Credit Card fraud detection

Ashis Sahu (D17129721)

# Introduction

Credit card fraud detection is an important topic of investigation today, to make sure each transaction made through credit card are fraudulent transactions are not. Many data mining algorithms are used today to detect fraudulent transactions in credit card transactions. The following dataset is from Risk analytics where banks today use **Risk Analytics Helps in Risk management finding the credit score of the clients to qualify customers or clients for providing them credit.** Banks today conduct comprehensive background checks of their [customers](https://www.educba.com/course/customer-analytics-course/) to prevent issues related to frauds and thefts. Data mining techniques and big data technologies with opportunities to address these challenges in real time, these techniques help companies to reduce and manage risk.

# Synopsis of the existing contributions/ Kernel (1 page)

The existing kernel used for this task1 is taken up from kaggle website: <https://www.kaggle.com/joparga3/in-depth-skewed-data-classif-93-recall-acc-now>. This kernel is the highest forked kernel. The kernel has been contributed by the user joparga3, Data science analyst at Hastings Direct, Londres, Inglaterra, Reino Unido. He has contributed this kernel to credit card fraud detection dataset using python nootebook in Kaggle. He has worked on increasing the recall accuracy of predicting the fraud in credit card transaction. Therefore, the kernel has been titled “In depth skewed data classif. (93% recall acc now)”. He has applied cross validation for hyperparameter tuning on classification models, using Logistic Regression. Jopagra3 suggested that, the recall score accurarcy is the most important metric to evaluate the machine learning model, because that will help us try to capture the most fraudulent transactions.

## Logistic regression with different C parameters

## Joparga3 has trained the classification model by taking different C parameters 0.01, 0.1, 1, 10,100 with penalty L1 to evaluate the recall accuracy of the logistic regression model.

## Normalising the amount column using standard scaler function.

The amount column was not in line with the anonymized features (v1-v28), therefore in the kernel; Jopagra3 has normalized the amount column using standard scaler function.

## Random under sampling

## As per Jopagra3, the dataset of credit card fraud has been heavily skewed due imbalanced number of class factors in credit card fraud dataset. The data column with name Class which is the target variable has imbalanced number of data; (class = 0 being non-fraud transaction, class = 1 being fraud transaction).

## Kfold cross validation

Jopagra3 has used cross validation to evaluate the recall score of each iteration with different c parameters to find the best C parameter for logistic regression performing the best on credit card fraud dataset. He trained the model with 5- fold cross validation (k=5).

## Confusion matrix with recall accuracy

As per the kernel, the prediction has been done on a test set, on which the data is not trained with.  The model showed an 93.2% recall accuracy on the generalized unseen data (test set), which is better than when the data is under sampled and trained by the logistic regression model.

## Plotting ROC curve and Precision-Recall curve.

## Jopagra3 has plotted the ROC curve and precision recall curve of logistic regression at different iterations to investigate Precision-Recall curve and area under this curve.

# Future Work:

# In his future update, Jopagra3 assured to do analysis on evaluating the recall scores on predicting the target variable using SVM and decision trees.

# Proposed planning on Machine Learning work

The work carried out by Jopagra3 was mainly only on one data mining algorithm i.e, logistic regression classifier. In our machine learning work we have added three more data mining models such as SVM, decision trees and k-nearest neighbor. We have used different parameters and configuration for KNN model, decision tree, SVM and linear regression to formulate a comparison between all the models.

To increase the accuracy of our models, we removed the "extreme outliers" from features or attributes which have a high correlation with our classes. This had a positive impact on the accuracy of our models. We took the most correlated data attributes and removed the outliers on whichever data falling outside the range of 1st and 3rd quartile range, to increase the strength of prediction of data mining model.

For SVM, decision tress and KNN, to evaluate with the best possible model to predict the fraudulent transactions, we have choose different parameters, such as varying the kernel type and c parameters in SVM and in decision tree varying the criterion and max depth for the tree formation and in k- nearest neighbors we varied the algorithms for KNN to predict the fraudulent transactions most accurately and provide the comparison test among all the models.

The existing work carried out only using random under sampling, whereas we worked with machine learning task with SMOTE technique and random under sampling technique and evaluate the difference between the two model with recall accuracy comparison.

