Detection of fraud using Classification (Credit Card Fraud Dataset)

Amena Afreen
Computer Science
Hood College
Frederick, Maryland, USA
aa24@hood.edu

ABSTRACT

Anomaly detection is the key for the future data mining. It is a data mining technique which allows us to find patterns in data that do not conform to expected behavior. In a world where technology is advancing at abnormal rate, credit cards are playing a vital role in today's economy. It is used everywhere such as in household, business and global activities. There are quite many benefits of using credit cards. Credit cards should be used legally but the illegal usage of it might lead to fraudulent activities and cause several damages. Fraudulent activities take place in every field of financial domain, credit card transactions are also sometimes fraudulent. Due to the widespread usage of credit card, there is an increase in credit card frauds. The credit card fraud detection dataset is highly skewed and imbalanced, so we first balance the dataset by equally distributing the frauds using some random sampling. The project aims detecting the fraudulent transactions by training the dataset and applying different classification models to predict the fraudulent behavior of the transactions.

KEYWORDS

Fraud Detection, K Nearest Neighbors, Support Vector Machines

1 Introduction

Datamining is about discovering insights from data that are statistically reliable, unknown previously and actionable (Elkan, 2001). The data must be available, and it should be clean to work on. The word "fraud" refers to the abuse of a profit organization's system without necessarily leading to direct legal significances. In an environment where there are competitions emerging, there are chances that fraud can become a critical problem if it is widespread and if the procedures undertaken for prevention are not fail-safe. One of the important parts of fraud control is "Fraud Detection", it automates and helps reduce the manual parts of screening/checking process.

Fraud detection has become one of the most reputable applications of data mining. The occurrence of anomalous behavior is not seen frequently, if they occur frequently then the consequences are negative. The advancement of e-commerce led to convenience in using credit cards and has become necessary for financial life. Fraud is increasing drastically with globalization and modern technology which results in major loss to the business.

When a credit card is used illegally without the knowledge of owner of the card then it is referred to as credit card fraud. There are two groups of frauds basically application fraud and behavioral fraud. Application fraud is a type of fraud where fraudsters apply new cards from bank using false identity. On the other hand, behavioral fraud is of four types such as stolen/lost, counterfeit card, mail theft and card holder not present.

The patterns of text in a dataset that do not conform to a normal behavior are termed as anomalies. These are the behaviors that are previously not known. Anomaly Detection is a technique used to find these patterns in the dataset. The data has two normal regions, N1 and N2, since most observations lie in these two regions. Points that are away from these regions or areas, such as the points o1 and o2, and points in region o3, are anomalies.

When a single data point can be considered as anomalous with respect to the rest of data, then the point is called point anomaly. When a data point is anomalous in a specific context, then it is termed as contextual anomaly. It comprises of contextual attributes and behavioral attributes. A collection of related data instance is anomalous with respect to the entire dataset, it is termed as collective anomaly. The unique data instances in a collective anomaly may not be anomalies.

2 Dataset Information

The credit card fraud dataset is from a European credit card company and is obtained from Kaggle. The transactions are done in 2013 September. The transaction history of the two days is shown in the dataset. The dataset is highly imbalanced, which means that the number of observations that belong to one class are higher than the observations belonging to other class. The total number of transactions in the dataset are 284,807 out of which 492 transactions are termed as fraud which is accounted as 0.172% of whole transactions. Due to confidentiality, the input variables are converted to numerical values by PCA transformation, the features V1,V2,., V28 are the principal components obtained from PCA and thus are not the original features. Feature Time represents the difference in the seconds between the particular transaction and the first transaction. 'Amount' is the amount of money for which the transaction has taken place. Feature 'Class' is the label that shows whether a transaction is fraud or not. The data labels indicate that the data point is anomalous or normal. When we plot time against the transaction, no abnormal behavior is shown. The columns "Time" and "Amount" are not scaled by PCA, so they should be scaled. We first determine that how imbalanced the dataset is, we tend to resample the data by using random sampling distributions so that there are equal number of fraud and valid transactions.

