Credit Approval Prediction

Aditya Rathi
adityamr@andrew.cmu.edu

Rishabh Pahuja
rpahuja@andrew.cmu.edu

Vatsal Joshi
vatsalkj@andrew.cmu.edu

Abstract

Increased access to financial technologies such as credit cards and their rapid proliferation in all socioeconomic backgrounds have made the demand for credit higher than ever before. However, a large section of the population still have little to no credit history due to multiple socioeconomic factors. Hence, to cater to this demographic we have developed a system that can predict an applicant's credit worthiness using commonly available data points such as marital status or number of dependants. For this task, we have implemented a hybrid machine learning model based on DBSCAN in the first stage and several supervised learning methods in the second. These models have been trained and tested on a host of data-sets from multiple sources to ensure the consistency of results and validate the effectiveness of our strategy. The data-sets were manually cleaned and filtered and their dimensionality was reduced using PCA to both improve the computational time and to allow visual tuning of DBSCAN. By tuning the DBSCAN parameters, viz. the maximum allowable distance between data-points and the minimum number of data-points per cluster we were able to successfully detect and remove the outliers from our data-sets. The various supervised learning methods yielded an accuracy of around 70% *on the data-sets before outlier removal and had an increase in accuracy of* $3-5\%$ *after outlier removal. It was also observed that different supervised learning methods have different sensitivity to outliers. Methods such as logistic regression show the most sensitivity due to the high bias of the outliers while ensemble methods such as random forest show the least because of the averaging effect of the voting between candidates.*

1. Introduction

With the increasing proliferation of financial technologies such as credit card, the demand for easy access to credit is higher than ever before. Thus, modern day credit evaluation systems cannot rely on judgmental approaches that subjectively evaluate every individual. Hence, methods that can objectively and reliably extract an overview of the applicant's creditworthiness without actually pulling up their credit record, are in increasing demand. Furthermore, in this regard, low-income and young individuals often do not have a credit score due to multiple factors such as lack of income history or improper documentation. In the US alone there are over 25 million Americans who were credit invisible in 2015 according to the CFPB [3].

The ability to quickly predict an applicant's creditworthiness and judge the risk associated with lending to him/her will naturally be very useful to financial institutions. As an added advantage, individuals will also have unprecedented opportunity to monitor their credit score and take steps to improve it. With demand from both institutions as well as individuals we believe that this is a meaningful challenge to tackle. We plan to build a credit score predictor targeting these demographics in particular. Hence, unlike typical credit score calculators such as FICO, we plan to use data that most people will be able to easily provide irrespective of their credit record such as their current income, marital status, number of dependents amongst others. The output of our model will be the predicted credit-worthiness of the applicant. By training our model with sufficient data we hope that our model will be able to predict the credit-worthiness of applicants with a high degree of confidence making the entire exercise practically useful.

There has been work carried out in this area focusing on trying to create models to predict credit-scores of applicants such as [8, 10, 9, 1]. The works done before either use a single model to test their results, do not try out hybrid models or do not compare with multiple data-sets making their results not as robust or relevant. We have carried out all of these techniques viz. compared our results over multiple algorithms and across multiple data sets and we have also implemented hybridized models for added robustness and efficiency. Lastly, we have also carried out principal component analysis on our data-points to reduce the number of required features while minimizing loss of information thereby allowing us to visualize our outlier classification more effectively.

Regarding the data, we plan to use data-sets that have these features to train our model. We have collected several data-sets from Kaggle, the UCI machine learning repository and other sources which are explained in detail in

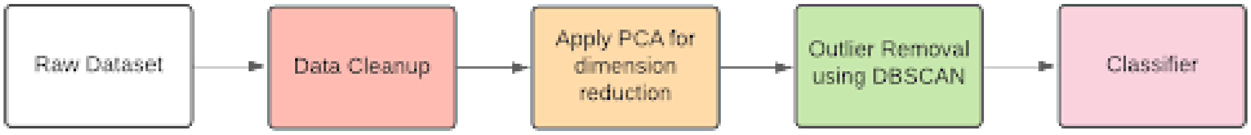

Raw Dataset
Data Cleanup
Apply PCA for dimension reduction
Outlier Removal using DBSCAN
Classifier

