Credit Card Fraud Detection

Akshaya A (CB.EN.U4CSE17401)
B Manish Reddy (CB.EN.U4CSE17409)
Gayathri E (CB.EN.U4CSE17420)
Thejaswagiri P(CB.EN.U4CSE17463)

Dept. of Computer Science & Engineering
Amrita School of Engineering
Amrita Vishwa Vidyapeetham

Nov 1, 2020

1 Introduction

Credit card fraud events take place very frequently in recent years and then result in huge financial losses. The number of online transactions has been growing in large quantities and online credit card transactions hold a huge share of these transactions. Therefore, financial institutions offer credit card fraud detection applications with much value and demand. Fraudulent transactions can occur in various ways and can be put into different categories. The major key area that we address in our project is real-time credit card fraud detection. For this, we take the use of predictive analytic done by the implemented machine learning model to decide if a particular transaction is genuine or fraudulent. Unfortunately, due to confidentiality issues, data does not have original features and extra background information, the data used in our model has the features which are the principal components obtained with principal component analysis. The model is built with the records that turned out to be a fraud in the past credit card transactions. The built model is then used to identify whether a new transaction made by the credit card is fraudulent or not. Thus, achieving classification of fraudulent and non-fraudulent.

2 Objective

The aim is to classifying fraudulent transactions present in the history of transactions by applying various machine learning techniques such as feature selection, re-sampling for highly imbalanced data set, logistic regression, random forest and thereby building a model to prevent further fraudulent transactions happening.

3 Problem Statement

To understand and and build a model with the records that turned out to be fraud in the past credit card transactions. The built model is then used to identify whether a new transaction made by the credit card is fraudulent or not.

