Credit Card Fraud Detection By Using Data Mining

1. Introduction

1.1 Background

Billions of dollars are lost annually due to credit fraud. The U.S is been reported the No.1 in cases with 38.6% of all reported card fraud losses in 2018 [1]. The estimate fraud losses reached $16.9 billion last year according to The Wall Street Journal, an increase dollar volume of fraudulent transactions leaped 35% year-over-year in April[2]. In fraud activity data, the transaction data may have a mix of numerical and categorical attributes, some of the categorical data may contain hundreds of categories. Data mining refers to extracting or analyzing the patterns from data, is been widely used to make intelligence from those complex data.

This study aims to conduct comparative analysis of identification of fraudulent activity on credit card by using logistic regression, random forest classifier and K nearest neighbor classifier. And explore the evaluate metrics changes by utilizing correlation and SelectKbest feature selection methods.

1.2 Dataset and Data Preprocessing

The dataset used in this study comes from Kaggle Credit Card Fraud Detection. This dataset contains transactions made by credit card in September 2013 by European cardholders. It contains only numerical input variables which are the result of PCA transformation. Due to confidentiality issue, features are represent as V1 to V28, except ‘Time’ and ‘Amount’. Feature ‘Class’ is target variable where 1 indicates fraud, 0 as otherwise.

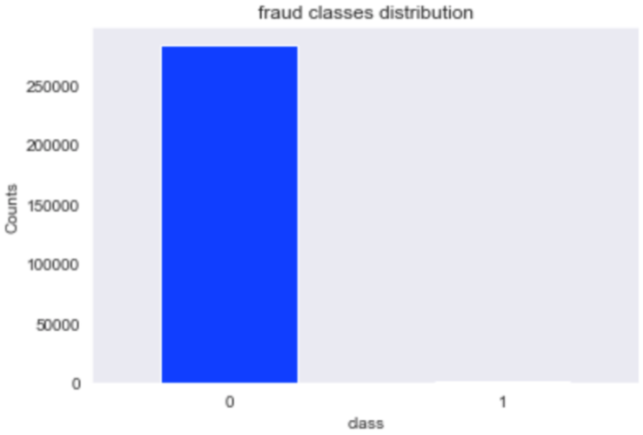


Figure 1 Distribution of Genene(0) and Fraud(1) Transactions

This dataset has 28,4807 data in total. It is highly unbalanced, 99.83% of the data is valid transaction, while 0.17% is fraud. The dataset is first been split as training and testing with ratio of 7:3. In order to train a model to predict valid and fraud transaction equally, I use resampling technique to deal with imbalanced dataset. I choose over-sampling by using RandomOverSampler from imblearn package. After this, I get fraud: valid=50:50 in the training dataset.

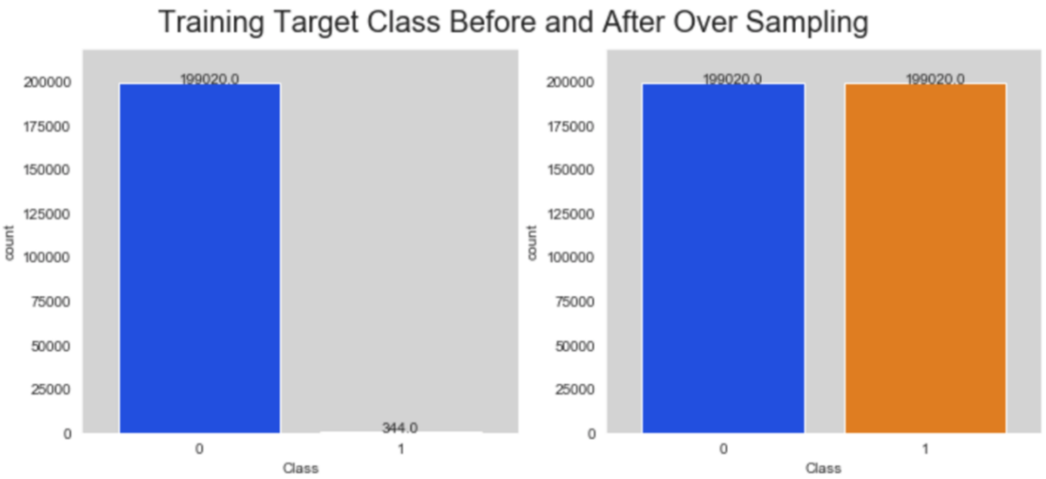


Figure 2 Distribution of Training Dataset After Over-Sampling

2. Methodology

In this study, I use three different classification data mining techniques to detect fraudulent transaction for credit card, as well as two feature selection techniques: correlation and KBest. After feature selection, results will be compared with new features and original features among those three models.

