced data (Torgo, 2010). The goal for this balanced data set is to leverage an improved accuracy performance.

One implementation method for this approach was to train data for samples based on imputed data and utilize an over-sampling method. A second implementation method for this approach was to train data for samples based on imputed data and utilize an under-sampling method.

To “control the amount of over-sampling of the minority class and under-sampling of the majority classes” (Torgo, 2010) a synthetic over sampling (SMOTE) function was chose and utilized to “handle unbalanced classification” data where the “majority class examples are under-sampled, leading to a more balanced dataset” (Torgo, 2010).

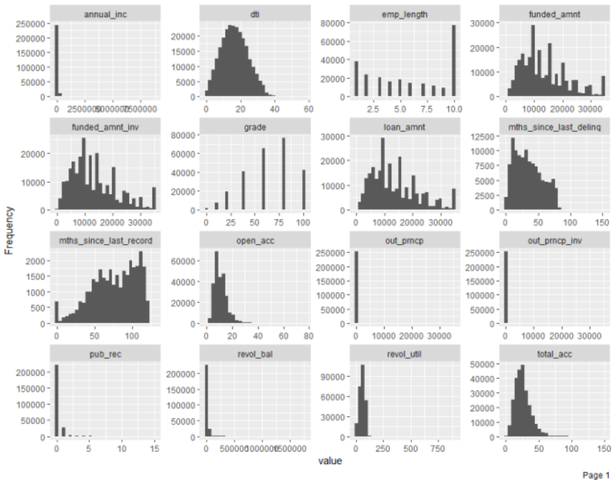
Results (Performance Accuracy Method Study)

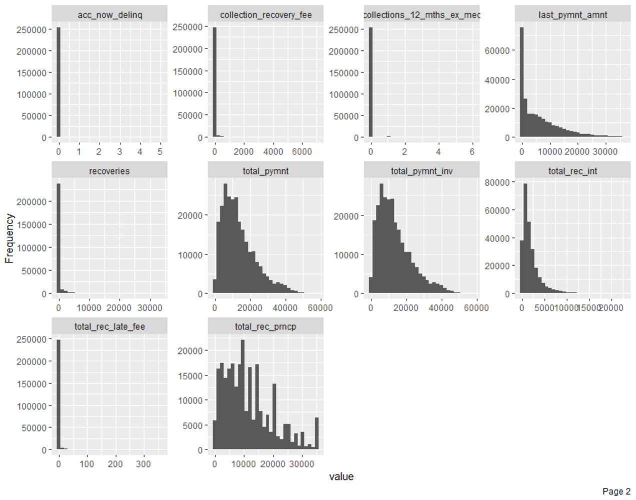
Exploratory Data Analysis

Exploratory Data Analysis for determining which continuous data is appropriate for the model build for the best feature variables. The data set source for this study is the Lending Club Loan Data (Kaggle).

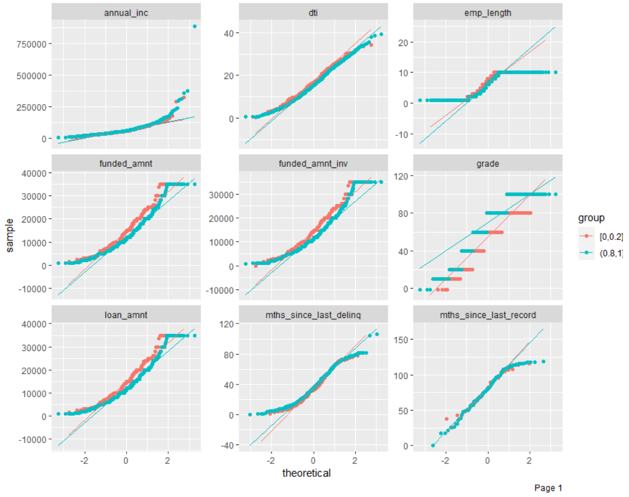
Generated was histograms summaries for determining the frequencies and value count for the continuous exploratory variables. Generated was QQ analysis charts, or quantile-quantile plots, to give us a visual view of the distribution on whether it is normal or exponential, if the response variable is normally distributed.