4 Architecture

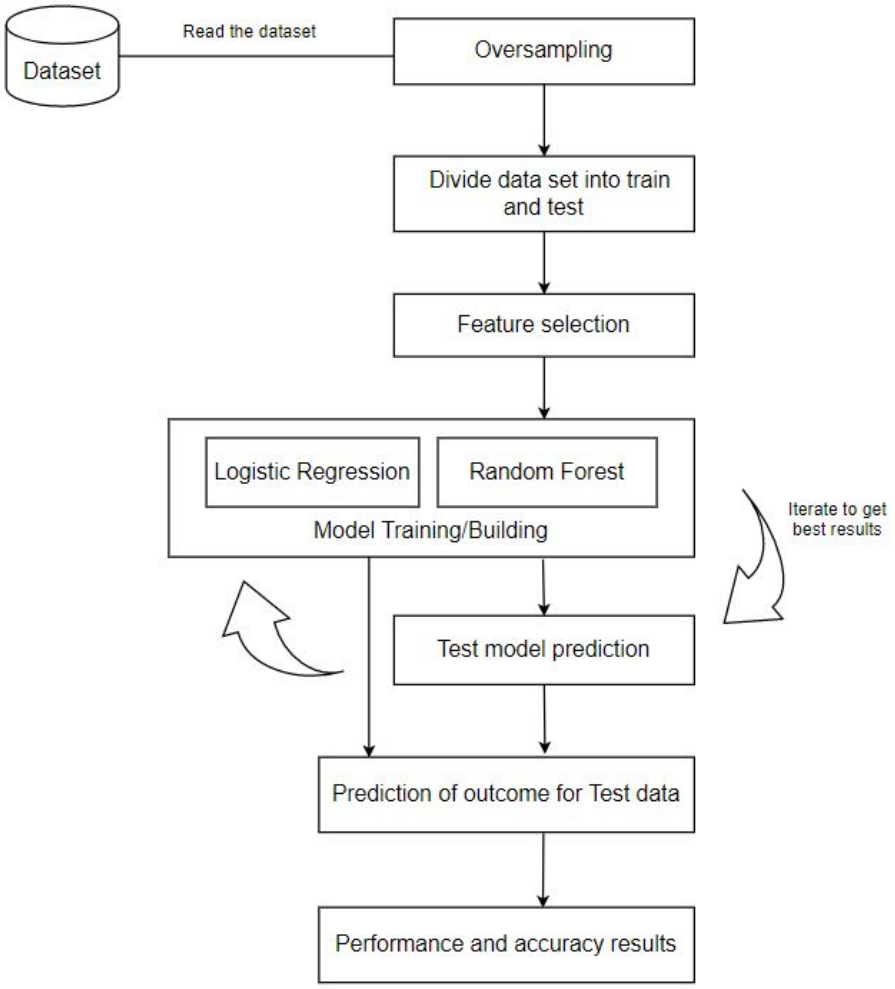

Figure 1: System Architecture

5 Related Work

J. O. Awoyemi, A. O. Adetunmbi and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis" - A hybrid technique of under-sampling and oversampling is carried out on the skewed credit card fraud data.Based on accuracy, sensitivity, specificity, precision, balanced classification rate and matthews correlation coefficient the performance is evaluated. The results shows of optimal accuracy for naive bayes, k-nearest neighbor and logistic regression classifiers are 97.92 , 97.69 and 54.86 percentile respectively.

W. Yu and N. Wang, "Research on Credit Card Fraud Detection Model Based on Distance Sum" - This research paper proposes a fraud detection model based on distance sum according to the infrequency and unreality of credit card transaction data, applying outlier mining into credit card fraud detection. Experiments show that this model is feasible and accurate in detecting credit card fraud.

6 Novelty of the Work

Due to large amount of data generated in electronic transactions, we have to find the best set of features to identify the frauds. Fraud detection is a specific application of anomaly detection, characterized by a large imbalance between the classes, which can be a detrimental factor for feature selection techniques. Here we evaluate the behaviour and impact of feature selection techniques to detect frauds in the web transactions. To measure the effectiveness we use odds ratio, confidence intervals and p values. Given the class imbalance ratio, recommended measure of accuracy is to use Area Under the Precision-Recall Curve (AUPRC).

Probability value for a given statistical model that if the null hypothesis is true then a set of statistical observations more commonly known as the statistical summary is greater than or equal in magnitude to the observed results is the p-value. We can remove different features and measure the p-value in each case. These measured p-values can be used to decide whether to keep a feature or not. We have used P-value based approach for feature selection which is a backward elimination method. Imbalance data refers to classification problem where the number of observations per class is not equally distributed. If the data set has severely imbalanced classes it can lead high overall accuracy without much effort but without generating any good sights. The overall accuracy might be high, but for the minority class will have very low recall. Dealing with imbalanced data sets entails strategies such as improving classification algorithms or balancing classes in the training data before providing the data as input to the machine learning algorithm. The later technique is preferred as it has wider application.

The main objective of balancing classes is to either increasing the frequency of the minority class or decreasing the frequency of the majority class. This

is done in order to obtain approximately the same number of instances for both the classes. So this can be obtained using sampling techniques.Two types under it being: Random under sampling - aims to balance class distribution by randomly eliminating majority class examples, which is done until the majority and minority class instances are balanced out; Random over sampling - Over-Sampling increases the number of instances in the minority class by randomly replicating them in order to present a higher representation of the minority class in the sample. We use Synthetic Minority Oversampling (SMOTE) which is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them. SMOTE synthesises new minority instances between existing minority instances.

7 Data Collection and Preparation

Building the data set involves decisions which can greatly affect the quality of the data mining project and leads to poor results. Moreover, one must transform all categorical variables into numerical, in order to use few machine learning algorithms such as Support Vector Machines. This section describes the reasoning behind collecting data and preparation of the data for this project. Handling the data was done with Python version 3.8.3, available through the Python Software Foundation at http://www.python.org. Two particular modules for Python, Pandas data structures and Scikit-Learn, were especially useful for running the algorithm.