2.1 Logistic regression

Logistic regression is a type probabilistic statistical classification that uses both the logistic regression function and the sigmoid function[3]. Instead of fitting a straight line or hyperplane, the model use sigmoid function to gives a value between 0 and 1. It is been used when the target class is categorical. Logistic regression is one of the widely used classification algorithm in machine learning. In fraud detection, the logistic function calculates the probability, then the probability is used to classify data into two classes by a setting threshold value.

2.2 Random forest classifier

Random forest is supervised machine learning technique used to solve both regression and classification problem. It is an ensemble of classification(regression) method which develops a set of models and aggregates their predictions in target class label for a data point[4]. Random forest is a computationally efficient method since each decision tree is built independently of the others. By using ensemble method, this algorithm is robust to overfitting and noise in the data.

In fraud detection, the random forest algorithm first extract the test features of incoming data and use the rules of each randomly created decision tree to predict the result and store the predicted result, then it calculates the votes for each prediction, finally it evaluates the high voted predicted target from different decision trees as the final result.

2.3 K-nearest neighbor classifier

K-Nearest neighbor(KNN) algorithm is a supervised machine learning algorithm that can be used to solve classification and regression problem. It assumes that similar things exist in close proximity[5]. This algorithm stores all data and classifies in coming data based on a similarity measure. The distance in KNN between two data instances can be calculate by Euclidean distance. The new data point is classified based on the majority of the classes of its neighbors. The classes are define based on a distance metric. The distance from test point to its nearest k neighbor points are calculated and it is classified based on the points which is closest.

It is been widely used in statistical estimation and patten recognition. The KNN rules achieves consistently high performance without a priori assumption about the distribution of training data. For credit card fraud detection, KNN calculates the transaction's nearest point, if this point is near the fraudulent transaction, KNN identifies this transaction as a fraud. The value of k is used to break the tire. The larger k values, the more it can help to reduce the effect of noisy data.

3. Feature selection

3.1 Correlation values methods

The most common way to select feature in a model is to use correlation type statistical measures between input and output variables as the basis. Features with high correlation are more linearly dependent and have almost the same effect on the dependent variable. Based on this, we can drop one of feature if it is highly correlated with the other one. The p-value can be used to decide whether to keep a feature or not, since the removal of different feature from the dataset will have different on the p-value for the dataset.

In this study, I first draw a correlation heatmap to filter out larger coefficient. The correlation map shows the correlation between various features. A correlation close to 1means a strong positive relationship between the target variables. A -1value means strong negative relationship. If the value is equal to 0, which mean no relationship between target variable. Unfortunately most coefficient are less than 0.2. So I use sm.Logit to check all coefficients. I use a 99% confidence level to choose important features, which means I choose a p-value less than 0.01. Among 29 features originally in the dataset, I filter out 9, which are **V4, V8, V10, V13, V14, V20, V21, V22** and **V27**.

