Final Project Data Science 6100

Credit card fraud detection using machine learning algorithms Logistic regression, Naïve bayes, SVM and Random Forest

Professor: Ayesha Ali

Student: Mohsen Selseleh

Student No.:

1207477

Fall 2021

Github: https://github.com/Mohsenselseleh/DATA-SCIENCE-6100-FIANL-PROJECT

1.Abstract

Nowadays, there are various new means to do fraud due to the advancement of communication networks. On the other hand, payment with credit card is the most common way of payment, therefore online frauds are increased, too. Financial losses hurt the merchants, banks, and individual clients. When the bank loses money, customers should pay higher interest rates, higher membership fees, etc. Fraud affects the reputation and image of a bank; customers may no longer trust that bank and choose a competitor. Traditional methods for detection this kind of frauds are not fit because there are lots of features should be considered, Therefore, banks use the data mining techniques to build trained model by learning samples and trained model is used to make decision in new situations. The machine learning algorithms help financial organisations to detect fraud in the credit card transactions. First, the bank terminal checks a transaction to sure that pin is correct, balance is sufficient, card is not blocked, Finally, banks use predictive models to label transactions as fraud or genuine and if it is necessary inform investigators and cancel credit card.The main goal of my study of fraud, it is so imperative task for banks. My goal is prediction fraud and comparison performance Classification models ;Logistic Regression, Naïve Bayes, Random Forest and Support Vector Machine. 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 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'. The result shows that although all models had good performance, but some models (RF and LR) had good accuracy and the highest AUC.

Introduction

Fraud is an ancient activity; it means cheating innocent people and make loss to their properties. Credit card fraud includes stealing the password, the credit card number, etc, from the cardholder by different methods as scams, phishing, phone calls or SMS, etc. Billions of dollars are stolen annually because of credit card fraud. Increasing in credit card transactions, as a share of the payment system, there has also been an increase in credit card fraud. Meanwhile, credit card fraud helps fund organized crime, international narcotics trafficking, and terrorist financing. Therefore, the banks are trying to decrease their losses resulting from card fraud. Banks generally use rule-based systems for credit card fraud detection. These systems are formulated based on the experience of fraud experts and the results of fraud investigations. Each credit card transaction is reviewed on basis to the rule set, and an alarm is sent if the transaction matches one or more rules. When new fraud cases occurred, new rules are added to these rule-based systems that its disadvantages is manual process. By contrast, artificial intelligence (AI) models learn from past transaction data. Methods AI models are divided into two types, supervised and unsupervised, both of which have been used in credit card fraud detection. In supervised fraud detection, both fraudulent and legitimate historical transactions are used for training. In unsupervised fraud detection, the spending behavior of each cardholder is modelled using that cardholder’s past transactions. When a new transaction does not fit the established behavior model, it is potentially fraudulent.

2.Background

Literature review

A comparison of logistic regression and naïve bayes is done by Ng et al (2002). Their results show that if the number of training examples is increased, naive Bayes perform better, but discriminative logistic regression generally overtakes the performance of naive Bayes.Kültür et al (2016) used six models decision tree, random forest, Bayesian network, Naïve Bayes, support vector machine, and K* models, to form an ensemble for the detection of credit card fraud. They focused on the voting mechanisms used by the ensemble and proposed optimistic, pessimistic, and weighted voting strategies. Their proposed model is called optimistic, pessimistic, and weighted voting in an ensemble of models. Maes et al (2002) have studied decision tree, neural networks, and logistic regression, they show that bayesian network is better than neural network in finding credit card fraud.Bhattacharyya et al (2011) used the dataset, which was obtained from an international credit card operation. They compare three advanced data mining approaches, support vector machines, random forests, and Logistic regression, to detect credit card fraud. Their result show that RF shows highest performance; SVM and LR are similar.Arya et al (2020) They proposed a Deep Ensemble ALgorithm (DEAL) framework for predicting fraud in real-time data set. results demonstrate DEAL framework superiority over few state-of -the-art methods in catching frauds. Li et al (2012) study applied data mining techniques to analyze detailed daily account transaction data from a selected bank to find fraudulent account. The training data set was pre-processed with Bayesian Classification and Association Rule techniques were applied to analyze transactional details to identify signs of fraudulent accounts. Bahnsen et al (2016) show that every year billions of Euros are lost worldwide due to credit card fraud. They compare state-of-the-art credit card fraud detection models and evaluate how the different sets of features have an impact on the results.