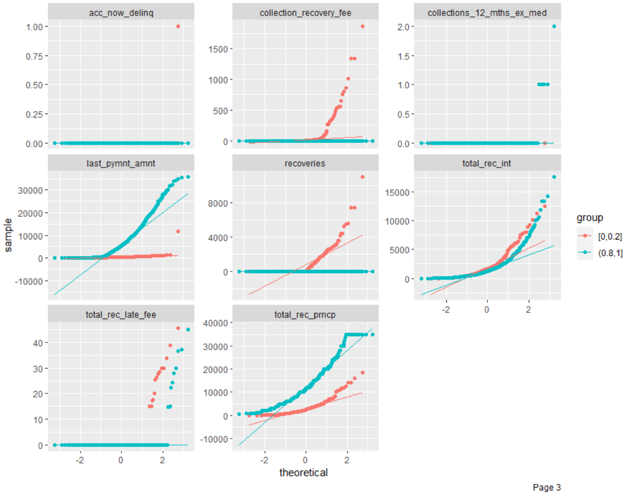
Histogram Frequencies for visual analysis on continuous exploratory predictor variables.

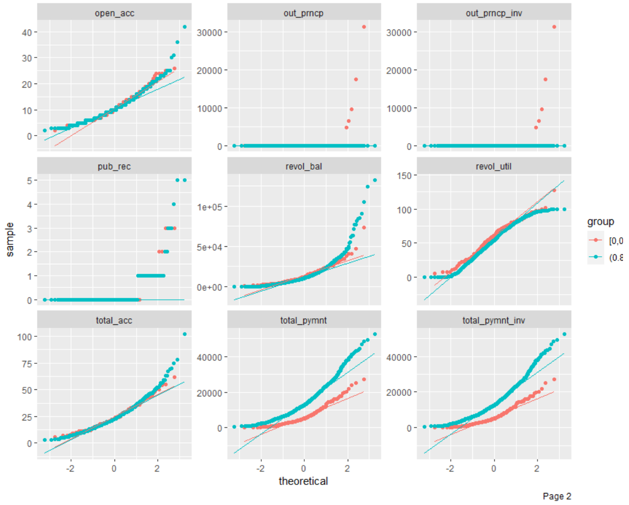




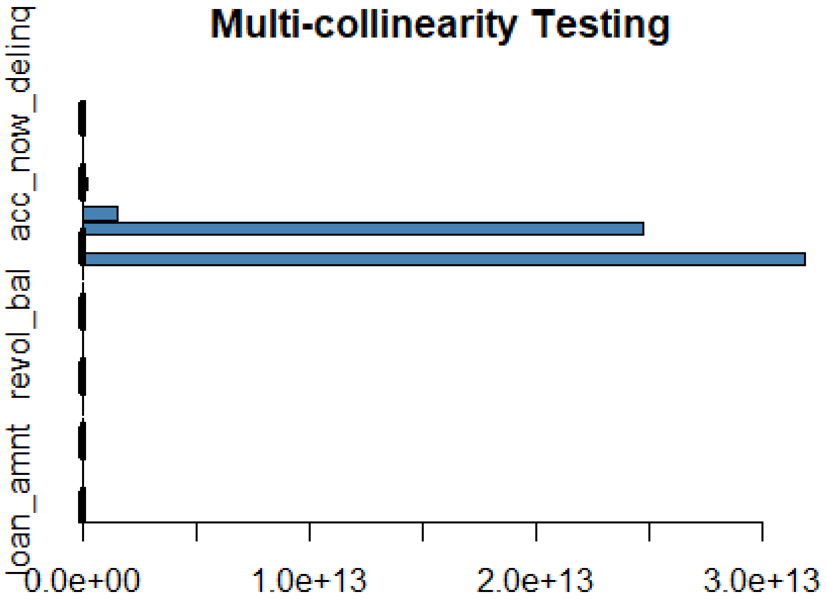
QQ analysis charts provide visual determination on whether the dependent variable is normally distributed or exponential distribution. Several variables seen in these QQ charts exhibit not to be normally distributed as highlighted by flat line at bottom of several QQ charts. Several variables succinctly appear to be normally distributed since the data points fall along a straight line together with the response variable.