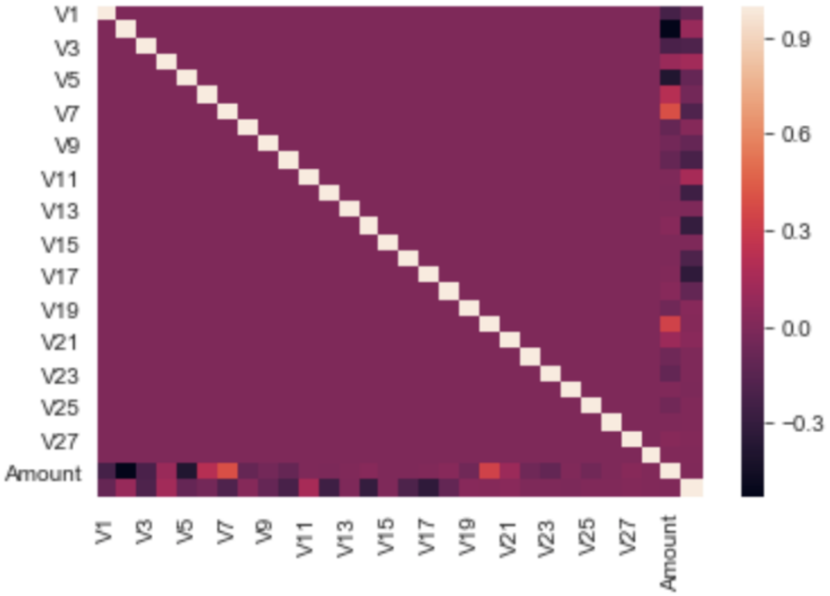
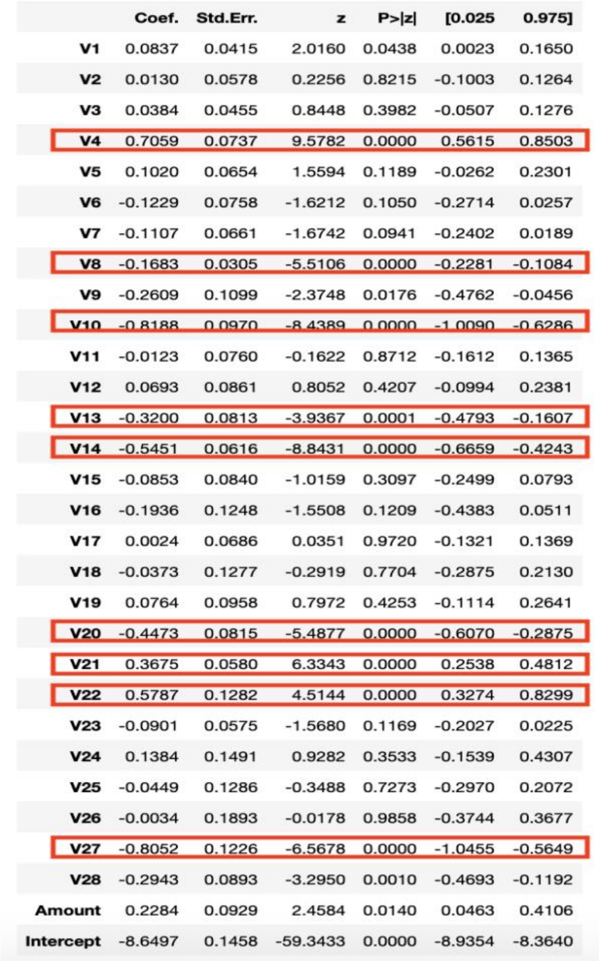