Questions of interest

In this research I try to find some answers by focusing on my questions such as: i) why and how under sampling is useful in the presence of class imbalance (because ii) how to assess performances of my models. iii)Which model(or models) are outperform than others?

Source of data

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 sourced from https://data.world/raghu543/credit-card-fraud-data.

3.Statistical Methods

This research is a Supervised machine learning one which there are pre-defined set of “training examples”, it facilitates its ability to receive to an accurate conclusion for new data set.Logistic Regression is a type of generalized linear models, it is a function of independent variables for the prediction the probability of a binary (nominal or ordinal) variable that the probability value changes between 0 and 1.

Random Forest is an ensemble of decision trees, generally trained via the bagging method, it is more convenient and optimized for decision trees. The random forest searches for the best feature among a random subset of features (Geron,2019)

Naïve Bayes is a classification method based on Bayes’ theorem. It assumes that the features are independent. It chooses the decision based on the highest probability (Geron,2019).Burkov (2019) stated that gradient boosting is one the most powerful ensemble machine learning algorithms(As Random Forest) which could manage big dataset, easily and quickly. Also, it creates very accurate models.

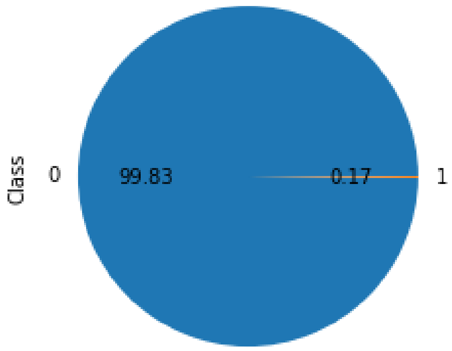
The important hyperparameters of it are the number of trees and the depth of trees. Majority of machine learning algorithms have their limitations, but ensemble learning approach could boost the performance prediction. Two principal ensemble learning methods are boosting and bagging.

4.Results

Descriptive statistics

I chose by sample random 10 percent of dataset as my sample and analyse it.

Table 1: Pie Chart for Y (variable fraud detection: 0 means no fraud and 1 is fraud)

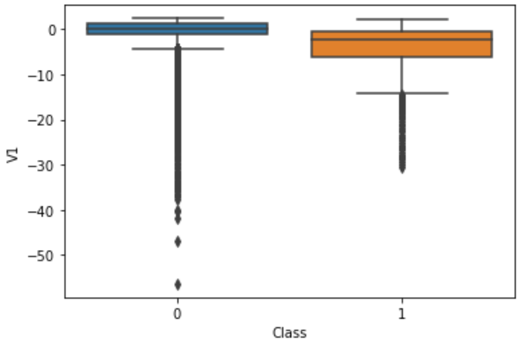


I see that dataset is highly inbalanced and I used undersampling method to balance it.

Table 2: Features

Features	Definition	Type of feature
V1 to V28	are the principal components obtained with PCA	Interval/Independent variable
Time	Contains the seconds elapsed between each transaction and the first transaction in the dataset.	Interval/Independent variable
Amount	The transaction Amount	Interval/Independent variable
Class	is the response variable and it takes value 1 in case of fraud and 0 otherwise.	Nominal/Dependent variable

Table 3: Box plot

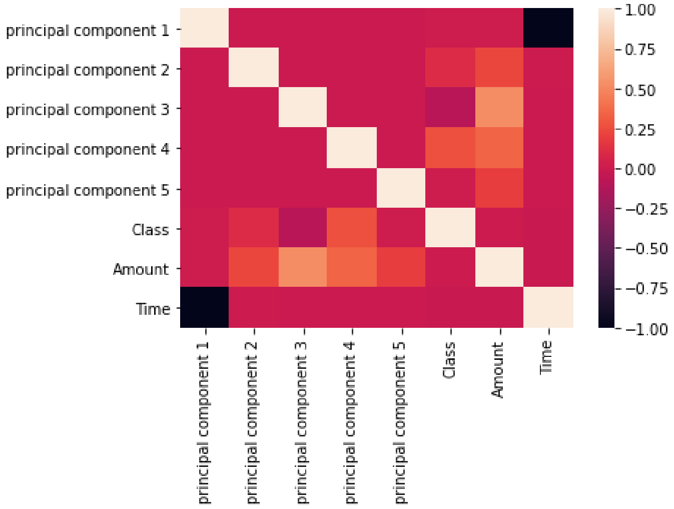


I see that there were outliers in features, and I handle them by Replacing outliers with Q1 and Q3

