Credit Card Fraud Detection Project Report

Osman Furkan Kınlı

[furkan.kinli@ozu.edu.tr](mailto:furkan.kinli@ozu.edu.tr)

**Abstract**

In this project, it has been worked on detecting the fraud transactions for imbalanced credit card transactions data with famous machine learning approaches in Scikit-Learn library. It has been used oversampling and undersampling methods from imbalanced-learn library. The main purpose of this approach is to balance the dataset to be able create a meaningful machine learning model. Finally, the performance of our models for detecting fraud transactions of oversampled and undersampled datasets is compared with each other by helping of K-fold Cross Validation method and is measured with the area under precision-recall curve.

1. **Introduction**

The main purpose of this project was to gain experience in taking up and detecting the fraud transactions of sample credit card transactions data with logistic regression approach. However, this dataset is highly unbalanced, it means that the positive class (frauds) account for 0.17% of all transactions. It requires to be balanced initially. First, this dataset has been loaded with method of reading csv files in Pandas library which creates a data frame with 30 feature-columns, 1 label-column and 284.807 rows. There are mainly 3 different kinds of features in this dataset. The first one is the time column which contains the seconds elapsed between each transaction and the first transaction in the dataset. Secondly, 28 columns are transformed just because of confidentially issue on credit card information. The last type of features is about the amount of the transactions. Following that, it was attempted to use logistic regression approach at our problem to be able to obtain the most accurate result for detecting the fraud transactions in this dataset. Finally, the training set results have been validated with the K-fold Cross Validation technique which splits the data into K parts and uses each small part as the test set of the remaining parts and obtains K different test accuracies trained with different parts of whole data and the accuracy has been measured with the area under precision-recall curve.

1. **Problem and Dataset Information**

The dataset has been collected and analyzed during a research collaboration of Worldline and the Machine Learning Group of ULB (Université Libre de Bruxelles) on big data mining and fraud detection. This dataset, retrieved from Kaggle, has been used in this project for fraud detection analyzing by using logistic regression algorithm.

This dataset has 30 features columns and 1 label column such as fraudulent and genuine. First, the time column has been converted to time zone column to be able to obtain more meaningful inferences from this feature. Moreover, the amount column has been rescaled to the properties of a standard normal distribution with a mean of zero and a standard deviation of one. The other 28 feature columns have been already transformed with Principal Component Analysis method to protect the information of the credit card owners.

* 1. **Credit Card Transactions Dataset**

**Description:** Credit Card Transactions

**Samples total:** 284.807

**Dimensionality:** 30 & -time, [confidential], amount-

**Features:** float numbers

**Targets:** integer numbers & -0 for genuine, 1 for fraudulent-

**Problem:** How could we detect the fraud in credit card transactions (as fraudulent or genuine) by the help of machine learning algorithms?

1. **Machine Learning Algorithms**
   1. **Logistic Regression**

Logistic Regression (LR) is a supervised machine learning algorithm which can be used for regression challenges. This algorithm is predictive analysis to conduct with binary dependent variable and basically tries to explain the relationship linearly between one dependent variable and single/multiple independent variables.

The code in SKLearn Library:

class sklearn.linear\_model.**LogisticRegression**(penalty=’l2’, dual=False, tol=0.0001, C=1.0, fit\_intercept=True, intercept\_scaling=1, class\_weight=None, random\_state=None, solver=’liblinear’, max\_iter=100, multi\_class=’ovr’, verbose=0, warm\_start=False, n\_jobs=1)

Some parameters of Logistic Regression:

* **penalty** (str): {‘l1’, ‘l2’}
* **dual** (boolean): Dual or primal formulation
* **C** (float): Inverse of regularization strength
* **fit\_intercept** (boolean): Whether to add bias/intercept to the decision function
* **solver** (str): {‘newton-cg’, ‘lbfgs’, ‘liblinear’, ‘sag’, ‘saga’}
* **max\_iter** (int): The maximum number of iterations that solver converges
* **tol** (‘auto’, True, False): Tolerance for stopping criteria
* **verbose** (int): verbosity mode
* **random\_state** (int): RandomState instance

1. **Balancing the dataset**
   1. **Oversampling the dataset**

Oversampling can be used to repeat some samples by different methods (randomly, smote etc.) to balance the number of samples between the dataset. In this experiment, SMOTE (Synthetic Minority Over-Sampling Technique) method has been used to create over-sampled dataset. SMOTE mainly proposes several variants by specifying the samples to consider during the resampling. While the borderline version detects which point to select which are in the border between two classes, the SVM version uses the support vectors found using an SVM algorithm to create new samples.