Figure 1: Workflow of the model

the methodology section. Several works show that correct feature selection greatly improves efficiency of the model [2, 4, 5, 6, 7]. Hence, we believe that devoting resources to finding the optimum set of features will help to increase the efficiency of the model that we create. Furthermore, most credit scoring datsets are drastically imbalanced with a large majority of the data points focusing on rejections rather than acceptances. This is just the nature of the data available as credit applicants are more likely to be rejected rather than accepted. Hence we have tested out various machine learning methods such as random-forests or logistic regression to figure out which perform well with such data imbalance. Furthermore, to improve the efficacy of the machine learning methods we have implemented hybrid machine learning models with outlier removal.

Model evaluation has been carried out using the Percentage Correctly Classified (PCC) metric. Dimensions of the input data were reduced using PCA, while keeping the variance explained over 97%. After applying the hybrid models their results were compared with the vanilla algorithms. From the results it was seen that all of the vanilla models perform consistently over multiple data-sets and $3-5\%$ of improvement was produced when applying the updated data-sets.

The rest of the paper discusses the methodology and the results in detail. It is organized as follows: Section Two talks about the methodology followed in the paper along with data collected, Section 3 talks about the results and discussion and finally section 4 talks about conclusion.

2. Methodology

The orignal data-set was processed into a form so that it can be used by different models without any bias. Dimensional reduction was carried out on the processed data to reduce calculation time and so that data can be visualized. After dimensional reduction, the outliers were removed using an unsupervised method. Subsequently, several supervised machine learning models were applied on data-sets with and without outliers. The accuracy scores on test dataset were recorded for each method.

2.1. Datasets Collection

For this project, we required data samples which comprise of features that not only indicate a generic financial situation of people, but also represent their lifestyle such as condition of their car (new model, second-hand, etc.), type of job and many more. 8 data-sets suited the aforementioned requirements and were chosen to execute this project. The German, Japan, and Australian data-sets were acquired from UCI machine learning repository. The PAKDD dataset was obtained from PAKDD 2009 Data Mining Competition. Econometric analysis data was taken from Kaggle. The mortgage and Home equity data-sets were obtained from Credit Risk Analytics. Table 1 given below gives information about all the data-sets used.

Table 1: data-sets Used

S.No.	data-set	data points	Features
1	UCI-German	1000	24
2	UCI-Australia	690	14
3	UCI-Japan	690	9
4	PAKDD-2009	50000	51
5	Kaggle-Econo. Analysis	1320	11
6	Thomas et al., 2002	1226	14
7	CRA-Home equity	5960	12
8	CRA-Mortgage	50000	22

As mentioned earlier, these data-sets consists of non financial features like age, gender, number of vehicles owned, number of dependents, etc. so that the possibility whether an applicant can get a loan or not can be predicted by using the non-financial data in case the applicant has no credit score.

2.2. Data-sets Processing

The raw data-sets had a lot of junk data, as well as nonessential features. Hence, these data-sets could not be used in their original format. Data cleanup was, therefore, performed to make the data usable to generate good results. The issues with the data were:

- Redundant features like serial number, applicant ID, etc.

- Missing values for some data-points
- Categorical/Non-numeric features

We overcame these issues in the following manner:

- Redundant features were removed from the data-set
- Missing values were either averaged out for a feature or the entire data point was removed in case the features could not be averaged
- One hot encoding was performed on categorical features to convert these features to vectors. No numerical label was assigned to the categorical features because that would have result in bias in the data-set

2.3. Dimensionality Reduction

As it can be seen from Table 1, the data-sets chosen had a lot of features, ranging from 9 to 51. Even though having so many features enables prediction in case of missing values for an applicant but at the same time, it gets difficult to visualize these features, and locate outliers. Moreover, Reducing the number of dimensions before outlier detection improved the computation time significantly. So the first step after pre-processing the data-set was to apply dimensionality reduction techniques. It was performed using **Principal Component Analysis** or **PCA**. The new number of dimensions were chosen on the basis of percentage of explained variances. A threshold value of 97% was taken and for all the 8 data-sets, two dimensions were taken, with variance values ranging from 97% to 99% for the two dimensions. Having high values of explained variances ensured that the most of the data was preserved, and we can safely assume that our models will provide us with accurate results even with lower dimensions.

