Credit Card Fraud Detection using Machine Learning

Name: Harshdeep Singh

Registration No.: 11902010

Section: KM016

Abstract

It's critical for credit card firms to be able to spot fraudulent credit card transactions so that customers aren't charged for things they didn't buy. Such issues may be solved with Data Science, which, together with Machine Learning, cannot be underestimated. With Credit Card Fraud Detection, this project aims to demonstrate the modelling of a data set using machine learning. Modelling prior credit card transactions with data from those that turned out to be fraudulent is part of the Credit Card Fraud Detection Problem. The model is then used to determine whether or not a new transaction is fraudulent. Our goal is to detect 100% of fraudulent transactions while reducing the number of incorrect fraud classifications. n. Detection of credit card fraud and scams Displaying prior credit card swaps with details of those that turned out to be misrepresentation is a problem.

That model is then used to determine whether or not another transaction is deceptive. Our objective is to identify 100 percent of fraudulent transactions and reduce the number of false fraud/scam classifications. Scam/fraud using credit cards A typical example of grouping is identification. here We had focused on analysing and pre-preparing data throughout this round. Set collections are nothing more than the transmission of a large number of inconsistency detection or identification algorithms.

I Introduction

In credit card transactions, 'fraud' refers to the unlawful and unwelcome use of an account by someone who is not the account's owner. To stop this misuse, necessary preventative steps should be adopted, and the behavior of such fraudulent acts can be analyzed to decrease it and defend against future occurrences. In other words, credit card fraud occurs when a person uses

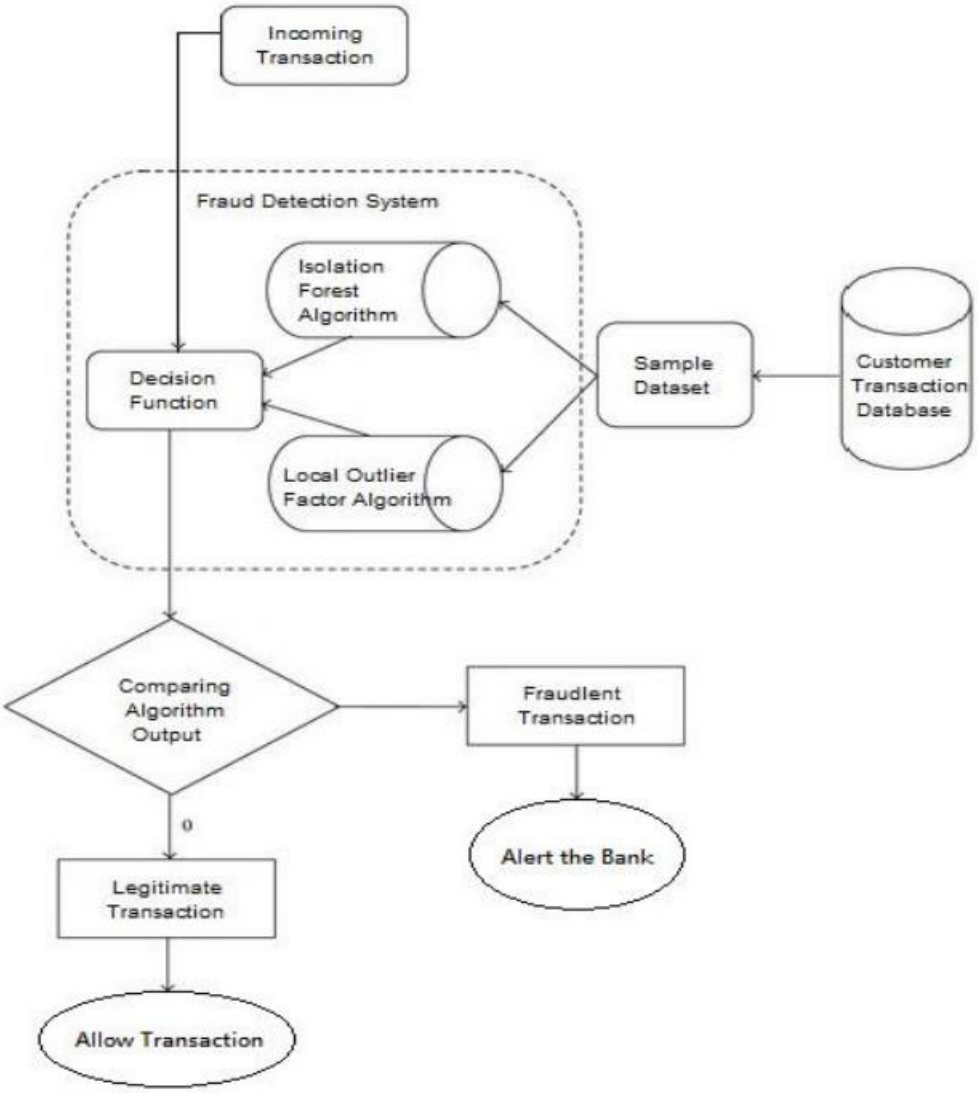
another person's credit card for personal gain while the owner and card issuing authorities are ignorant of the transaction.