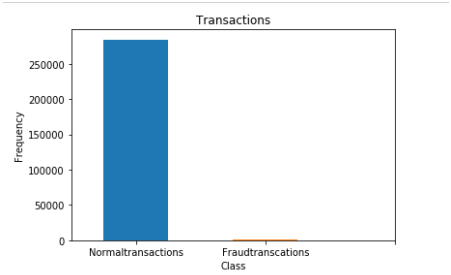

Figure 1: Plot of valid transactions and fraud transactions

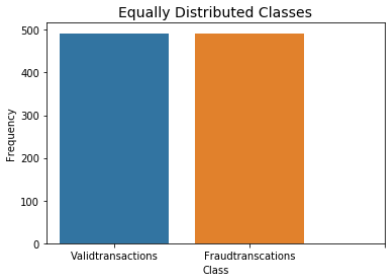

Figure 2: Plot of valid transactions and fraud transactions after balancing the dataset

Correlation Matrix

Correlation Matrices helps us understand our data. If there are any features that influence severely in whether a transaction is fraud or not. For it, we use the sampled data to see the features having positive and negative correlations with regard to fraud detection
First for the imbalanced class the correlation matrix for the dataset shows that there is no correlation between any of the V1 to V28 components however there is positive and negative correlation between 'Class' and other V components. Also, Class has no correlation with 'Time' and 'Amount'. The correlation matrix after sampling the data distribution there are negative correlations seen in the attributes V17, V14, V12, V10 and The attributes V2, V4, V11, V19 are positively correlated that means that the higher the values for these there is possibility that the transaction is termed as fraud. The attributes that will be used for fraud detection are Time, Amount, V17, V14, V12, V10, V2, V4, V11, V19 and class label.

3 Machine Learning Based Fraud Detection

3.1 Density Based Anomaly Detection

In Density-based anomaly detection density of the neighborhood of each data instance is estimated. An instance is said to be anomalous when its density is low, and it is said to be normal when it lies in dense neighborhood. For any instance, the distance to its K^{th} nearest neighbor is equivalent to the radius of hypersphere. So, basically the distance to its K^{th} nearest neighbor is estimated to be the inverse of the density of the instance in the dataset.

3.2 K-Nearest Neighbor

Nearest-Neighbor based fraud detection there needs to be some distance or any similar measure between two data points. K-Nearest Neighbor is a supervised learning algorithm. It requires valid as well as fraudulent samples of the data fro training. It was first introduced by Aha, Kibler and Albert (1991). The result of any given query is classified based on the K-Nearest Neighbor category. The performance of the K-Nearest Neighbor is best because of certain factors such as the distance metric used in locating the nearest neighbors, a distance rule in deriving the classification from K-Nearest Neighbor and the number of neighbors used in classifying the new sample. Credit card fraud detection requires distance between the two data instances. The incoming transaction is classified by calculating the nearest point to new coming transactions. If the nearest neighbor is fraudulent then the transaction is termed a fraud. The k values is a small value or is any odd value.

3.3 Logistic Regression

Logistic Regression is used for different classification problems such as fraud detection, diabetes prediction, school grade prediction. It is most commonly used Machine learning algorithms. The output of logistic regression is continuous. In this discrete set of classes are assigned observations depending on the data. It transforms its output using a logistic function to return a value which is most probable, and which can be mapped to two or more discrete classes. The basic idea is that the input space is divided into two regions, one for each class by a linear boundary. The linear boundary is a straight line.
Logistic regression is termed as statistical method for predicting binary classes. The target variable is usually dichotomous in nature, which means that there are only two possible chances as in fraud detection. It is either '1' for fraud and '0' for non-fraudulent.

3.4 Support Vector Machines

An SVM is a Machine learning model that fits for binary classification of patterns that are linearly separable. In this one can draw some possible hyperplanes between the two classes. The idea that lies behind SVM is to derive a unique separating hyperplane that maximizes the separating margin between the two classes. The separating margin in the jargon of linear algebra is the feature vector that lies on the boundary and is termed as Support Vectors. Classifiers that exploit this property are termed as support vector machines. SVM is a binary classifier. It is a supervised learning model that analyzes and recognizes patterns for classification and regression problems. The approach behind SVM for the instances of two classes was to find an optimal hyper-plane. The gap between some marginal instances is called as the support vectors, and it is the place where the hyper-plane was located. This was all done linearly. For linearly inseparable data, there were kernel functions introduced. In a high dimensional space. A kernel function represents the dot product of two data points. In classification problems, for a given training sets, marked with some label, SVM algorithm finds the hyper-plane that can assign new incoming instances into either of the two classes. SVMs haves been applied to many fraud detection problems and credit card fraud detection has been the very known. In credit card fraud detection, a model was developed by Gosh and Reilly and it was better than neural networks.
The steps taken are:

1. First set the training data for creation of model
2. Set the parameter's to be used by SVM
3. SVM Trainer
4. SVM Predicto

4 Performance Measures

4.1 Confusion Matrix

In binary classification model instances are classified into two classes, that is true and a false class. This indeed gives rise to four possible classifications for each data instance namely true positive, true negative, false positive and false negative. The confusion matrix contrasts the observed classifications for a phenomenon with the predicted classifications of the model.