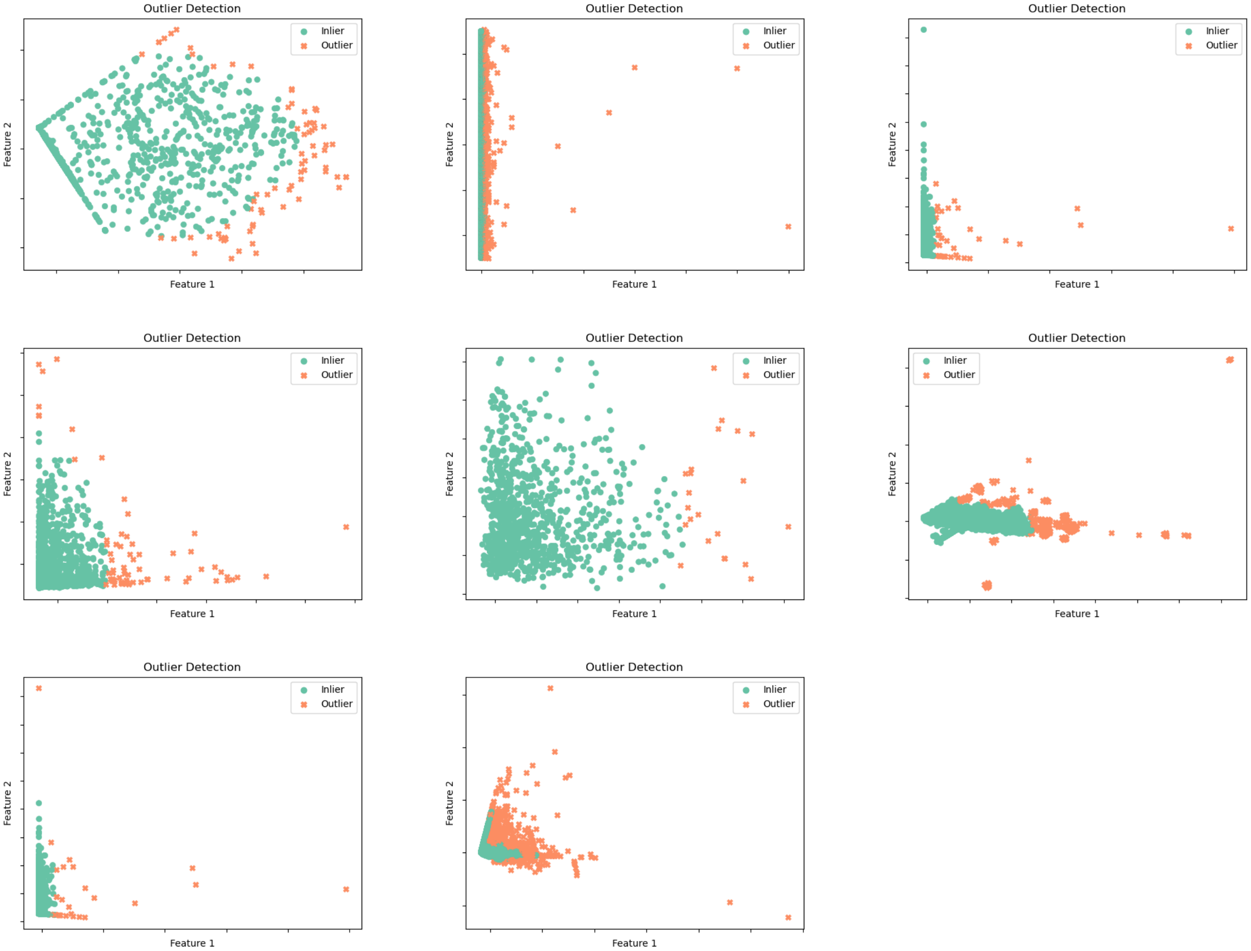
