CS4035 – Fraud Detection

student number:4795245

May 2019

1 Introduction

This assignment will focus on fraud detection of credit card. The challenging points of this task lie in the properties of fraudulent data records:

- Data set is imbalanced due to the majority of the data is benign
- Criminals try to hide unusual data among normal data
- Data set to deal with can be extremely large
- Data set is sometimes unlabeled

In this assignment only first two properties are taken into consideration. This report contains three parts. The first part will give an overview of the data. Analysis on data will be based on several data visualization methods. The second part covers data processing methods applied on data set, which aims for tackling data imbalance problem. In the third part, classifiers will be trained on processed data so as to detect fraud in credit card transaction.

2 Data analysis

2.1 Introduction to data set

As can be seen in image 1, raw data set of fraud detection is a labeled data set with 17 attributes. Among them, **Simple journal** is used as label. The three status – **Settled,Refused** and **Chargeback** represent normal, not sure, fraud respectively.

Figure 1: A slice of raw data

txid	bookingdate	issuercountrycode	txvariantcode	bin	amount	currencycode	shoppercountrycode	shopperinteraction	simple_journa	cardverificationcodesupplied	cvcresponsecode	creationdate	accountcode	mail_id	ip_id	card_id
1	11/9/2015 14:26	MX	mccredit	530056	64800	MXN	MX	Ecommerce	Chargeback	TRUE	0	7/1/2015 23:03	MexicoAccount	email68370	ip111778	card184798
2	11/9/2015 14:27	MX	mccredit	547046	44900	MXN	MX	Ecommerce	Chargeback	TRUE	0	7/2/2015 4:50	MexicoAccount	email101299	ip78749	card151595
3	11/23/2015 16:34	MX	mccredit	528843	149900	MXN	MX	Ecommerce	Chargeback	TRUE	0	7/2/2015 14:30	MexicoAccount	email278604	ip70594	card242142
4	11/23/2015 16:34	MX	mccredit	547146	109900	MXN	MX	Ecommerce	Chargeback	TRUE	0	7/3/2015 7:53	MexicoAccount	email47409	ip113648	card181744
5	11/9/2015 14:26	MX	visaclassic	477291	89900	MXN	MX	Ecommerce	Chargeback	TRUE	0	7/8/2015 18:35	MexicoAccount	email205501	ip83553	card97271

2.2 Data visualization

Four different plots are shown in image 2.2.

1. Observation based on currencycode is shown in (a). As can be seen, the average amount of fraud is higher than normal transactions. And it shows that countries with better economy tend to have higher fraud amount.

2. figure (b) shows the distribution of different currency. Although there are some outliers(can be seen as the purchase of luxury goods) in normal transactions, the distribution of them are similar in general. However the distribution is not constant in fraud amount.

3. Plot (c) shows the density estimation of fraud(blue color) and normal(orange color) transactions. Normal transactions gather around 9000 EUR while fraud is much higher.

4. Plot (d) shows the number of fraud each hour. Most fraud transactions happen between evening and next day's morning. So transactions during that time period may have higher chance to be fraud.

2.3 Data processing

Feature selection: Not all features provided are related to fraud detection. Only following features are used in this assignment:

1. simple journal as its the label of samples. It is relabeled to 1 if its fraud transaction and 0 otherwise.