Figure 3 Correlation Heatmap



	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
V1	0.0837	0.0415	2.0160	0.0438	0.0023	0.1650
V2	0.0130	0.0578	0.2256	0.8215	-0.1003	0.1264
V3	0.0384	0.0455	0.8448	0.3982	-0.0507	0.1276
V4	0.7059	0.0737	9.5782	0.0000	0.5615	0.8503
V5	0.1020	0.0654	1.5594	0.1189	-0.0262	0.2301
V6	-0.1229	0.0758	-1.6212	0.1050	-0.2714	0.0257
V7	-0.1107	0.0661	-1.6742	0.0941	-0.2402	0.0189
V8	-0.1683	0.0305	-5.5106	0.0000	-0.2281	-0.1084
V9	-0.2609	0.1099	-2.3748	0.0176	-0.4762	-0.0456
V10	-0.8188	0.0970	-8.4389	0.0000	-1.0090	-0.6286
V11	-0.0123	0.0760	-0.1622	0.8712	-0.1612	0.1365
V12	0.0693	0.0861	0.8052	0.4207	-0.0994	0.2381
V13	-0.3200	0.0813	-3.9367	0.0001	-0.4793	-0.1607
V14	-0.5451	0.0616	-8.8431	0.0000	-0.6659	-0.4243
V15	-0.0853	0.0840	-1.0159	0.3097	-0.2499	0.0793
V16	-0.1936	0.1248	-1.5508	0.1209	-0.4383	0.0511
V17	0.0024	0.0686	0.0351	0.9720	-0.1321	0.1369
V18	-0.0373	0.1277	-0.2919	0.7704	-0.2875	0.2130
V19	0.0764	0.0958	0.7972	0.4253	-0.1114	0.2641
V20	-0.4473	0.0815	-5.4877	0.0000	-0.6070	-0.2875
V21	0.3675	0.0580	6.3343	0.0000	0.2538	0.4812
V22	0.5787	0.1282	4.5144	0.0000	0.3274	0.8299
V23	-0.0901	0.0575	-1.5680	0.1169	-0.2027	0.0225
V24	0.1384	0.1491	0.9282	0.3533	-0.1539	0.4307
V25	-0.0449	0.1286	-0.3488	0.7273	-0.2970	0.2072
V26	-0.0034	0.1893	-0.0178	0.9858	-0.3744	0.3677
V27	-0.8052	0.1226	-6.5678	0.0000	-1.0455	-0.5649
V28	-0.2943	0.0893	-3.2950	0.0010	-0.4693	-0.1192
Amount	0.2284	0.0929	2.4584	0.0140	0.0463	0.4106
Intercept	-8.6497	0.1458	-59.3433	0.0000	-8.9354	-8.3640

Figure 4 Feature Selection By Correlation Method

3.2 Select KBest method

Select KBest is an univariate feature selection method. It uses statistical tests to select features that have the strongest relationship with the output variable. There are many different statistical test scan can be used with the selection. In this study, I use ANOVA f-value method which is appropriate for numerical input and categorical output. I also choose a k value equal to 6 to select the top 6 best features. Those selected feature are: **V4, V10, V11, V12, V14, V16**.

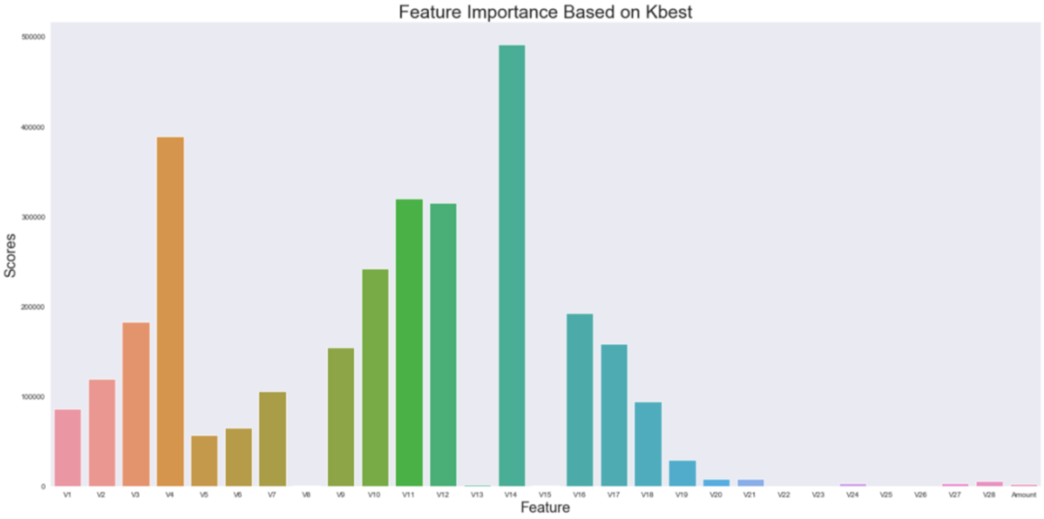


Figure 5 Feature Selection by SelectKBest Method

4. Result

To evaluate machine learning models, I use confusion matrix, AUC-ROC Curve and Precision-Recall Curve.