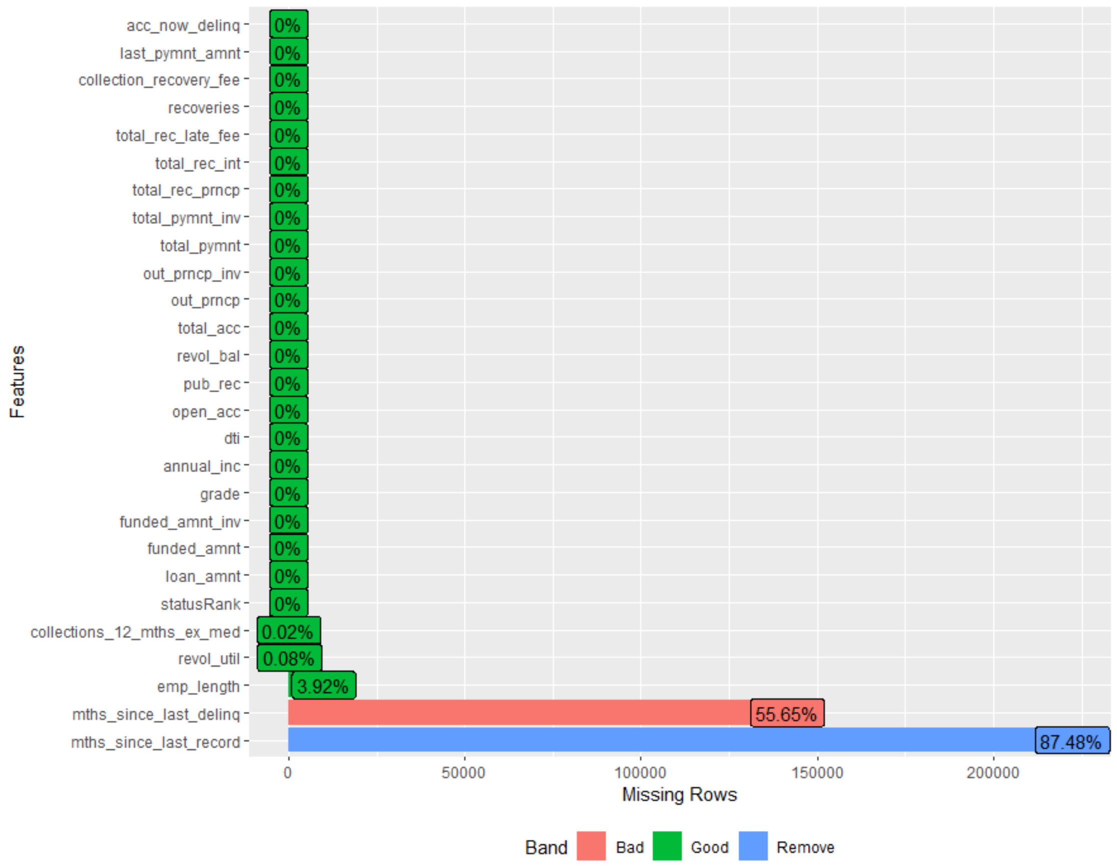


Multi-collinearity test to ascertain if any of the variables are linearly related by examining the correlation of variables for high correlation. From this test most variables were not sources of multicollinearity. This multi-collinearity test determined accurately the exploratory feature variables to select for the project modeling phase.



Missing Data Analysis

Missing data analysis for the exploratory continuous predictive variables data shows only four variables with slight missingness of values; two of which have significantly high missingness. One candidate, at 55% and another at 87%, respectably. The missing imputation algorithm for this study data set utilized the outperforming predictive mean matching method for imputing missing data values.