# Methodology

This section should outline the work carried out in developing a solution through the stages of pre-processing and model selection, paying particular attention to document the reason for decisions taken to make progress. (4 pages)

## Dataset description

The dataset has been acquired from Kaggle dataset. The data was collected on big data mining and fraud detection from  a research collaboration of Worldline and the Machine Learning Group ([http://mlg.ulb.ac.be](http://mlg.ulb.ac.be/)) of ULB (University of Brussels) . The data consists of 31 variables with all the features went through a PCA transformation (Dimensionality Reduction technique) V1 to V28 except time and amount attributes. The target variable is class where class = 0 for non-fraudulent transaction and class = 1 for fraudulent transaction. The variables V1 to V28 are PCA transformed under GDPR act., as it includes many sensitive information of clients which can also be misused. Each row in the dataset is information about each credit card detection of customers of the bank which has been transformed for anonymity. The dataset includes 285,000 rows and 31 columns of credit card transaction.

## Unbalanced dataset

The target variable is highly imbalanced with 99.83 % of non- fraudulent transaction and with only 0.17% of fraudulent transaction. Below the histogram showing the unbalanced dataset:

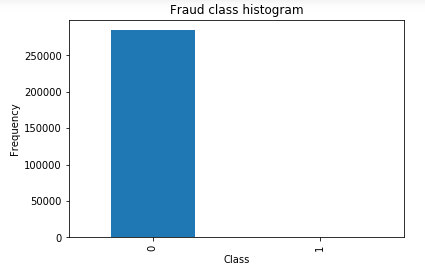


Figure 1

## Sampling

### Collecting more data

### To resolve the imbalanced dataset, it would have been a nice strategy to collect more data but not applicable in this case, as we cannot collect more data.

### Random Under Sampling

To create an under-sample data, we randomly select records from the majority class and import it into the dataset with the same number of records in minority class.In under sample dataset there will be 50/50 ratio of target varaible by removing majority class randomly

### SMOTE

SMOTE (Synthetic Minority Over-Sampling Technique), which is a combination of oversampling and under sampling. The oversampling approach is not undertaken by replicating minority class but constructing new minority class data instance through an algorithm. SMOTE creates new synthetic points to have an equal balance of the classes.

In this experiment, we will use random under-sampling and SMOTE.

## Data Mining models

### SVM

Support vector machines (SVM) is a supervised learning algorithm which has been used in classification and regression problems. Support vectors or also called kernels are the data points that fall near to the decision surface which transforms the data points in to non-linearly separable transformation. It uses optimal hyperplane for linearly separable patterns. It extends patterns to non-linearly separable by transformations of original data to map into new space using the Kernel function. SVM maximizes the separation margin around the separating hyperplane by transforming the data into transformed decision surface or so-called hyperplane. Basically, SVM performs like linear regression though it takes non-linear transformation to better predict the outcome variable. The algorithm used for this approach is used with different C parameters such as 'C': [0.5, 0.7, 0.9, 1] and different kernel function such as 'rbf', 'poly', 'sigmoid', 'linear'; to evaluate the model and find the best performing model by varying hyper parameters.

### Decision tree

Decision tree constructs classification models like a tree structure. It forms a tree by breaking down a dataset into smaller and smaller subsets at the same time building an associated decision tree. The tree formed with leaf and decision nodes. This algorithm for formation of tree structure is called ID3 for building decision trees which uses greedy search and a top-down. We have changed the parameters to perform in the prediction for decision tree with setting and varying its criterion to “gini" and "entropy". We have also varied max\_depth, min\_samples\_leaf parameter of the decision tree to evaluate the best performing models.

### K- nearest neighbor

The k-nearest neighbors (KNN) data minning model is a supervised machine learning algorithm that is used for classification and regression problems. KNN falls under **lazy learning**, which means that there is **no explicit training phase before classification**. In KNN, the data points are separated on basis of clusters formed at feature space. The Euclidian distance between the data points forms the cluster. In KNN, K specifies the number of cluster to be formed for classification task and in unsupervised learning its different. The KNN model used in this algorithm performs 2 clusters to perform the classification task as there is binary varaible for the target varaible for classification task for KNN. KNN is trained with different parameters such as 'auto', 'ball\_tree', 'kd\_tree', 'brute' algorithm to check for the best performing algorithm.

### Logistic Regression

Logistic regression is a regression analysis data mining model which is applied of a dataset where when the dependent variable is dichotomous variable. Logistic regression is a classifier algorithm which is defined by the logarithmic function

http://www.statisticssolutions.com/wp-content/uploads/2010/01/log23.jpg

In this experiment, we took logistic regression with varying c parameters such as 'C': [0.001, 0.01, 0.1, 1, 10, 100, 1000] and penalty of L1 and L2 to formulate the best performing model.

## Cross Validation

## Cross-validation is a method used for resampling which is used to evaluate machine learning models. In cross validation the dataset is divided into k-folds; k-1 of the subset becomes training dataset and 1 of them becomes validation set. Then this process is run iteratively through out of the learning or training process to evaluation of the model. This procedure consists of a parameter called k; which implies to the number of sub datasets the dataset must be split into. In this experiment, we have made k=5; which means the dataset will be splitted into 5 subsets for training of data.