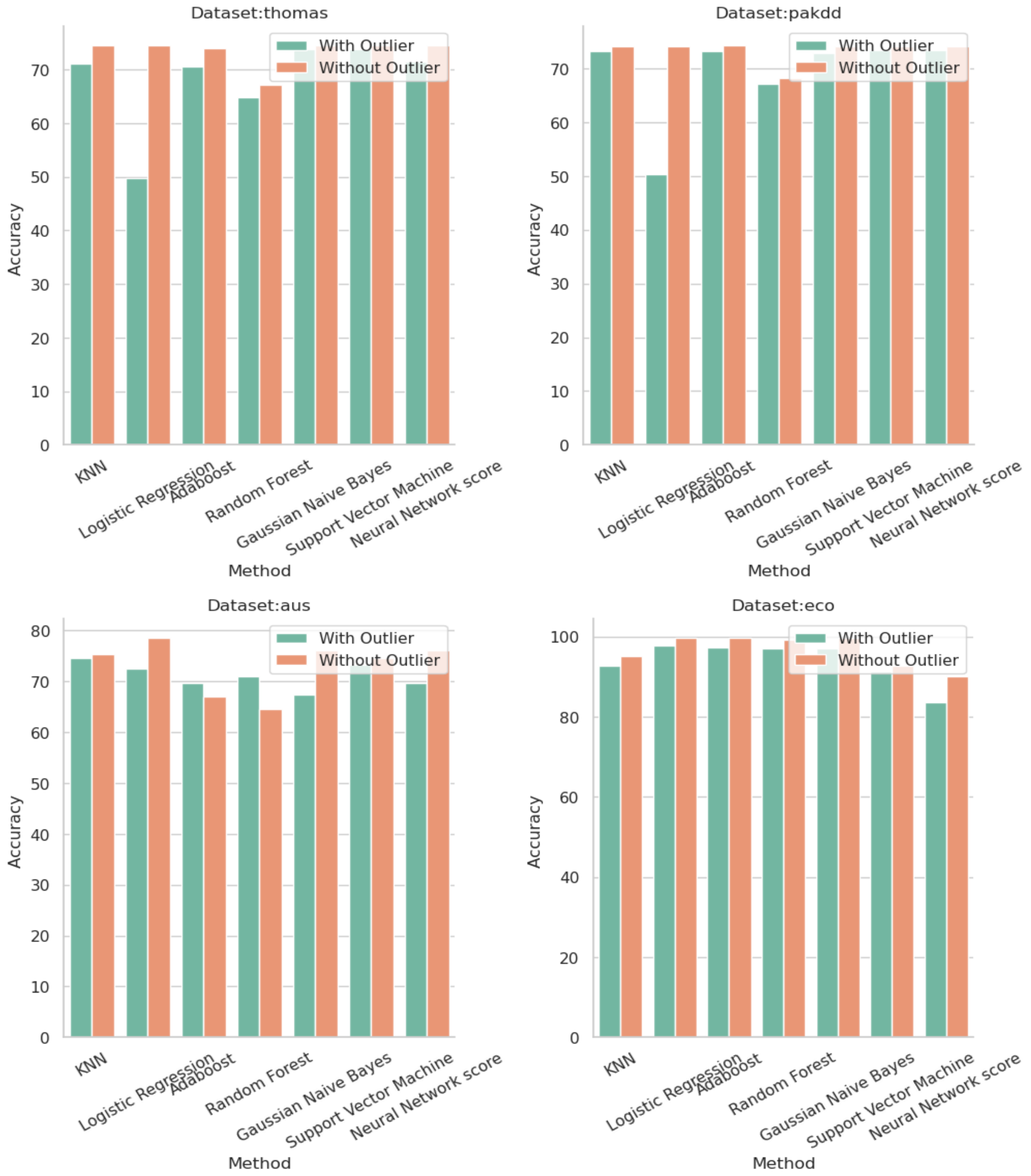