Formulate Hypothesis Tests

For the hypothesis test in this project the project asked the question: can some similar algorithm approach demonstrate some improvement to the model accuracy. The hypothesis question is:

H0: Does a similar algorithm approach show improved model accuracy (AUC)

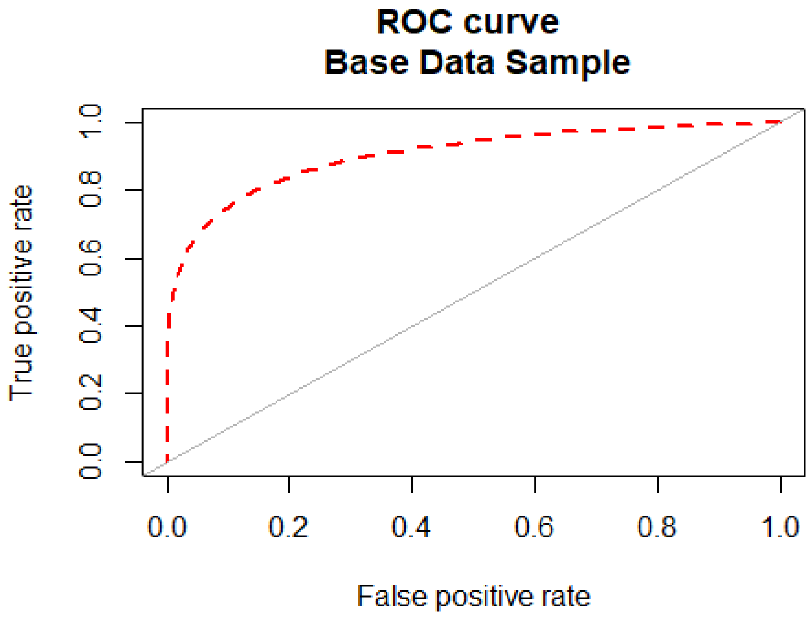
HA: Does a similar algorithm approach fail to show improved model accuracy (AUC)

Results of Classifier Methods

AUC for baseline data accuracy measure outcome

The baseline data, e.g., raw data. Before starting on the various methods of un-balanced data for accuracy performance, it is necessary to establish the baseline un-balanced data and accuracy (AUC) which is manifested without any treatment methods. The baseline data reported precision, recall, F statistic, and Area under the curve. The baseline accuracy AUC is 0.904, which is below the 95%, or the alpha 0.05 standard level.

precision	0.871	recall	0.969
F statistic	0.459	AUC	0.904

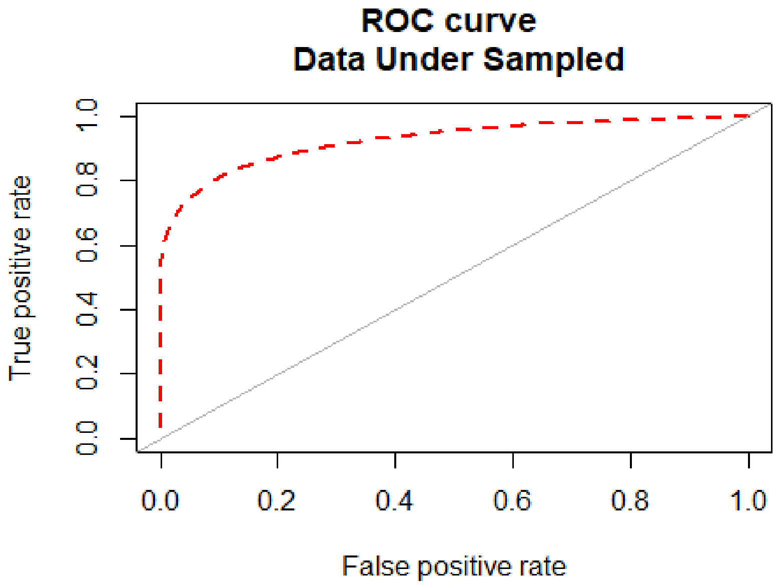


AUC Under Sample Balancing measure outcome

Under sample balancing approach is done without replacement. In this method, good transactions are equal to fraud transactions. Hence, no significant information can be obtained from this sample.