## Performance Evaluation

### Recall

### This is a clear example where we cannot use accuracy score to evaluate our classification algorithm. By checking the accuracy, we will still be having a high accuracy, but we will not e able to predict the CLASS = 1 correctly i.e., the main fraudulent transactions. Therefore, in order to evaluate the data mining models we check the recall accuracy.

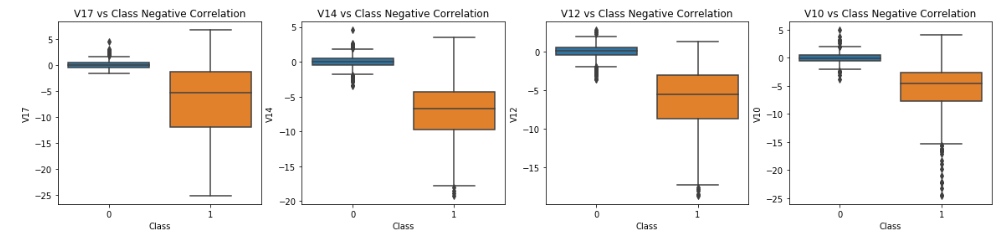
Recall: True Positives/(True Positives + False Negatives)

### ROC curves

### Performance evaluation of data mining model for such a case where thee is imbalanced dataset can be done through ROC curves, to find the datamining model with higher area has better classification strength. ROC takes sensitivity/specificity ratio into consideration which is required in this case.

## Normalizing correlated variables

Normalizing correlated variables are essential in increasing the efficiency of data mining model. Since the correlated variables consisting in the dataset having outliers can deviate the performance of the data mining model. The variables in the dataset V17, V14, V12 and V10 are negatively correlated and V2, V4, V11, and V19 are positively correlated. We checked the correlated variables using boxplot to have a better understanding of the normality distribution as these variables brings the main variance in the target variable classify in fraudulent and non-fraudulent transactions. To make sure the correlated variables attaining the normality distribution to improve the efficiency of the data mining models, we checked the distribution of feature with target variables with the boxplot to check the outliers. The observations from boxplot (figure 2):



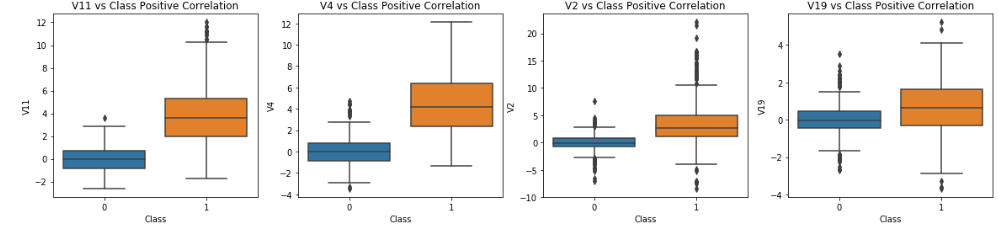


Figure 2

After checking with the boxplots, we found that variables V14, V2, V10 are having outliers which can deviate the accuracy of the models. Therefore, to increase the efficiency of the model we implemented outlier reduction with taking 1st and 3rd quartile ranges. The lower and upper range are changed to the 25th percentile and 75th percentile of the corresponding variables. Whenever the data points go beyond the threshold 75th and 25th percentile, in that case the instance will be deleted which improves the accuracy.

## Scaling Amount and time attributes:

The amount and time attributes needed to be scaled as they are highly skewed. Following are ate distribution of time and amount variables in the dataset (Figure 3):

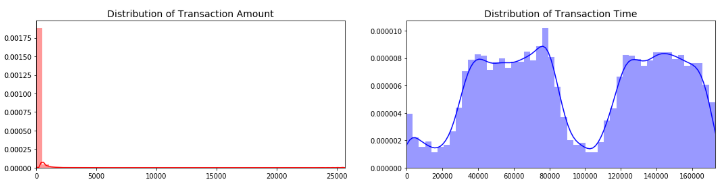


Figure 3

As we see from the graph, the distribution chart of Transaction Amount and time, they are highly skewed, which can push down the efficiency or accuracy of the model. Therefore, to normalize we needed to scale the time and amount variables.

Feature Scaling is used to standardize the range or normalize the independent variables or features of data. Feature selection can be done through various ways, such as;

1. *Standardization*: It can scale features where the distribution is centered around 0, with a standard deviation of 1. It can only work with the features with gaussian distribution.
2. *Normalization*: It shrinks the range between 0 and 1 if there are if there are negative values -1 to 1. The problem with normalization is that gets influenced heavily by extreme values or outliers.
3. *Robust Scaler*: This is similar to normalization, rather it uses the interquartile range for scaling the feature variable, therefore it becomes robust to outliers. Moreover, it only focuses on the parts where
4. the bulk data is present and doesn't take the median into account.