Fraud detection is tracking the behaviors of large groups of people in order to predict, detect, or avert unacceptable behaviors such as fraud, intrusion, or defaulting. This is a highly important subject that requires the attention of fields like machine learning and data science, where the answer may be automated.

This issue is particularly difficult to solve from the standpoint of education since it is characterized by many elements such as class imbalance. The number of legitimate transactions considerably outnumbers the number of fraudulent transactions. Furthermore, transaction patterns frequently modify their statistical features over time. These aren't the only difficulties that come with putting a system in place.

However, there is no real-world fraud detection mechanism. In real life, for instance, the huge influx of payment requests is rapidly processed. Automatic tools are used to identify which transactions should be scanned.

To analyze all of the data, machine learning techniques are used. Authorized transactions should be reported, whereas questionable transactions should be reported. These professionals evaluate reports and make contact with the appropriate authorities. The investigators give input to the automated system, which is utilized to train and upgrade the algorithm over time to enhance fraud detection effectiveness.



Fraud detection methods are continuously developed to defend criminals in adapting to their fraudulent strategies. These frauds are classified as:

- Credit Card Frauds: Online and Offline
- Card Theft
- Account Bankruptcy
- Device Intrusion
- Application Fraud
- Counterfeit Card
- Telecommunication Fraud

II Literature Review

Fraud is defined as an illegal or criminal deceit designed to gain financial or personal gain. It is a purposeful act committed in violation of a law, regulation, or policy with the intent of obtaining unlawful financial advantage.

A large number of literatures on anomaly or fraud detection in this sector have previously been published and are available for public use. Data mining applications, automated fraud detection, and adversarial detection are among the strategies used in this sector, according to a comprehensive survey undertaken by Clifton Phua and his colleagues. Suman, Research Scholar, GJUS&T at Hisar HCE, proposed strategies for credit card fraud detection such as Supervised and Unsupervised Learning in another study.