Confusion matrix: decriable the performance of a classification model with four essential measurements:

- Ture Positive (TP) the cases for which the classifier predict valid transaction and the transaction were actually valid
- Tue Negative (TN) the case for which the classifier predict fraud transaction and the transaction were actually fraud
- False Positive (FP) the case for which the classifier predict fraud transaction and the transaction were actually valid
- False Negative (FN) the case for which the classifier predict valid transaction and the transaction were actually fraud

AUC-ROC Curve: this measurement is used to tell how much the model is capable to distinguish between classes. ROC is a probability curve, AUC represent to degree of separability. The higher AUC, the better model can distinguish between fraud and valid transaction. The ROC curve is plotted with Ture Positive Ratio (TPR) against False Positive Ratio (FPR). Where TPR=TP/(TP+FN), FPR= FP/(TN+FP)

Precision-Recall Curve (PR Curve): Precision is a measure to quantify the number of correct positive prediction, with equation of Precision= TP/(TP+FP). Recall calculates the number of true positive divided by the total positive prediction, with equal of Recall=TP/(TP+FN). Precision-Recall curve is plotted with value of precision and recall with different probability thresholds.

4.1 Comparison between methods

Table1 Accuracy, Recall and Precision Before Feature Selection

Models	Accuracy(%)	Recall(%)	Precision(%)
Logistic Regression	97.584	92.568	6.256
Random Forest Classifier	99.960	83.784	93.537
K-nearest neighbor Classifier	99.906	87.162	67.895