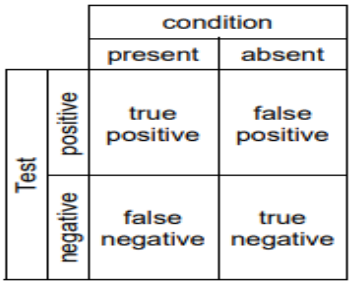

Figure 3: Confusion Matrix

True Positives (TP): These are the correctly predicted positive values saying that the actual class value is yes, and the predicted class value is also yes. For detecting fraud if the actual class says that the transaction is fraud and the predicted class also says the same thing.

True Negatives (TN): These are the correctly predicted negative values saying that the actual class value is no, and the predicted class value is also no. For detecting fraud if the actual class says that the transaction is not fraud and the predicted class also says the same thing.

False Positives (FP): These are the incorrectly predicted positive values saying that the actual class value is no, and the predicted class value is also yes. For detecting fraud if the actual class says that the transaction is not fraud and the predicted class says that the transaction is fraud.

False Negatives (FN): These are the incorrectly predicted negative values saying that the actual class value is yes, and the predicted class value is also no. For detecting fraud if the actual class says that the transaction is fraud and the predicted class also says that the transaction is not fraud.

Accuracy: It is an important performance measure and is just the ratio of correctly predicted observation to the total observations.

$$\text{Accuracy} = (TP+TN)/(TP+FP+FN+TN) \quad (1)$$

Precision: The ratio of correctly predicted positive observations. This shows that the number of credit card transactions that are labelled as fraud how many are actually fraud. High precision relates to low positive rate.

$$\text{Precision} = TP/(TP+FP) \quad (2)$$

Recall: The ratio of the correctly predicted positive observations. This shows that of the total transactions that are fraud, how many did we label.

$$\text{Recall} = TP/(TP+FN) \quad (3)$$

F1 Score: F1 is the weighted average of precision and recall. It considers both false positives and false negatives into account. It is not easy to understand as accuracy but is most useful than accuracy when the class distribution is imbalanced. If the costs of false positives and the false negatives are different then we should consider precision and recall.

$$\text{F1 Score} = 2*(\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision}) \quad (4)$$

5. Results

5.1 K-Nearest Neighbor

Train/Test model	Accuracy	F1 Score	Precision	Recall
Train	0.95%	0.95%	0.97%	0.93%
Test	0.92%	0.93%	0.98%	0.88%

Figure 4: Accuracy Scores

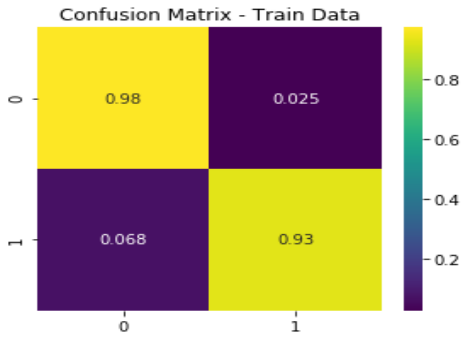

Figure 5: Confusion Matrix using KNN (train data)

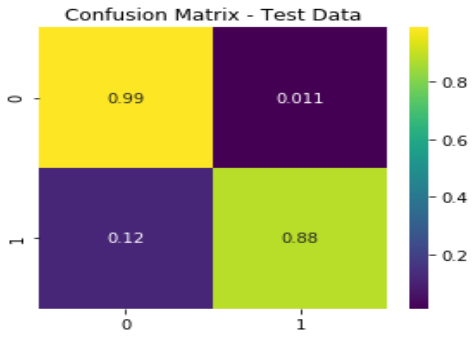

Figure 6: Confusion Matrix using KNN (train data)

5.2 Logistic Regressor

Train/Test model	Accuracy	F1 Score	Precision	Recall
Train	0.94%	0.94%	0.97%	0.91%
Test	0.93%	0.94%	0.98%	0.90%

Figure 7: Accuracy Scores

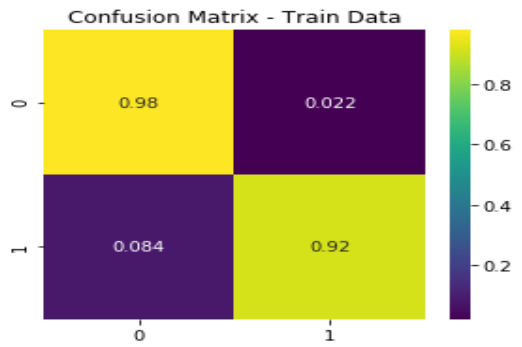

Figure 8: Confusion Matrix using Logistic Regressor (train data)