Figure 2: From top to bottom and left to right: (A) Thomas data-set, (B) PAKDD data-set, (C) Australian data-set, (D) Econometric data-set, (E) German data-set, (F) CRA-HMEQ data-set, (G) Japanese data-set, (H) CRA-Mortgage data-set

2.4. Outlier Removal

After dimension reduction, all the data-sets were plotted to visualize the presence of outliers in the data-set. Outliers were found in all the data-sets and were recognized and removed by **Density-based spacial clustering of applications with noise** (DBSCAN). DBSCAN is an unsupervised clustering algorithm that groups together points that



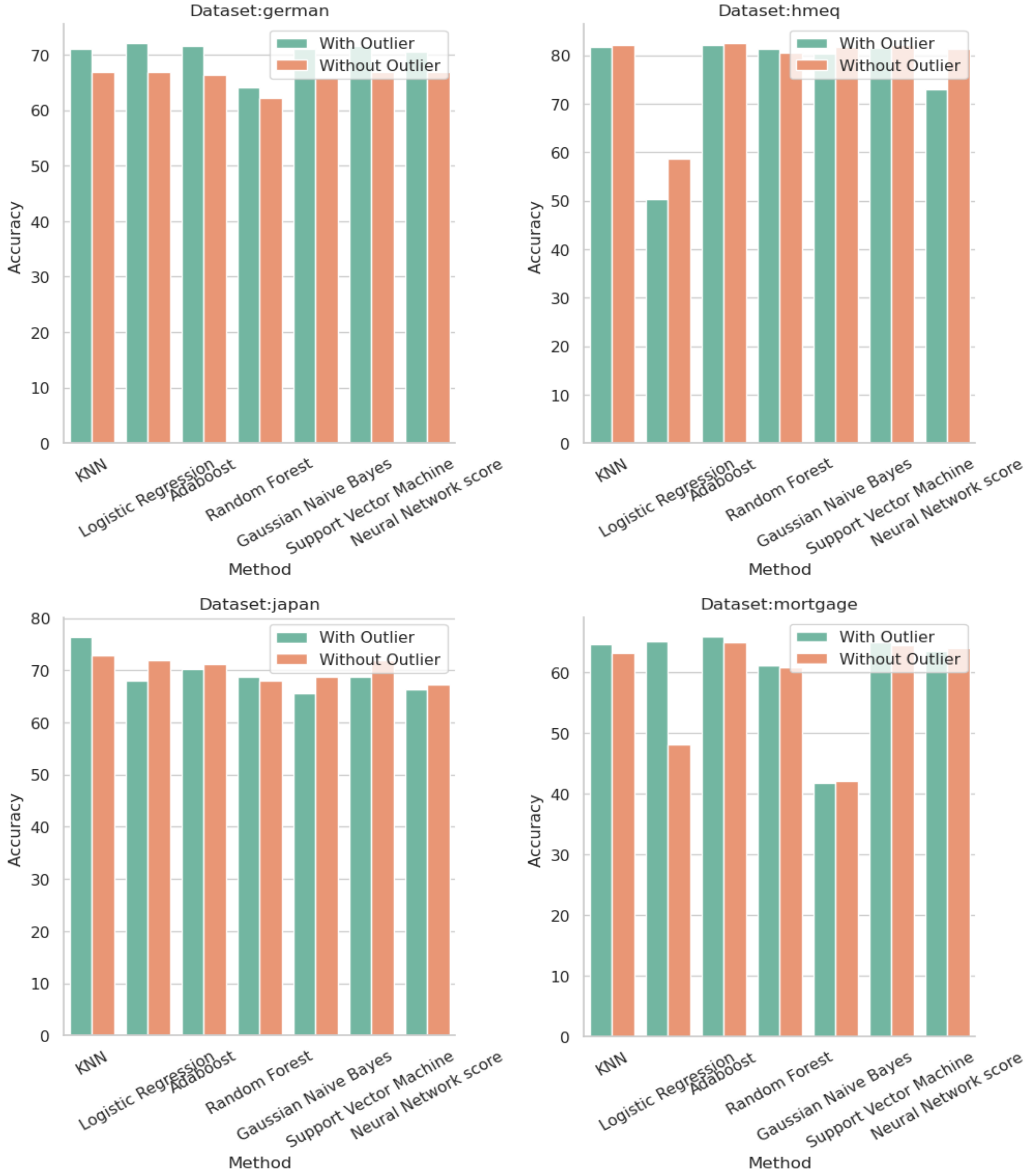


Figure 3: Accuracy obtained before and after removing outliers. Data-sets from top to bottom and left to right: (A) Thomas, (B) PAKDD, (C) Australian, (D) Econometric, (E) German, (F) CRA-HMEQ, (G) Japan, (H) CRA-Mortgage

are closely packed together and marks points in low density region as outliers. One of the major reasons why DBSCAN was used for outlier removal is that one does not need to inform DBSCAN the number of groups into which the data-set should be divided in. This reduces the chances of over-fitting. DBSCAN was performed using scikit-learn package which has in built class to perform DBSCAN. The parameters used for DBSCAN were eps and $min\_samples$, where eps stands for the maximum distance between two samples for one to be considered as in the neighborhood of the other and $min\_samples$ refers to the number of samples in a neighborhood for a point to be considered as a core point. Both of these parameters were tuned for all the data-sets so that maximum outliers and minimum inliers were removed from the data-sets. Fig. 2 shows the data points that were considered outliers by DBSCAN and were removed from the respective data-sets. The data-sets shown in Fig. 2 are obtained by applying PCA to the pre-processed data. It can be clearly observed that the nature of all the data-sets is different from each other, and hence, DBSCAN is required to be tuned for individual data-sets. For example, the PAKDD data-set shown in Fig. 2(B) required a lot of fine tuning of parameters as large numbers of data-points were concentrated in one region. On the other hand, data-set such as the one shown in Fig. 2(A) was more straightforward to remove outliers.

2.5. Supervised Learning