One chosen method for to the challenge of un-balanced data is a standard under-sample approach. The under-sample data reported precision, recall, F statistic, and Area under the curve. The under-sample approach accuracy AUC is 0.925, which is below the 95%, or the alpha 0.05 standard level.

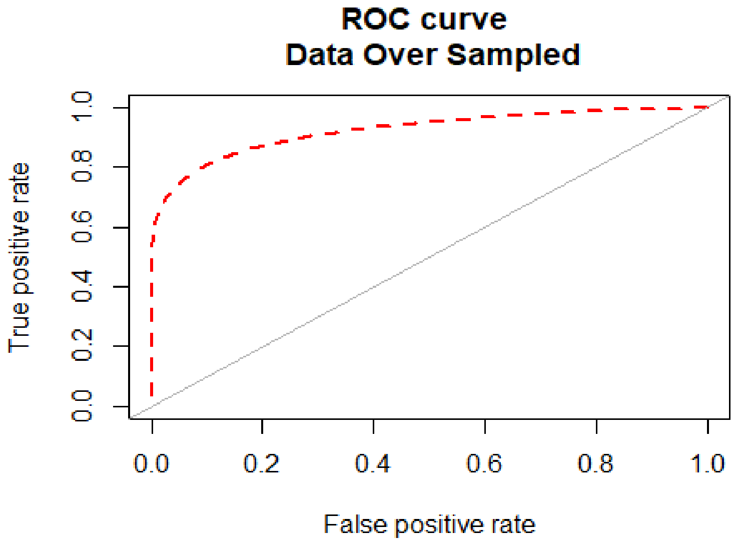
precision	0.895	recall	0.804
F statistic	0.423	AUC	0.925



AUC Over Sample Balancing measure outcome

Over sample method approach performs delivery over sampled in the minority class data. The over-sample data reported precision, recall, F statistic, and Area under the curve. The over-sample approach accuracy AUC is 0.924, which is below the 95%, or the alpha 0.05 standard level.