Figure 9: Confusion Matrix using Logistic Regressor (test data)

5.3 Support Vector Machines

Train/Test model	Accuracy	F1 Score		Precision	Recall
Train	0.95%	0.94%		0.98%	0.91%
Test	0.93%	0.94%		0.98%	0.90%

Figure 10: Accuracy Scores

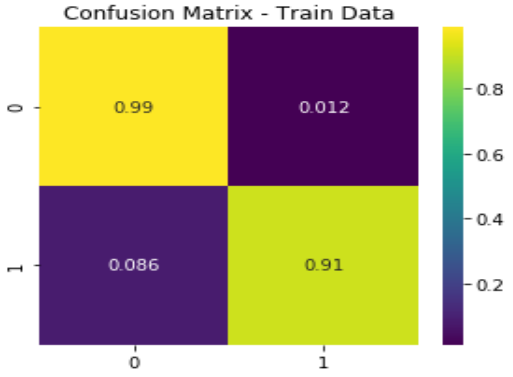

Figure 10: Confusion Matrix using SVM (train data)

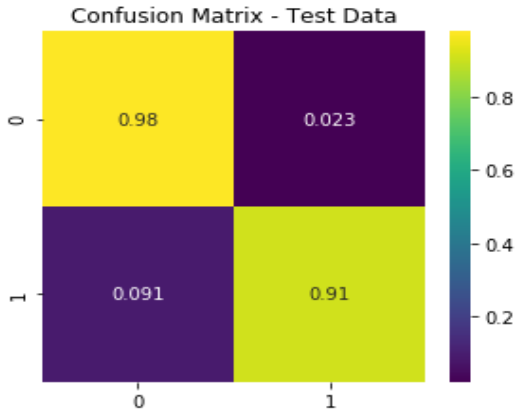

Figure 11: Confusion Matrix using SVM (test data)

6. Conclusion

The fraud detection model is trained using the European credit card fraud data. The credit card fraud detection model using Support Vector Machines, K-Nearest Neighbor and Logistic Regression for credit card fraud detection has been successfully implemented. The accuracy scores, for all the three classifiers turned out to be outstanding.

ACKNOWLEDGEMENT

I thank my Professor Dr. Liu for taking the class Data Mining and making me learn different techniques to solve the problems and apply them in real world.

REFERENCES

[1] Machine Learning Group. 2018. Credit Card Fraud Detection. (March 2018). Retrieved December 20, 2018 from https://www.kaggle.com/mlg-ulb/creditcardfraud

[2] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2016. Anomaly Detection. *Encyclopedia of Machine Learning and Data Mining* (2016), 1–15. DOI:http://dx.doi.org/10.1007/978-1-4899-7502-7_912-1

[3] Roberto Marmo. Data Mining for Fraud Detection System. *Encyclopedia of Data Warehousing and Mining, Second Edition*, 411–416. DOI:http://dx.doi.org/10.4018/978-1-60566-010-3.ch065

[4] Victoria Hodge and Jim Austin. 2004. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review* 22, 2 (2004), 85–126. DOI:http://dx.doi.org/10.1023/b:aire.0000045502.10941.a9

[5] Roberto Marmo. Data Mining for Fraud Detection System. *Encyclopedia of Data Warehousing and Mining, Second Edition*, 411–416. DOI:http://dx.doi.org/10.4018/978-1-60566-010-3.ch065

[6] Ankit Mishra and Chaitanya Ghorpade. 2018. Credit Card Fraud Detection on the Skewed Data Using Various Classification and Ensemble Techniques. *2018 IEEE International Students Conference on Electrical, Electronics and Computer Science*

(SCEECS) (2018).
DOI:http://dx.doi.org/10.1109/sceecs.2018.8546939
[7] Aihua Shen, Rencheng Tong, and Yaochen Deng. 2007. Application of Classification Models on Credit Card Fraud Detection. *2007 International Conference on Service Systems and Service Management* (2007). DOI:http://dx.doi.org/10.1109/icsssm.2007.4280163
[8] Shiguo Wang. 2010. A Comprehensive Survey of Data Mining-Based Accounting-Fraud Detection Research. *2010 International Conference on Intelligent Computation Technology and Automation*(2010). DOI:http://dx.doi.org/10.1109/icicta.2010.831