* 1. **Undersampling the dataset**

Undersampling can be used to remove some samples by different methods (randomly, neighborhood rule etc.) to balance the number of samples between the dataset. In this experiment, “RandomUnderSampler” method has been used to create under-sampled dataset. RandomUnderSampler is the naivest way of performing such selection by randomly selecting a given number of samples by the targeted class.

1. **Evaluation**

In this section, the performances of logistic regression approach on original, oversampled and undersampled datasets are presented and to assess the performance, K-fold cross validation method has been chosen for logistic regression model. The accuracies are measured with the area under the precision-recall curves.

* 1. **Performances of Logistic Regression Approach**
     1. **The original dataset**

For the original dataset, logistic regression approach has been used to detect fraudulent transactions in this project. After preprocessing data by converting the time column to time zone and rescaling the amount column, the logistic regression model has been trained with default parameters. Therefore, 10-fold cross validation accuracies for the original dataset:

10-Fold Cross Validation Scores for Original Dataset:

[0.98483253 0.99957868 0.99982444 0.99950844 0.99929778 0.99908708 0.99968399 0.99870084 0.99834972 0.99880618]

According to the classification report method in Scikit-Learn library:

label precision recall f1-score support

0 1.00 1.00 1.00 85307

1 0.71 0.65 0.68 136

avg / total 1.00 1.00 1.00 85443

The precision-recall curve:

![metin, harita içeren bir resim

Çok yüksek güvenilirlikle oluşturulmuş açıklama]()

* + 1. **The oversampled dataset**

For the oversampled dataset, logistic regression approach has been used to detect fraudulent transactions in this project. After preprocessing data by converting the time column to time zone and rescaling the amount column, the logistic regression model has been trained with default parameters. Therefore, 10-fold cross validation accuracies for the oversampled dataset:

10-Fold Cross Validation Scores for Oversampled Dataset:

[0.91949212 0.9559827 0.95941193 0.95935917 0.9587085 0.96217157 0.968362 0.9687489 0.9684851 0.97451725]

According to the classification report method in Scikit-Learn library:

labels precision recall f1-score support

0 0.93 0.99 0.96 85149

1 0.99 0.93 0.96 85440

avg / total 0.96 0.96 0.96 170589

The precision-recall curve:

![ekran görüntüsü içeren bir resim

Çok yüksek güvenilirlikle oluşturulmuş açıklama]()

* + 1. **The undersampled dataset**

For the undersampled dataset, logistic regression approach has been used to detect fraudulent transactions in this project. After preprocessing data by converting the time column to time zone and rescaling the amount column, the logistic regression model has been trained with default parameters. Therefore, 10-fold cross validation accuracies for the undersampled dataset:

10-Fold Cross Validation Scores for Undersampled Dataset:

[0.97 0.99 0.97959184 0.90816327 0.89795918 0.92857143 0.96938776 0.8877551 0.89795918 0.85714286]

According to the classification report method in Scikit-Learn library:

labels precision recall f1-score support

0 0.90 0.96 0.93 150

1 0.96 0.89 0.92 146

avg / total 0.93 0.93 0.93 296

The precision-recall curve:

![ekran görüntüsü içeren bir resim

Çok yüksek güvenilirlikle oluşturulmuş açıklama]()

1. **Conclusion**

In this experiment, the main purpose is to get hands-on experience in using different balancing methods on imbalanced credit card transactions to detect the fraudulent ones by the help of logistic regression approach and how to handle these kinds of problems. It is obvious that this project has served this purpose.

**References**

* Scikit-Learn Package for Python

[http://scikit-learn.org/stable/supervised\_learning.html#](http://scikit-learn.org/stable/supervised_learning.html)

* Imbalanced-Learn Package for Python

<http://contrib.scikit-learn.org/imbalanced-learn/stable/index.html>

* Logistic Regression

<http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>

* Oversampling with SMOTE

Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015

* Undersampling with Random Under Sampler

<http://contrib.scikit-learn.org/imbalanced-learn/stable/auto_examples/under-sampling/plot_random_under_sampler.html>