precision	0.893	recall	0.802
F statistic	0.422	AUC	0.924

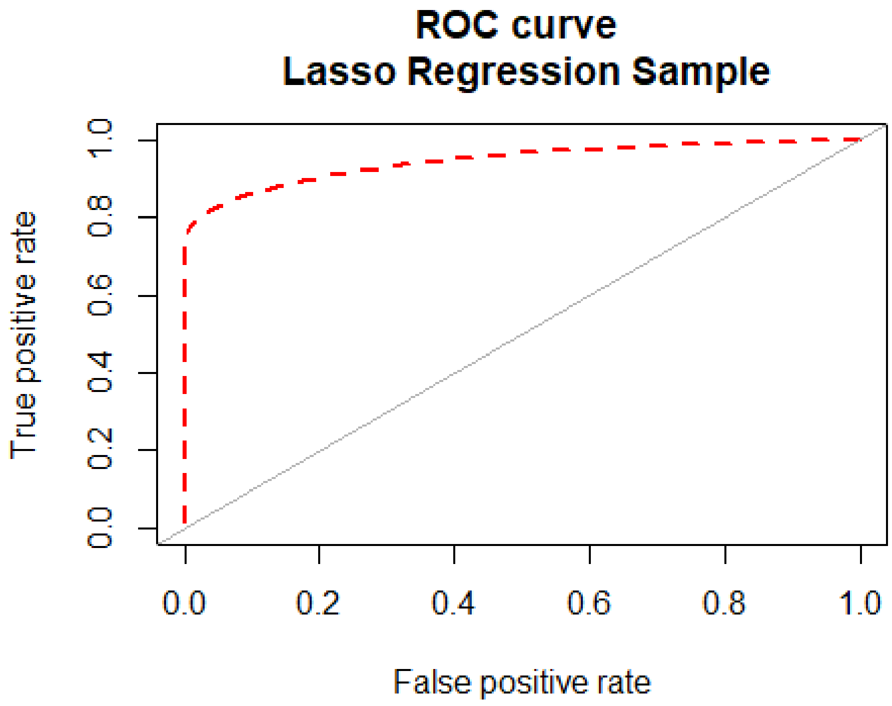


However, the conclusion of the under-sample and over-sample test reveal slightly higher AUC accuracy performance in comparison with the baseline model. Therefore, from the perspective of both under-sample and over-sample tests, there is sufficient evidence, below the alpha 0.05, as a result, to say that the conclusion analysis is sufficient to reject the Null hypothesis since evidence is not sufficiently found in the AUC measured test of the accuracy performance which is lower than the alpha 0.05 significance level. (H0: Does a similar algorithm approach show improved model AUC accuracy)

AUC Lasso Regression Model Balancing measure outcome

A Lasso regression model, which is the baseline algorithm that Wang (Wang, 2015) expanded on in his research experiments, Lasso-logistic regression. A standard classifier Lasso regression shall be utilized in this project also. This Lasso regression algorithm implements a penalization approach to reduce, under specific conditions, to set some coefficients to zero. Furthermore, the Lasso model is instrumental to assist with feature selection for identifying fewer more succinct predictor variables. For our model balancing accuracy AUC is set at 0.945, which is excellent for this data set, and a noted improvement over standard over/under balance method (as seen on previous pages).

precision	0.952	recall	0.903
F statistic	0.464	AUC	0.945



AUC Synthetic Over Sampling SMOTE Method measure outcome

The next chosen method to the challenge of un-balanced data the Synthetic Over Sampling (SMOTE) algorithm, which is an instrumental algorithm to assist in handling classification problems on un-balanced data sets (Torgo, 2010). This was for comparative analysis for measuring classification accuracy performance between over/under and SMOTE unbalanced classification approaches.

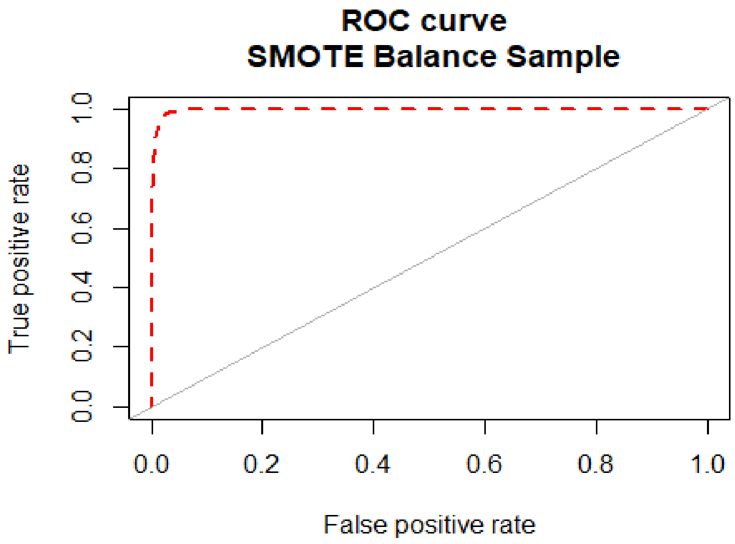
The SMOTE algorithm is utilized to measure the unbalance original data set by applying “new examples of the minority class using the nearest neighbors” for the resulting balanced data cases (Torgo, 2010). The SMOTE library is directed at problems of class unbalance, and therefore may be a suitable alternative in this study’s Hypothesis test (Torgo, 2010).

In this SMOTE Algorithm test, the percentage over-sample is set such that for each case in span of minority class instances, the percentage over-sample configuration will create a random proportion to be added to this minority class of data instances. The percentage under-sample sets a proportion of minority cases to be randomly selected for a newly generated balanced data set (Torgo, 2010).

The majority class cases in this sample split outcome was 38,515 for “1” records, and the minorit y class cases was 8,559 for “0” records; this represents a ratio that is unbalanced.