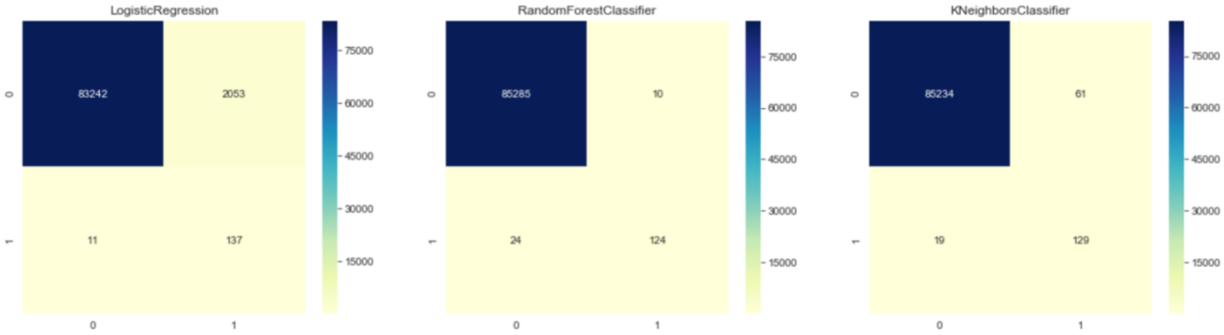


Figure 6 Confusion Matrix Before Feature Selection

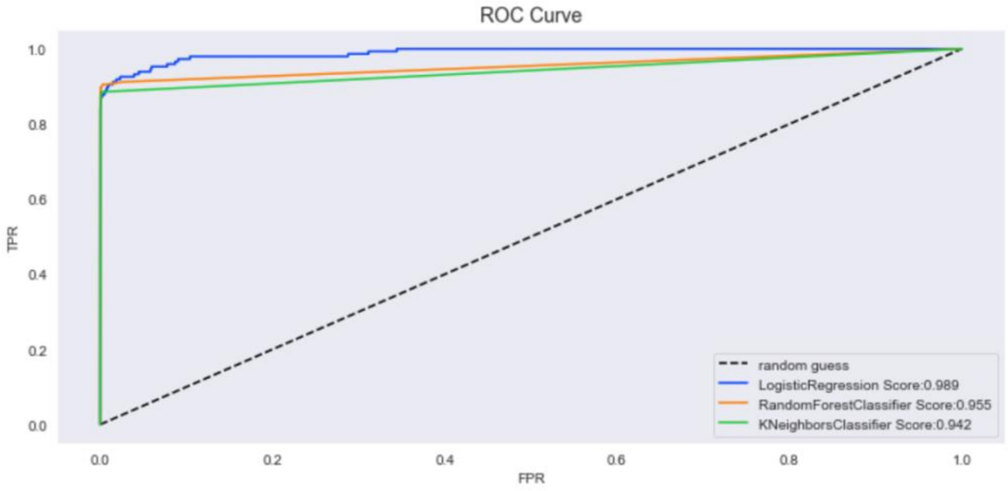


Figure 7 AUC-ROC Curve Before Feature Selection

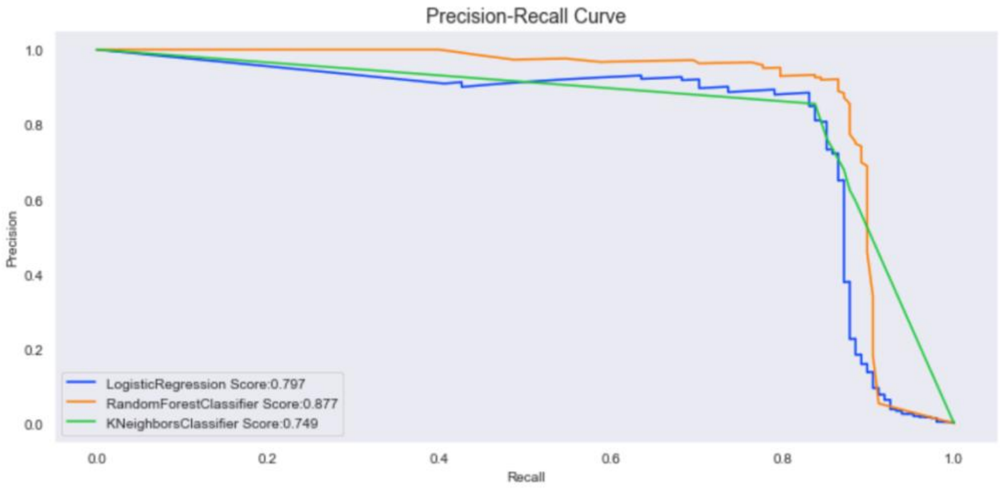


Figure 8 Precision and Recall Curve Before Feature Selection

4.2 Comparison between all features and selected features

Table 2 Accuracy, Recall and Precision After Feature Selection

	After Correlation(%)			After KBest (%)		
Models	**Accuracy**	**Recall**	**Precision**	**Accuracy**	**Recall**	**Precision**
Logistic Regression	97.088	91.892	5.207	97.623	91.216	6.270
Random Forest Classifier	99.959	82.432	91.130	99.965	84.459	94.697
K-nearest neighbor Classifier	99.886	85.135	62.687	99.905	85.811	67.914

After Correlation:

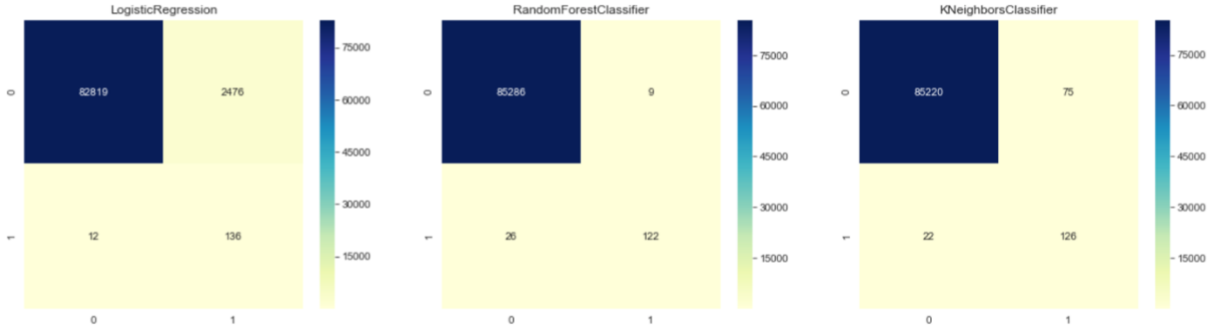


Figure 9 Confusion Matrix After Correlation Feature Selection

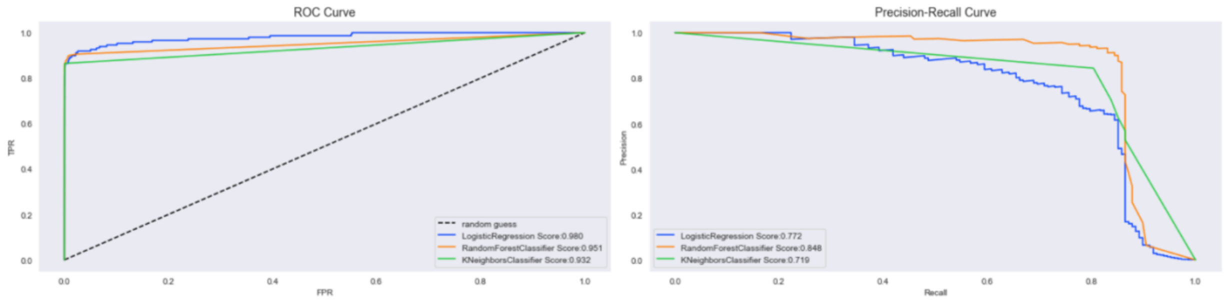


Figure 10 AUC-ROC Curve, Precision-Recall Curve After Correlation Feature Selection

After Select KBest:

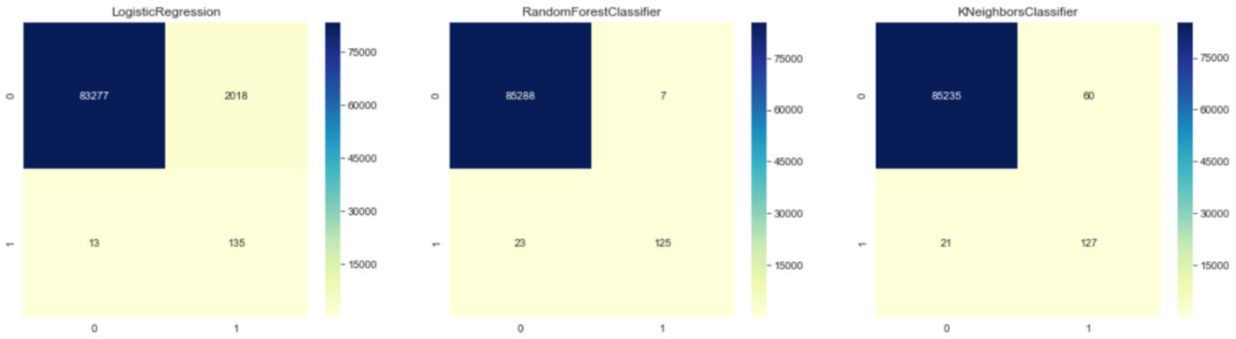


Figure 11 Confusion Matrix After SelectKBest Feature Selection

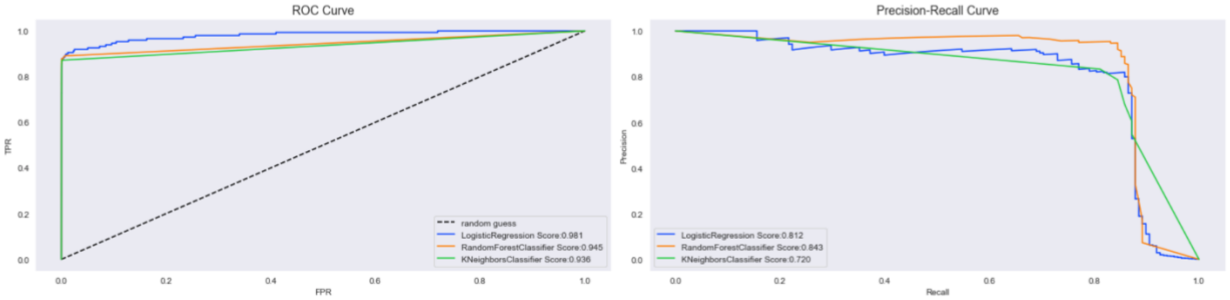


Figure 12 AUC-ROC Curve, Precision-Recall Curve After SelectKBest Feature Selection

5. Discussion

5.1 Comparison between models

Through the confusion matrix in Figure 6 and Table 1: three models all have very high performance on accuracy. This means they are able to predict true valid transaction over 97%. Since our data is highly imbalanced, with over 99% proportion of valid transaction data, the measurement of accuracy is weak in this study. Recall value represent the ability to capture fraud transaction in all fraud cases. Logistic Regression predicts 137 cases out of 148, which is outperform to random forest and K nearest neighbor models. Precision value represents the percentage of true fraud transaction in all predicted fraud transactions. In this study, random forest gets 93.537% compared to the others. In random forest model, there is only 9 true valid transaction been misclassified as fraud transactions. While in logistic regression model, there are

2053 valid transactions been misclassified as fraud transaction. Based on the analysis of confusion matrix, random forest classifier has a better ability to correctly predict fraud transactions.