The data set used in our project is taken from Kaggle. It does not have original features and extra background information, the features in the data set are V1, V2, V3,...,V28 which are the principal components obtained with principal component analysis. "Time" containing the seconds elapsed between each transaction and "Amount" being the transaction amount are the only two features which have not been transformed with PCA. The data collected is then split in two halves: a training and a testing set. The training set is then used for the algorithm to learn how to classify the transactions into fraudulent and non-fraudulent. Later, the testing set is used to validate the effectiveness of the algorithm, removing the over-fitting effect, i.e., the increase in performance that the algorithm has on the instances it has based its learning, as it describes the random error including the underlying relationship.

7.1 Feature Selection

Imbalanced classes are a common problem in classification problems in data mining. Skewed class proportions with unequal ratio of observations in each class makes the data set imbalanced. Class imbalance can be found in many several areas including filtering spams and diagnosis in medicine. This data set presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The data set is highly unbalanced, the positive class

(frauds) account for 0.172% of all transactions. The feature selection we have used is based on backward elimination method. The significance of using this method is to keep only those features which are relevant to the data set i.e. those features doing considerable amount of change to the dependent variable are only considered.

A significance level is chosen and the model is fit with all features. P-values of different features are checked and if p-value is higher than significance level, the feature is removed. The steps are repeated with the reduced features until the features having p-values less than significance level alone remains. From the data frame, the dependent variable and a list of column names are taken and regression is run repeatedly eliminating feature with the highest P-value above alpha one at a time and returns the regression summary with all p-values below alpha and drops the other features with p-value greater than alpha.

7.2 Data visualization

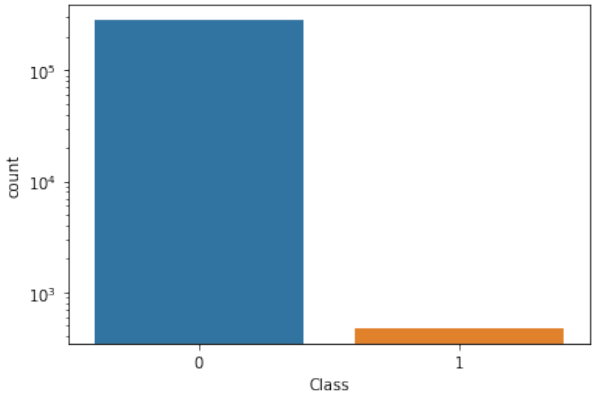

Figure 2: Imbalanced data - frauds(minority class)

Data visualization involves presenting data in graphical or pictorial form which makes the information easy to understand. It helps to explain facts and determine courses of action. It will benefit any field of study that requires innovative ways of presenting large, complex information. Real-world data sets related to telecommunications fraud, computer network intrusion, and credit card fraud were evaluated. The results were displayed with visual appeal to data analysts as well as non-experts, as high-dimensional data samples were projected in a simple 2-dimensional space. To observe the frequency of a particular value, histogram is plotted. X axis has the column/feature values; y axis has the frequency for each feature.