There is enough evidence from the SMOTE algorithm to warrant AUC performance accuracy improvement on its minority class.

precision	0.578	recall	1.000
F statistic	0.366	AUC	0.996



The conclusion of the SMOTE algorithm test revealed the best AUC accuracy performance, over the other respective models inside this study. Therefore, from the perspective of this test, the hypothesis test does have sufficient evidence to support accepting the hypothesis, therefore, the conclusion is to accept the Null hypothesis since evidence is sufficiently found in the AUC measured test of the accuracy performance. The SMOTE algorithm did have evidence of AUC performance accuracy improvement on its minority class.

Summary of Collection of Classifier Method Approaches

This study performed a pattern of experiments using some standard R packaged classifiers, the results in terms of AUC accuracy performance, and did demonstrate better performance for AUC (and F statistic measure) for both Lasso Regression and Synthetic Over Sampling (SMOTE) algorithm for unbalanced data in this study. This outcome shows measurement using standard R classifiers can delivery some decent performance for AUC. Yet our project was not able to confirm the equal experiment results that Wang (Wang, 2015, p. 5) demonstrated, but this is promising even without using Wang's LLRE algorithm.

As with Wang's research, this study has found at least one algorithm method that can closely simulates and realizes an advantage to increase a noticeable performance improvement in the AUC for unbalanced data, namely, the Synthetic Over Sampling (SMOTE) algorithm for unbalanced data. The SMOTE algorithm reported AUC 99.6% provided sufficient evidence to permit this project to accept the Null Hypothesis, e.g., H0: Does a similar algorithm approach show improved model AUC accuracy?

Classifier Method	Generated AUC	Generated F Statistic	
Under Sample	0.925	0.423	
Over Sample	0.924	0.422	
Lasso Regression	0.945	0.464	
SMOTE	0.996	0.366	

In comparison to this study, Wang's research used "experiment results in terms of AUC" used AUC algorithm measurements for the following algorithms: LLRC, LLR, RF, CART (Wang, 2015, p. 5), seen in the performance on two variable sets, see Wang's paper Table 3.

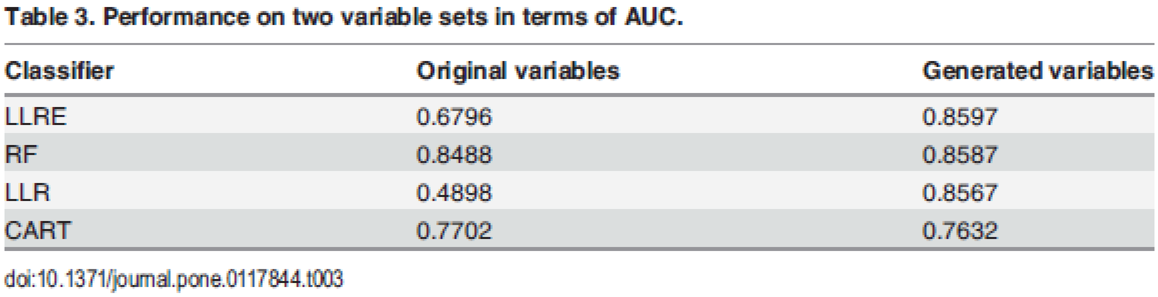


Table 3. Performance on two variable sets in terms of AUC.

Classifier	Original variables	Generated variables
LLRE	0.6796	0.8597
RF	0.8488	0.8587
LLR	0.4898	0.8567
CART	0.7702	0.7632

doi:10.1371/journal.pone.0117844.t003

Wang’s research experiments for the Lasso-logistic regression ensemble (LLRE) learning algorithm had “demonstrated overall better performance across the popular AUC and F-measure for unbalanced credit scoring” Wang’s LLRE for AUC measures “outperforms RF, LLR and CART significantly” (Wang, 2015, p. 16). It is evident that Wang’s experiments for Lasso-logistic regression ensemble (LLRE) delivered “better performance” and “should make an adequate gain in profits for financial institutions” (Wang, 2015, p. 16).