Initially, we planned to use unsupervised learning on all the data-sets to create clusters and then followed by linear regression to predict the credit score of each individual. To execute this we wanted to compare the accuracy of the hybrid model. Since credit score is a sensitive information, we did not find data-sets containing the credit scores of each individual to compare our results with. Therefore, we proceeded with the good/bad credit classification problem.

All data-sets were divided into randomized training and testing datsets with 80-20 split using in built function train_test_split(). Several machine learning models such as logistic regression, k-nearest neighbors (KNN), Adaboost, Random Forest Classifier, Gaussian Naive Bayes, Support Vector Machine, and Neural Networks were used to classify the samples with and without outliers. The test accuracy of all the models were recorded and plotted against each other. All these models were applied using scikit-learn.

3. Results and Discussion

The authenticity of our model was measured by PCC plots which compares the test accuracy of both the hybrid model and the supervised learning model with removing the outliers. These are shown in Figure 3. We found some interesting results upon observing the difference in accuracy of both the models. Firstly, most of the supervised learning models achieved around 70% accuracy on the testing data-set before outlier removal. After removing outliers, most of the methods were able to generate an accuracy improvement of $3-5\%$ in all of the data-sets.

We also observed that there was a massive increase in accuracy for some of the training methods. For example, logistic regression is observed to be the most susceptible to outlier removal. This can be explained by the fact that outliers contribute to a lot of bias while constructing a decision boundary. Therefore when the outliers were removed, the biases generated by those samples were also eliminated. Hence, we observe a huge boost in the test accuracy when the DBSCAN was applied in case of logistic regression. However, we observed a few exceptions in some of the models for a few data-sets. In the mortgage data-set, we observe that the accuracy of logistic regression plunges significantly. The plausible reason for this anomaly is the strong resemblance of the testing data to the outliers in the training data-set.