2. Time of transaction: Transactions happen during evening to morning are more likely to be fraud

3. Amount of money: Fraud transactions tend to be higher than normal transactions.

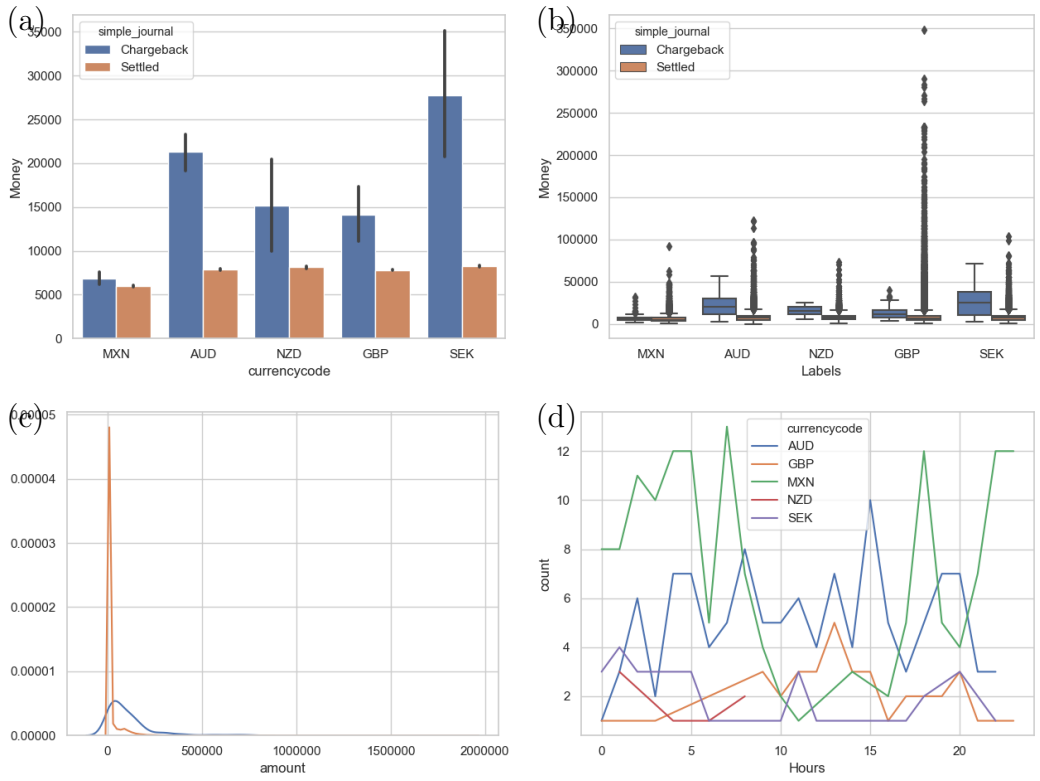

Figure 2: Data visualization based on 4 different methods.(a):Transaction amount of each currency (b): distribution of each currency (c): density estimation of transactions (d): number of transactions each hour

4. Whether the issuercountrycode matches shoppercountrycode will be considered as the difference rate in fraud transaction is 4% ,while in normal transaction it's only 2.9%. This column is set to one if the two countrycodes match and zero otherwise.

5. Currencycode: Fraud happens more often in some places and the amount of money is also related to currency. One-hot encoding is used here as not all classifiers can handle non-numerical data.

Other features seem to be less significant.For example, The rate that currency matches the shoppercountrycode is almost the same in both fraud data(2.8%) and normal data(2.9%). So only 5 features are used in later classifier training process.

Noise reduction: The status "Refused" of label **Simple journal** is uncertain. Although the probability of being fraud is higher for credit cards with fraud history, its still hard to analyse(semi-supervised learning might provide some improvement). So data with "Refused" label are removed so as to wipe out noise. After that, data size reduced from 290382 to 237036.
Credit card fraud is normally a cross national crime. Thus, different currencies('MXN' 'AUD' 'NZD' 'GBP' 'SEK') are transferred to the same unit(EUR) so as to be compared easily.

3 Dealing with imbalance

There are in all 345 fraud transactions, which is a quite small number compared with all 237036 samples in this data set(after pre-processing). To deal with this problem, an oversampling method SMOTE combined with cleaning method tomek are applied and compared with undersampling method using random selection and original data. 4 different machine learning algorithms, namely logistic regression, random forest, 3-NN and SVM are used. performance comparison is done using ROC curve. Image is shown in figure 3. It's noticeable that the time complexity of svm is more than quadratic with the number of samples, so it doesn't work well with so many samples in this assignment. Thus only 3 graphs are generated.
Although cross-validation is not appropriate for time series data set, 10-fold cross validation is still used in this assignment as otherwise a incredibly large number of models needs to be trained. Before using any sampling method, the minority data only account for 1.5% of all data. After SMOTE+TOMEK, percentage goes up to 16.3%. And undersampling increases the percentage to 16.7%.
As can be seen in the figure, classifiers like random forest and logistic regression performs better than KNN. Although the improvement on already well-performed classifier is not significant, in general sampling methods perform better. And Smote+tomek shows satisfying performance among all classifiers. The reason lies in that the data set is imbalanced and lack of positive samples, which can easily lead to a wrong classifier. Oversampling method like smote can adjust the balance of data by adding positive samples. To some extend it can be seen as an ensemble learning method and is more resistent to noise points. Although it may bring some impractical

points, by using tomek (can be seen as a combination of oversampling and undersampling) we can exclude some of the them.
There are also some disadvantages of sampling methods. They may change the original distribution and over-emphasize the effect of particular samples. So sometimes, they will cause overfitting.