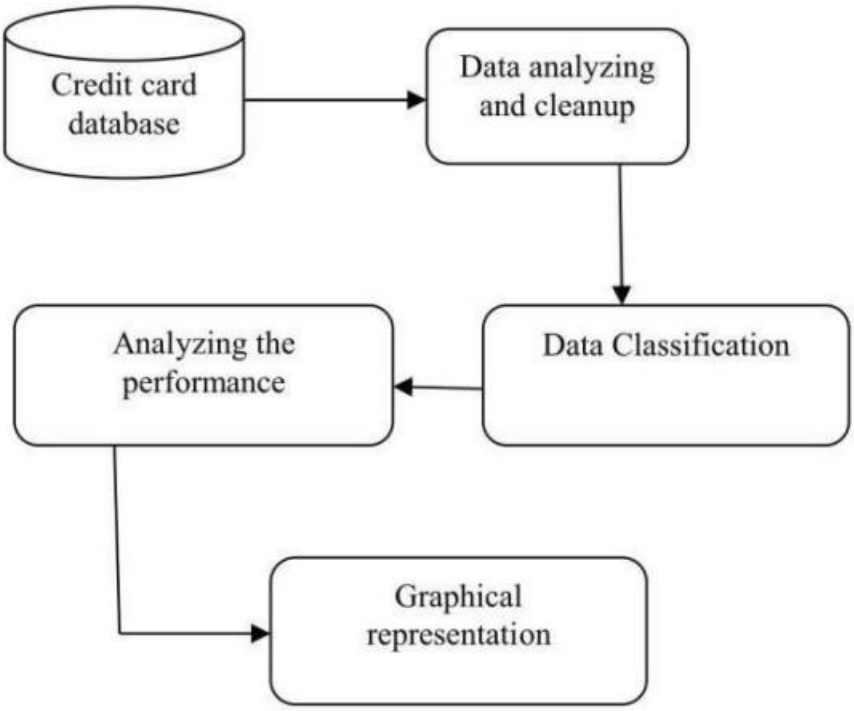
Despite their unexpected success in some areas, these approaches and algorithms failed to provide a long-term and consistent answer to fraud detection. Wen-Fang YU and Na Wang presented a similar study topic in which they employed Outlier mining, Outlier detection mining, and Distance sum algorithms to correctly forecast fraudulent transactions in an emulation experiment using credit card transaction data from a single commercial bank. Outlier mining is a type of data mining that is commonly utilized in the financial and internet industries. It deals with identifying items that are disconnected from the main system, such as fraudulent transactions.

They took aspects of consumer behavior and estimated the distance between the observed value of that attribute and its predetermined value based on the value of those attributes.

Unconventional approaches, such as hybrid data mining/complex network classification algorithm, have shown effective on medium-sized online transactions, based on network reconstruction algorithm that allows building representations of the divergence of one instance from a reference group. There have also been attempts to go forward from an entirely other perspective. In the event of a fraudulent transaction, efforts have been made to enhance the alert feedback interaction. In the event of a fraudulent transaction, the authorized system would be notified, and a report would be provided to the customer.

III Methodology

In this approach all the 5 classification models of machine learning are applied on the training dataset to predict the results. 70% of the data from the dataset is used to train the system and results are predicted by using 30% of the test data. Features are selected from the dataset to improve the results. after executing all the fifteen models, top three models with best performance are chosen. Ensemble of the top best three models is done and the results are evaluated to enhance the results of the resultant prediction. Cross validation is then taken into consideration. K-fold validation is a category of cross validation which 5 measures the robustness of the model.



I. Dataset

The dataset contains transactions made by credit cards in September 2013 by European cardholders.
This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Given the class imbalance ratio, we recommend measuring the accuracy using the Area Under the Precision-Recall Curve (AUPRC). Confusion matrix accuracy is not meaningful for unbalanced classification.

II. Data Pre-processing

Since this dataset is about transactions that were done by the European cardholders and the original dataset contained the all-personal information of all the users like Name, age, residential address, annual income, profession etc. All the columns containing personal and sensitive information all being dropped already by the data provider to maintain the privacy of their customers.

Rest of the columns are not provided by the data provider to tell us which feature it is. Only the last 2 columns are kept unchanged Amount and the target column containing either 0 which means false, the transaction is not a fraud, it is a genuine one or 1 which means, it is a fraud and not made by the owner of the card.

III Feature Selection

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection techniques are used for several reasons:

- Simplification of models to make them easier to interpret by researchers/users,
- Shorter training times
- To avoid the curse of dimensionality
- Improve data's compatibility with a learning model class
- Encode inherent symmetries present in the input space.

IV Feature Extraction

Feature Extraction aims to reduce the number of features in a dataset by creating new features from the existing ones (and then discarding the original features). These new reduced set of features should then be able to summarize most of the information contained in the original set of features.