Unlike logistic regression, the ensemble methods were nearly immune to outlier removal. These methods did not produce much difference in the test accuracy for all the data-sets. We believe that since these methods take a vote from multiple candidates to classify the data, the overall effect of the outliers does not impact the nature of the decision boundary significantly. This behavior is completely opposite to what we saw in case of logistic regression.

4. Conclusion and Future Work

In this study, we applied machine learning algorithms on data-sets with non-financial data to predict whether an individual has a good or bad credit for loan approval. Using these type of features has not only allowed us to generate substantially accurate results, but also has increased the usability without using sensitive information of an individual. Hybrid models and standalone supervised machine learning techniques were applied to compare the accuracy for both the models. To apply these models, the raw data from various resources was cleaned up and pre-processed using PCA. PCA allowed us to visualize the data that helped in tuning the parameters of DBSCAN by providing better representation of data without much loss of information. We also found out that algorithms similar to logistic regression are sensitive towards outlier removal. On the contrary, ensemble learning methods do not show much change in accuracy with or without outlier removal.

In future, we plan to implement Bayesian Optimization on supervised models to auto-tune hyper-parameters to find the maximum possible accuracy from each hybrid model. Moreover, we plan to use Linear Discriminant Analysis

(LDA) for dimensionality reduction instead of PCA. Since LDA is used to minimize the within-class variance, we are hoping for improved results as compared to PCA. We also aim to use a variety of loss metrics such as F-score, precision and recall to further evaluate the hybrid models.

Individual Contributions

All of the team members have contributed equally to the project. Initial identification and selection of the project was carried out through group brainstorming sessions where everyone was equally present and contributing. While working on the actual project, the division of work was as follows: The 8 data-sets were divided between the 3 members in a 3-3-2 fashion. Everyone carried out the pre-processing of the data individually to make it suitable for application of the machine learning methods. Further, we had 6 supervised learning methods which were also divided equally and modular codes were written for all of them so that they can be easily run by everyone on the project. The one with the 2 data-sets created the DBSCAN code for the first stage of the hybrid model. Finally during compilation, everyone ran all of the codes, in the following order: we carried out PCA on the data-sets to reduce the dimensionality, then we tuned the DBSCAN parameters to ensure effective outlier removal for each data-set and finally we carried out the supervised learning methods and collected all of the relevant results.

Code

All of the code used in the project was uploaded to a GitHub repository and is available here: https://github.com/rishabhpahuja/Credit-Score-

References

[1] A. G. Armaki, M. F. Fallah, M. Alborzi, and A. Mohammadzadeh. A hybrid meta-learner technique for credit scoring of banks' customers. *Engineering, Technology & Applied Science Research*, 7(5):2073–2082, 2017.

[2] K. Bijak and L. C. Thomas. Does segmentation always improve model performance in credit scoring? *Expert Systems with Applications*, 39(3):2433–2442, 2012.

[3] K. P. Brevoort, P. Grimm, and M. Kambara. Credit invisibles. *Bureau of Consumer Financial Protection Data Point Series*, (15-1), 2015.

[4] I. Brown and C. Mues. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3):3446–3453, 2012.

[5] F.-L. Chen and F.-C. Li. Combination of feature selection approaches with svm in credit scoring. *Expert systems with applications*, 37(7):4902–4909, 2010.

[6] H. Chen and Y. Xiang. The study of credit scoring model based on group lasso. *Procedia computer science*, 122:677–684, 2017.

[7] W. Chen and L. Shi. Credit scoring with f-score based on support vector machine. In *Proceedings 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC)*, pages 1512–1516. IEEE, 2013.

[8] X. Dastile, T. Celik, and M. Potsane. Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, 91:106263, 2020.

[9] Z. Li, Y. Tian, K. Li, F. Zhou, and W. Yang. Reject inference in credit scoring using semi-supervised support vector machines. *Expert Systems with Applications*, 74:105–114, 2017.

[10] C.-F. Tsai and M.-L. Chen. Credit rating by hybrid machine learning techniques. *Applied soft computing*, 10(2):374–380, 2010.