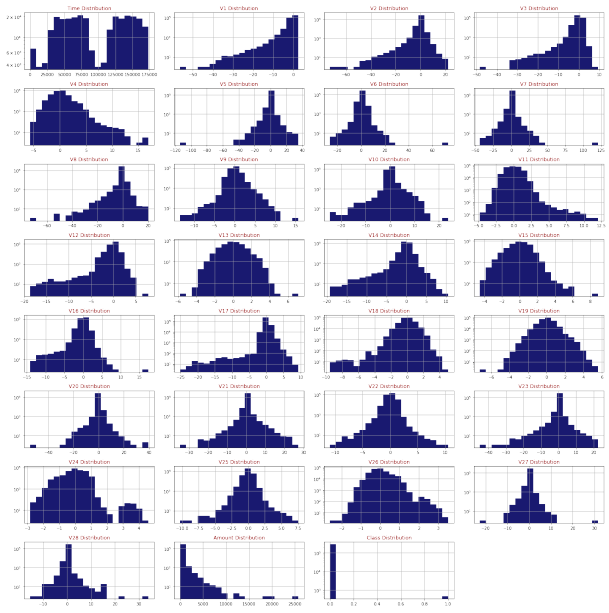

Figure 3: Histogram - distribution across all the columns

7.3 Data Cleaning & Pre-processing

The original data has a lot of dirty data, such as incorrect attribute values, duplicate records, null value, inconsistent values, various abbreviations, violations of referential integrity and so on. In order to make better use of the data for data mining and decision support, it should be changed into high-quality data. Therefore, a data pre-process procedure, which is known as data cleaning is to clean up the dirty data before using the data. Noticing how imbalanced our original data set is, most of the transactions are non-fraud. Happening to use this data frame as the base for our predictive models and analysis it might result to a lot of errors and algorithms will probably over-fit, as it "assumes" that most transactions are not fraud. But, we don't want our model to assume, we want our model to detect patterns that give signs of fraud!

Redundant entries are removed and the time in seconds is converted to hours and minutes form for convenience and to better understand the pattern. We will then scale the columns comprising of Time and Amount. Time and amount should be scaled as the other columns. On the other hand, we need to also create a sub sample of the data frame in order to have an equal amount of Fraud and Non-Fraud cases, helping our algorithms better understand patterns that determines whether a transaction is a fraud or not.

8 Data Modeling and Inference

The given data set is a result of result of a Principle Component Analysis transformation. It is a dimension-reduction method that is often used to reduce the dimensions of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.Then data quality is ensured by checking for NULL/missing values and duplicate values eradication. After this logistic regression and random forest classifier is done on the data with over-sampling(up-sampling) using SMOTE as a part of Exploratory Data Analysis. Accuracy, F2 score and Recall score are recorded for both the approaches. Histogram visualization has been done on all the columns to understand the frequency distribution of the data. Correlation heat-map matrix is visualized for the independent variables.The heat-map clearly shows which all variable are multicollinear in nature, and which variable have high collinearity with the target variable. The heat-map is referred back-and-forth while building the linear model so as to validate different correlated values along with p-value. It helps to determine the statistical significance of our results/research.

The results in the model SMOTE (Over Sampling) + Random Forest has the best score, on performing a f2 score calculation which resulted 0.8252 approximately. There is a considerable difference by the second best model that is 0.7831 that uses just Random Forests with some Hyper Parameters. Also it is clearly better logistic regression model, even though 0.64 of accuracy is not too bad, we prefer Random Forest Model which happens to have a higher value.

8.1 Improvements/Enhancement of models

A data set which is highly imbalanced needs to be handled, SMOTE being a oversampling technique has been performed making the data set balanced. Before oversampling the normal data distribution shows Counter(0: 283253, 1: 473) and after SMOTE (oversampling) data distribution shows Counter(0: 283253, 1: 283253). The model is further improved by performing hyper parameter tuning. The accuracy of the model with SMOTE and Random Forest classifier and accuracy of the other model with tuned hyper parameters are taken for comparison to choose a better model. Also, considered Logistic Regression with parameter tuning of 'C' and scoring as 'f1_micro' to get proper f1 values for imbalanced data set classes and Logistic Regression with parameter tuning of 'C' and 'class_weight'. The 'class_weight' parameter helps in a imbalanced data set by penalising the wrong classification of smaller classes more, 'roc_auc' being a scoring metric. The models were compared and considered the one with higher accuracy as the best model.

8.2 Performance Evaluation and result discussion

We use confusion matrix to find how many fraud cases can be predicted accurately. From the matrix we are able to classify 56678 cases as valid where as

68 are records as fraudulent. From the matrix we abstract True Positive, True Negative, False Positive and False Negative values which helps in calculating the evaluation of statistics such as accuracy, specificity and sensitivity. In logistic regression the odds ratio represents the constant effect of a predictor X, on the likelihood that one outcome will occur. The confidence interval is accurate if the sample size is large enough that the distribution of the sample odds ratios follow a normal distribution. A common way to visualize the trade-offs of different thresholds is by using an ROC(receiver operating characteristic) curve, a plot of the true positive rate (true positives/total positives) versus the false positive rate (false positives/total negatives) for all possible choices of thresholds. A model with good classification accuracy should have significantly more true positives than false positives at all thresholds.