For the feature scaling, we did use Robust scaler as it fits for the feature data. Scaled amount and scaled time are the transformed variables of amount and time with scaled values.

Evaluation and analysis of our observation:

## Under-sampling training accuracy score:

We trained the model with random undersampled dataset with 50-50 majority and minority class and this is the first evaluation of models, therefore we trained the model with the default parameters.The following observation has been observed:

|  |  |
| --- | --- |
|  | Accuracy |
| Logistic Regression | 93% |
| Knn | 93% |
| Support vector classifies | 92% |
| decisison tree | 88% |

From the accuracy score, we can conclude that Logistic regression and KNN performed the best when trained with default parameters.

## Cross validation accuracy score

For the next iteration we performed k-fold cross validation to check the accuracy score and we changed the hyperparameters of the models as described above to find the best performing algorithm. The following observation was observed when we trained the model with under sampled dataset:

|  |  |  |
| --- | --- | --- |
|  | Mean Accuracy Score | Max. accuracy score |
| Logistic Regression | 93.4% | 94.74% |
| Knn | 92.47% | 94.08% |
| Support vector classifies | 93% | 94.74% |
| decisison tree | 90.89% | 92.76 |

Using cross validation and changing the hyper parameters has increased the accuracy score as that can be observed from comparison.

After changing the hyperparameters and doing cross validation we see that Logistic classifier , SVC, Decision tree has increased the accuracy score. Whereas KNN performed shows a degradation in accuracy with cross- validation, the main reason could be KNN is a clustering algorithm and their it forms random clusters and using cross validation it does’nt perform that well. Though we can conclude from the observation here, logistic regression performs the best overall according to accuracy score.

## ROC curve:

Following is the ROC curve graph from all trained model with under sampled dataset:

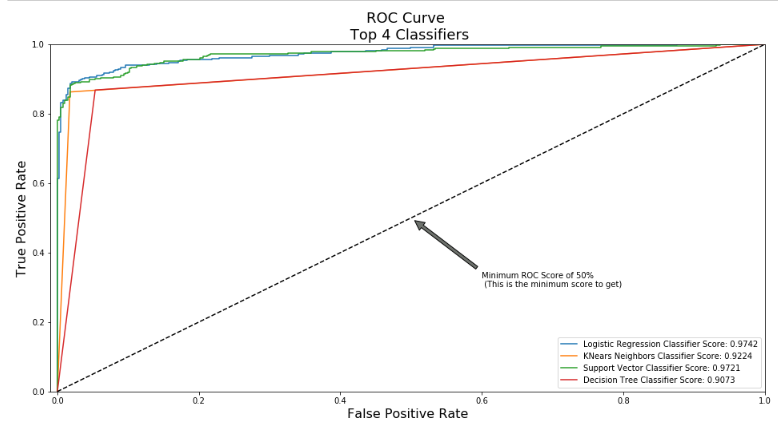


Figure 4

From the ROC curve graph , it is clear that logistic regression performs the best with ROC- area under curve score = 0.9741. Below is the table of ROC – Area under curve scores of all the model; from here we conclude that Logistic regression performs the best on undeersampled data set.

|  |  |
| --- | --- |
|  | ROC AUC score |
| Logistic Regression | 0.9741 |
| Knn | 0.9224 |
| Support vector classifies | 0.9720 |
| decisison tree | 0.9073 |

## Comparison between under sampled and SMOTE

As per the kernel by Jopagra3, he performs test to check the best performing logistic regression model with different c-parameters to check which model performs the best. For the novelty, we will evaluate the model trained with Under sampled dataset as well as SMOTE dataset with c-parameter, to find which sampled dataset must be used to training for best performance of the model. We trained the model with both the dataset using cross validation and observed the following result:

|  |  |  |
| --- | --- | --- |
|  | Accuracy | Recall |
| Undersampled | 0.957 | 0.91 |
| SMOTE | 0.977 | 0.87 |

As from the observation we observed that, the accuracy in SMOTE sampling dataset the accuracy is higher than under sampled dataset but the Recall is lower than under sampled dataset. Therefore, in order to train model to get the best prediction for fraudulent transactions, we must use under sampled dataset which has almost 4% higher recall value. Therefore, for this case study the best model to train to get optimum results in credit card fraud detection is to use logistic regression with minimum C- parameter using cross validation on random under sampled dataset.