Another commonly used technique to reduce the number of features in a dataset is Feature Selection. The difference between Feature Selection and Feature Extraction is that feature selection aims instead to rank the importance of the existing features in the dataset and discard less important ones (no new features are created).

V Machine Learning Models Used

5.1 Logistic Regression

Logistic regression, despite its name, is a classification model rather than regression model. Logistic regression is a simple and more efficient method for binary and linear classification problems. It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes. It is an extensively employed algorithm for classification in industry. The logistic regression model, like the Adaline and perceptron, is a statistical method for binary classification that can be generalized to multiclass classification. Scikit-learn has a highly optimized version of logistic regression implementation, which supports multiclass classification task.

5.2 Decision Tree

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. Decision tree can be computationally expensive to train. The process of growing a decision tree is computationally expensive. At each node, each candidate splitting field must be sorted before its best split can be found. In some algorithms, combinations of fields are used and a search must be made for optimal combining weights. Pruning algorithms can also be expensive since many candidates' sub-trees must be formed and compared.

5.3 Support Vector Machine

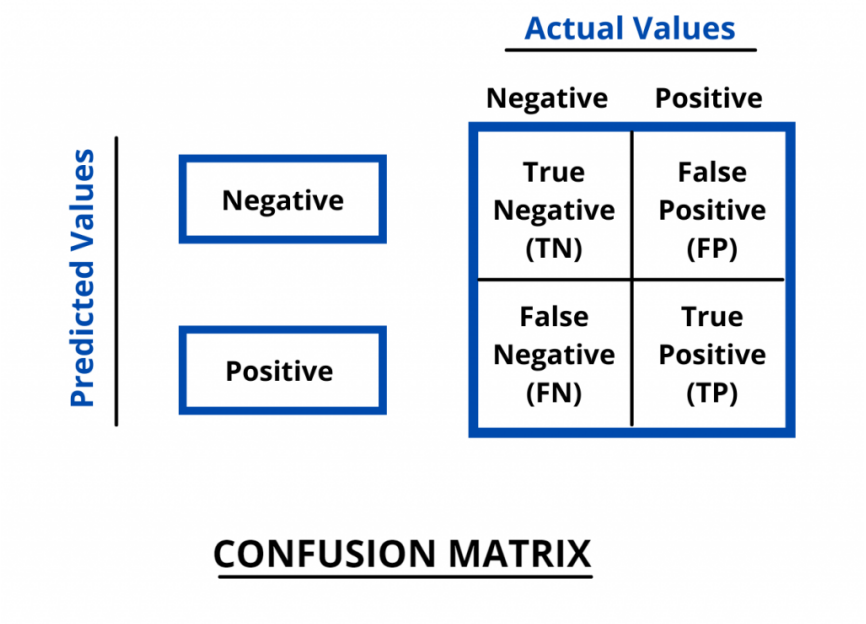
The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane. The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector.

VI Model Evaluation

6.1 Confusion Matrix

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well

our classification model is performing and what kinds of errors it is making.



6.2 Accuracy Score

Classification Accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

6.3 F1 Score

The F1 Score is the 2*((precision*recall)/(precision+recall)). It is also called the F Score or the F Measure. Put another way, the F1 score conveys the balance between the precision and the recall. To evaluate model performance comprehensively, we should examine both precision and recall. The F1 score serves as a helpful metric that considers both of them.

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

The equation for F1 score | Image by author

VII Result Analysis

7.1 Performance Comparison

Machine Learning Model	Accuracy Score	F1 Score
Logistic Regression	99.91 %	0.83
Decision Tree Classifier	99.89 %	0.84
Support Vector Machine	99.90 %	0.83

7.2 Ensemble Model

Ensemble learning involves combining multiple predictions derived by different techniques in order to create a stronger overall prediction. The goal of any machine learning problem is to find a single model that will best predict our wanted outcome. Rather than making one model and hoping this model is the best/most accurate predictor we can make, ensemble methods take a myriad of models into account, and average those models to produce one final model.