8.3 Model Tweaking, Regularization, HyperParameter Tuning

The function has trained itself to get the correct target values for all the noise induced data points and thus has failed to predict the correct pattern. This function may give zero error for training set but will give huge errors in predicting the correct target values for test data set. To avoid this condition regularization is used. Regularization is a technique used for tuning the function by adding an additional penalty term in the error function. The additional term controls the excessively fluctuating function such that the coefficients don't take extreme values.

Hyperparameters are points of choice or configuration that allow a machine learning model to be customized for a specific task or data set. Machine learning models also have parameters, which are the internal coefficients set by training or optimizing the model on a training data set. Parameters are different from hyper parameters. Parameters are learned automatically; hyper parameters are set manually to help guide the learning process. Typically a hyperparameter has a known effect on a model in the general sense, but it is not clear how to best set a hyperparameter for a given data set. Further, many machine learning models have a range of hyperparameters and they may interact in nonlinear ways. As such, it is often required to search for a set of hyperparameters that result in the best performance of a model on a data set called hyperparameter optimization, hyperparameter tuning, or hyperparameter search.

An optimization procedure involves defining a search space. This can be thought of geometrically as an n-dimensional volume, where each hyperparameter represents a different dimension and the scale of the dimension are the values that the hyperparameter may take on, such as real-valued, integer-valued, or categorical. A range of different optimization algorithms may be used, although two of the simplest and most common methods are Random Search - which defines a search space as a bounded domain of hyperparameter values and randomly sample points in that domain; Grid Search - defines a search space as a grid of hyperparameter values and evaluate every position in the grid. We used Logistic Regression with parameter tuning of 'C' and 'class_weight'.The

"class_weight" parameter helps in a imbalanced data set by penalising the wrong classification of smaller classes more. The "class_weight" parameter is tuned using RandomizedSearchCV and the way it takes the value for class_weight is the weights associated with classes in the form class_label: weight.

9 Conclusion

To classify fraudulent transactions present in the history of transactions by applying various machine learning techniques in spite of data without background information as the data is mostly private and the data being a Imbalanced Data i.e most of the transactions (99.8%) are not fraudulent which makes it really hard for detecting the fraudulent ones is very challenging. Feature selection enables the model to arrive at a much better accuracy by considering only those features which contribute the most to the prediction variable or output in which we are interested in. Logistic Regression model, has 0.64 of accuracy which is not too bad, we prefer Random Forest Model which happens to have a higher value. On performing a f2 score calculation for SMOTE (Over Sampling) + Random Forest model, resulted 0.8252 approximately. There is a considerable difference by the second best model that is 0.7831 that uses just Random Forest with some Hyper Parameters. Clearly, the results in the model SMOTE (Over Sampling) + Random Forest has the best score and turns out to be a better model. Overall model could be improved with more data.

References

[1] J. O. Awoyemi, A. O. Adetunmbi and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis". In *ICCNI-2007*.

[2] W. Yu and N. Wang, "Research on Credit Card Fraud Detection Model Based on Distance Sum". In *JCAI-2009*.

[3] Tao Y H,Rosa Yeh C C. Simple database marketing tools in customer analysis and retention. International Journal of Information Management,2003,23:291-301

[4] F.S. Maes, K. Tuyls, B. Vanschoenwinkel, and B. Manderick.Credit Card Fraud Detection Using Bayesian and Neural Networks[C]. Proceedings ofNeuro Fuzzy, Havana, Cuba, 2002.

[5] A. Shen, R. Tong, Y. Deng, "Application of classification models on credit card fraud detection", Service Systems and Service Management 2007 International Conference, pp. 1-4, 2007.