In AUC-ROC Curve as shown in Figure 7, Logistic regression has higher score of 98.9% then come with Random forest and KNN. In theoretically, the higher score, the better performance of classifier will be. But in our study with highly imbalanced data, the value of FPR is relatively small compare with the number of all valid transactions numbers. So I think AUC-ROC is not enough to measure the goodness of classifier.

PR Curve (Figure 8) provides alternative of model performance by switching from FPR to precision. The higher AUC, the better performance. From the result of PR Curve, Random Forest classifier outperforms with the others. So based on all the comparison between model, I would like to choose random forest classifier.

5.2 Comparison between feature selection and non-feature selection

The reason why we use feature selection is this method enable to reduce complexity of model and help model train faster.

The comparison of performance between all features and correlation feature selection are in Figure 9 and Figure 10. From confusion matrix, AUC-ROC curve and PR curve: all three classifiers have slight worse performance than non-feature selection, especially for logistic regression. This may because in logistic regression, it uses a single line to separate two classes based on thresholds probability. A slightly changes may cause big difference on the performance.

The comparison of performance between all features and Select KBest feature selection method are in Figure 11 and Figure 12. The oval results in three classifier are slightly better than non-feature selection method, especially for Random Forest classifier, which gets highest score in PR curve. From confusion matrix in Figure 11, the Random Forest gets balanced to correctly predict true fraud transaction and true legit. It also decrease the chances to generate FP and FN cases. Kbest method uses statistics tests to select feature which is strongest relationship with target feature. Feature V14 and V4 are the top 2 which are also been selected from correlation method. It seems that Kbest is better to distinguish feature importance than correlation, it results similar or better performance than original dataset.

5.3 Meaning of results

In credit card fraud detection. The purpose is finding a way to detect fraud transaction correctly. Accuracy and recall score are not the only way to measure the performance of model, since misclassifying fraud transaction to legit or misclassifying legit transaction to fraud will have costs in real life. Finding a balance between recall and precision is hard in fraud detection. In this study, Random Forest outperforms the other two classifiers because it maintain recall and precision score with default threshold. PR curve is an alternative measurement when we encounter imbalanced dataset. Our three classifier in ROC curve have less difference, while PR curve is able to differentiate the performance based on precision. In the feature selection, Kbest by using F-test seems to have better performance compared to original dataset and correlation feature selection method.

Reference:
[1] “15 Disturbing Credit Card Fraud Statistics,” *CardRates.com*, 10-Aug-2020. [Online]. Available: https://www.cardrates.com/advice/credit-card-fraud-statistics/. [Accessed: 10-Dec-2020].
[2] “Credit-card fraud surges 35% as coronavirus freezes the economy and wipes out jobs,” *Business Insider*. [Online]. Available: https://markets.businessinsider.com/news/stocks/credit-card-account-fraud-skyrockets-coronavirus-pandemic-recession-economy-layoffs-2020-5-1029246107. [Accessed: 10-Dec-2020].
[3] “Logistic regression,” *Wikipedia*, 07-Dec-2020. [Online]. Available: https://en.wikipedia.org/wiki/Logistic_regression#:~:text=Logistic regression is a statistical,a form of binary regression). [Accessed: 10-Dec-2020].
[4] “Random forest,” *Wikipedia*, 09-Dec-2020. [Online]. Available: https://en.wikipedia.org/wiki/Random_forest. [Accessed: 10-Dec-2020].
[5] “K-nearest neighbors algorithm,” *Wikipedia*, 29-Nov-2020. [Online]. Available: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm. [Accessed: 10-Dec-2020].