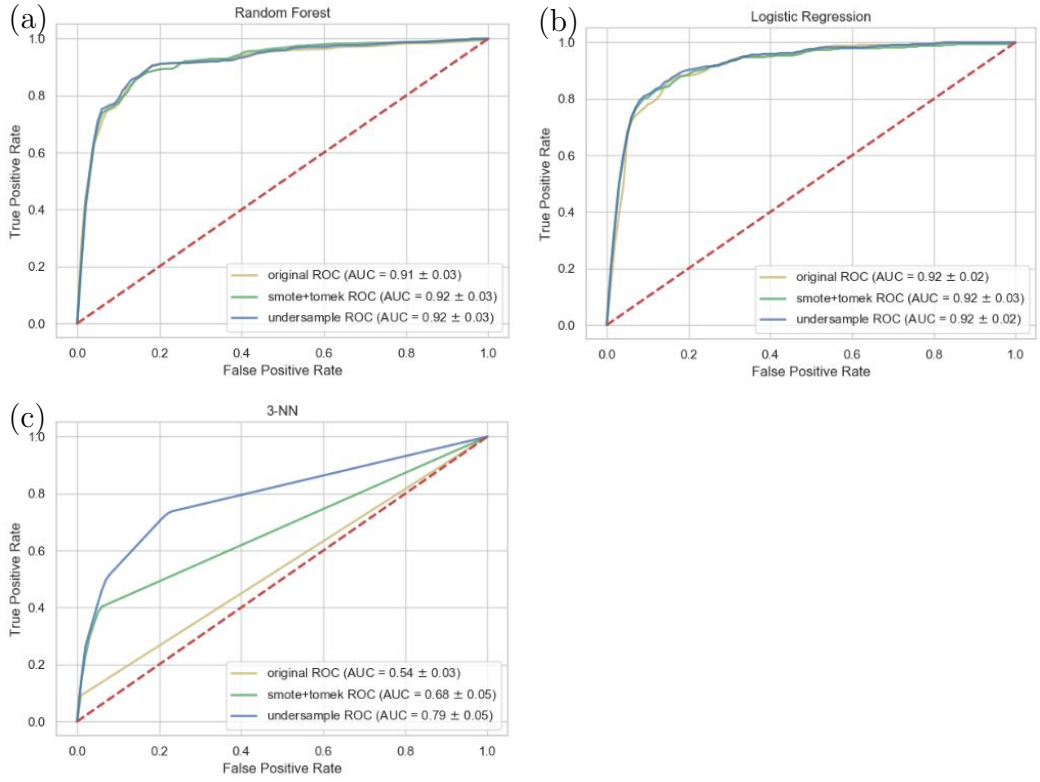

Figure 3: (a):ROC - Random Forest (b): ROC - Logistic Regression (c): ROC - 3NN

4 Training of classifiers

4.1 Black box:

The black box algorithm used here is random forest algorithm. Random forest is an ensemble learning method, gathering the votes of several decision

trees to predict an unknown label so as to provide better performance.
Single tree is often insufficient to correctly classify new data set. This is due to the simplicity of single decision tree brings bias to the system. On the countrary, Random forest can handle bias-variance-tradeoff more properly. That's why aggregating trees can provide better performance than a simple model.
Apart from that, different trees in the forest are independent as the training of each tree is based on bootstrapping method. Samples are random selected, which can effectively decrease the probability to overfit and at the same time become resistent to noise. This random selection also decrease the entropy of training set, thus brings information gain to the system.
These advantages of Random forest, and its good performance shown in image 3 are the reasons why I choose this algorithm. For its implementation, data pre-processing step is the same as mentioned in previous section but shuffled afterwards. Then training data is processed using SMOTE+tomek. Threshold is optimized. 10-fold cross-validation is used to find the best classification threshold. The test result of black box algorithm is shown below in figure 4.

4.2 White box:

I choose KNN as white box classifier. As can be seen in previous section, this algorithm shows great improvement after using sampling method. Also this algorithm is relatively easy to implement.
The main concept of this algorithm is that for an input unlabeled sample, classifier tries to find its k nearest labeled samples and assign it to the majority class among the nearest samples.
For this assignment, as the amount of positive samples are very small, K should be tuned to a small number. I tried three K and among them K = 3 performs best. It means the unknown data will be simply classified as the majority class of the 3 samples next to it. And the sampling method used here is undersampling. Figure 4 shows that when K is extremely small, it can be easily affected by noise. And by using optimized threshold and undersampling method, KNN shows satisfying result.

	AUC	F1	Accuracy	Precision	Recall	Confusion	Threshold
Logistic	0.92	0.05	0.97	0.02	0.45	23059,610,19,16	0.6
White Box	0.94	0.05	0.98	0.03	0.34	23266,403,23,12	0.8
Black Box	0.92	0.05	0.98	0.03	0.45	23119,550,19,15	0.4
3-NN	0.68	0.04	0.99	0.02	0.19	23362,307,28,7	0.8
2-NN	0.66	0.03	0.98	0.02	0.21	23260,409,27,7	0.8
1-NN	0.63	0.03	0.97	0.01	0.27	23044,625,25,9	0.7

Figure 4: Test Result of different classifiers, In Confusion Column, from left to right are namely true negative, false positive, false negative, true positive. Test data size is $\frac{1}{10}$ of all data. Apart from White Box(3-NN), other classifiers all use SMOTE+TOMEK