Table 4. Descriptive statistics of features

	principal component 1	principal component 2	principal component 3	principal component 4	principal component 5	Class	Amount	Time
count	2.84e+05	2.84e+05	2.85e+05	2.84e+05	2.85e+05	284807.	284807.00	284807.0
mean	-5.19e-10	-1.08e-17	-2.34e-16	8.475e-18	6.708e-17	0.0017	88.35	94813.86
std	4.74e+04	1.94e+00	1.65e+00	1.44e+00	1.40e+00	0.04	250.12	47488.14
min	-7.8e+04	-2.3e+00	-2.2e+01	-1.1e+01	-3.1e+01	0.0000	0.0	0.00
25%	4.45e+04	-1.5e+00	-7.99e-01	-8.78e-01	-6.82e-01	0.0000	5.60	54201.5
50%	1.01e+04	4.79e-02	-6.56e-02	9.23e-02	-6.47e-02	0.0000	22.00	84692.00
75%	4.06e+04	8.69e-01	6.00e-01	7.30e-01	8.808e-01	0.0000	77.165	139320.50
max	9.48e+04	5.71e+01	7.24e+01	8.34e+01	6.21e+01	1.00	25691.16	172792.0

Table5. Heatmap dimensionality reduction



I have used PCA for Reducing the number of input variables (Time and v1 to v28 to five components) for a predictive model and calculate correlation between them. I could see multicollinearity here, there is strong relation between Amount and principal component 3. I dropped principal component 3.

Table 6. Metrics for Machine learning models

	Accuracy	Precision	Recall	F1	Mean Absolute Error	AUC
RF	91.2	92.2	92.2	92.20	7.7	96
SVM	90.5	94.59	45.45	61.40	29.73	71.32
LR	89.8	94.59	45.45	61.40	29.73	89.8
NB	82.26	94.59	45.45	61.40	29.73	71.32

After split data set (30 percent for training dataset) and normalised dataset, I have trained models and test them on test data set and calculate Metrics to compare models.

5.Conclusion

I have changed the training data set a few times and results did not change a lot. Therefore, models are stable, therefore all the models are stable. I could see that although all models had good performance, but some models (RF and LR) had good accuracy and the highest AUC. I could see that RF as an ensemble method, outperformance all the models because I could see that it has the highest accuracy and AUC. I suggest that by performing RF, banks could find Fraudulent transactions. For future research, I think that it is better that try to select different Cross-validation method technique and training dataset to detect possibility decrease bias and increase accuracy in the models and on the other hand because use another kind of ensemble methods as XGBoosting to detect whether accuracy in prediction fraud will be increased. If this study could do on another kind of dataset, it is possible to consider possibility generalization its result.

6.Data Ethics Impact Statement

There is an important ethical issue with my research, if Banks detect fraud with machine learning models that makes decision-making automated and removes human agents, therefore it is possible that banks just consider some customers and do not give them credit card because they think maybe they do fraud in future. It hurts to some marginalised groups that deprived them from services

References

Arya, M., & Sastry G, H. (2020). DEAL–'Deep Ensemble ALgorithm'framework for credit card fraud detection in real-time data stream with Google TensorFlow. *Smart Science*, *8*(2), 71-83.

Bahnsen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. *Expert Systems with Applications*, *51*, 134-142.

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision support systems*, *50*(3), 602-613.

Burkov,A.(2019). *The hundred-page machine learning book*. Andriy Burkov

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Kültür, Y., & Çağlayan, M. U. (2017). Hybrid approaches for detecting credit card fraud. *Expert Systems*, *34*(2), e12191.

Li, S. H., Yen, D. C., Lu, W. H., & Wang, C. (2012). Identifying the signs of fraudulent accounts using data mining techniques. *Computers in Human Behavior*, *28*(3), 1002-1013.

Maes, S., Tuyls, K., Vanschoenwinkel, B., & Manderick, B. (2002). Credit card fraud detection using Bayesian and neural networks. In Proceedings of the 1st international naiso congress on neuro fuzzy technologies, 261-270.

Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information procesing systems* (pp. 841-848).

Appendix

Data set is sourced from

https://data.world/raghu543/credit-card-fraud-data