We have made an ensemble model of three machine learning models. These are Logistic Regression, Decision Tree and Support Vector Machine.

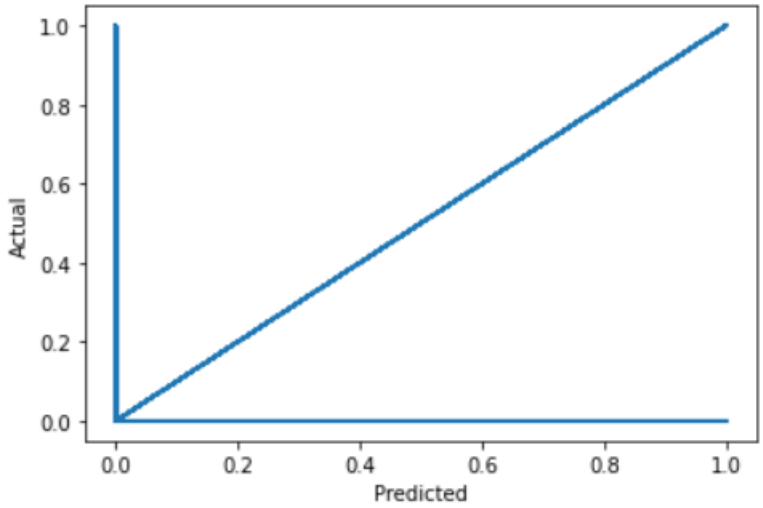
Voting Parameter	**Accuracy Score**	**F1 Score**

Voting soft	99.92%	0.85
Voting hard	99.91%	0.85

7.3 Cross Validation

In machine learning, we couldn't fit the model on the training data and can't say that the model will work accurately for the real data. For this, we must assure that our model got the correct patterns from the data, and it is not getting up too much noise. For this purpose, we use the cross-validation technique. Cross validation technique is used to validate the predictive models and analyse statistical results. It estimates how accurately any predictive model will perform. In this technique the original sample is partitioned into a training set to train the model, and a test set which is used for system evaluation. In this methodology validation approach of cross validation is used, in which data get shuffled on random basis. The goal of the cross validation is to define a test dataset which is used for testing the system and it also reduces the problem of overfitting. The dataset is shuffled ten times and the results are cross validated.

7.4 Scatter Plot



Ensembled Machine Learning Plot

VIII Conclusion

This approach tells us that we can predict credit card fraud detection using various machine learning models. And after comparing the accuracy of all the machine learning models, we have found that Logistic Regression, Decision Tree Classifier and Support Vector Machine is giving the highest accuracy out of all the other machine learning algorithms. It gave 99.91% , 99.89 % and 99.90 respectively.

These machine learning models are also performed very well when applied through an ensembled model. Moreover, the ensembled model gave the highest accuracy out of these 3 models. It performed very well combined as compared to working individually. The ensembled model gave 99.92 % accuracy which is the highest among all the three models.

Github link: https://github.com/coderhersh/Credit-Card-Fraud-Detection

References

[1] J William Langston. Parkinsons disease: current and future challenges. Neurotoxicology, 23(4):443–450, 2002.

[2] Lonneke ML de Lau and Monique MB Breteler. Epidemiology of parkinson's disease. The Lancet Neurology, 5(6):525–535, 2006.

[3]Dejan Varmedja; Mirjana Karanovic; Srdjan Sladojevic; Marko Arsenovic; Andras Anderla ; 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH) 20-22 March 2019

[4]Pranali Shenvi; Neel Samant; Shubham Kumar; Vaishali Kulkarni ; 2019 IEEE 5th International Conference for Convergence in Technology (I2CT) 29-31 March 2019

[5] N. S. Halvaiee and M. K. Akbari published“A novel model for credit card fraud detection using Artificial Immune Systems,” Applied Soft Computing, vol. 24, pp. 40–49, in 2014