In order to reach this level of success in delivering a better performance model with the Wang’s Lasso-logistic regression ensemble (LLRE) learning model, the performance outcome they observed was established and built upon a strong foundation of a careful exploratory data analysis and feature selection for predictive variables, and the utilization of domain subject matter experts (Wang, 2015, p. 17).

Challenges

One such challenge is to find and utilize Synthetic Over Sampling (SMOTE), Random Forest, Lasso Logistic Regression algorithms to match a similar comparative analysis for measuring classification accuracy performance. This paper provides a base line for further ensemble algorithm testing and comparative analysis between Synthetic Over Sampling (SMOTE), Random Forest, Lasso Logistic Regression, CART, and other unbalanced algorithms.

Further challenge was to analyze the predictive variables and perform a deep careful exploratory data analysis for building a model for ensemble logistic regression (Wang, 2015, p17).

Since the authors’ pointed out that their Lasso-logistic regression ensemble (LLRE) learning algorithm in actual experiment executions has “demonstrated that LLRE beats CART, LLR and RF in terms of AUC and F-measure”, therefore, to evaluate on further studies then any effort would require at least the same utilization of this Lasso-logistic regression ensemble (LLRE) learning algorithm. This translates into either finding another existing implementation algorithm to be suitable for a comparative analysis or enlist the authors to share their algorithm.

In order to build a sound credit scoring analysis approach for testing and performing comparative analysis then it would require proper “clustering the majority data and bagging the minority data…[in order] to generate balanced training data and [to] maintain data diversity” (Wang, 2015, p17). This emphasis is necessary for all and any logistic regression ensemble learning study when high ratio un-balanced data is represented in the domain data samples. This shows why careful exploratory data analysis is an essential as a building block for selecting important predictive variables that impact the model accuracy (such as: AUC, confusion matrix, or other preferred (or required) performance accuracy assessment tool).

Future directions

The authors, Wang, Zu, Zhou, conclude that it is imperative for future direction to focus on more research and collaboration with other researchers in order to devise innovative fast training models and designs that are able to yield improved majority class sub-groups classifier solutions. This was pointed out due to the observation there has been little research developing innovative logistic regression solutions as the base classifier for credit scoring studies, which should “alleviate the class imbalance problem”, once such learning algorithms are in common use in the industry (Wang, 2015, p1, 2, 17). “Neural networks, support vector machines, logistic regression models are frequent” choices “as base learners in credit scoring ensemble algorithms”. The authors proposed that many or most learning algorithms today focus *more incorrectly* on the majority class and for that cause results in lower accuracy performance when the minority class are not equally studied in credit scoring applications with large ratio un-balanced data (Wang, 2015).

Reflection

A important surprise that I learned is that most all standard learning algorithms focus more on the majority class, and not much on the minority class, and that this often causes poor performance because the analysis is not thinking broadly regarding the focus on the minority class. In any un-balanced data, the major impact on this un-balanced data is the focus on both majority class and minority class. Therefore, for any ideal credit scoring model the team thinking must focus equally on both on the minority class and majority class data. This succinct thinking, majority class and minority class focus, is necessary to select an appropriate classifier method, e.g., such as logistic regression, support vector machines, random forests, gradient boosting. Hong Wang, Qingsong, Xu, Lifeng Zhou, authors delineated this principal inside their research paper as a call to other researchers to invest into future classifier solutions that can delivery to “alleviate the class imbalance problem” (Wang, 2015, p1, 2, 17).

References

Wang H, Xu Q, Zhou L (2015). Large Unbalanced Credit Scoring Using Lasso-Logistic Regression Ensemble. *PLoS ONE 10(2)*: e0117844. https://doi:10.1371/journal.pone.0117844

Luis Torgo, https://rdrr.io/cran/DMwR/man/SMOTE.html, “SMOTE algorithm for unbalanced classification problems: in DMwR: Functions and data for Data Mining with R”

Torgo, L. (2010), *Data Mining using R: learning with case studies*, CRC Press

Kaggle. Lending Club Loan Data: 2007 through current Lending Club accepted and rejected loan data, Kaggle, https://www.kaggle.com/wordsforthewise/lending-club