Credit Card Fraud Detection

Maria Doda

Fall 2020

—

CSC 5825 – Machine Learning

—

Dongxiao Zhu

Section 1. Background/Introduction

Credit card fraud represents a big concern for global finance institutions. Due to the high dependency on the internet and the rise in both online transactions as well as e-commerce platforms, the number of cases of credit card frauds has exponentially increased. Since traditional rules-based systems have proven to not be effective to uncover increasingly sophisticated fraud techniques, the financial sector is looking to develop machine learning algorithms to help prevent or mitigate this issue. Many institutions have focused on integrating basic machine learning algorithms such as support vector machines, linear approaches and clustering classifiers to predict unknown fraud patterns, which traditional rules-based systems fail to do. In fact, detecting fraudulent transactions is a challenging task because frauds data are highly imbalanced, fraudsters always look for new fraud schemes or patterns to avoid being caught and lastly, they try to either conceal or blend in their illegal activities.

Through this project, I will analyze a dataset of credit card transactions occurred over a two-day period by European Cardholders during the month of September. The dataset contains 284,807 transactions. Among those 492 were identified as fraudulent. Each of the transaction has 31 numerical features, which are the result of a PCA transformation due to privacy concern. The only features that have not been transformed are “Time” and “Amount”. In addition, features have been scaled in order to implement PCA. The Time feature corresponds to the time elapsed since the first transaction whereas the Amount feature contains the transaction amount. The feature Class is the response variable, and it has value 1 in the case of fraud, and 0 otherwise. The dataset is highly imbalanced with a class imbalance ratio of 0.172%.

The goal of this project is to build models to predict whether a credit card transaction is fraudulent using a supervised imbalanced binary classification and unsupervised anomaly detection. In addition, visualizations techniques will be used to better understand the data and unveil interesting patterns.

Section 2. Methods

The models used to predict credit card fraud are Logistic Regression, Support Vector Classifier and random forest. First, I import the data and libraries. Then, I perform exploratory data analysis and plot the results to check for common or interesting patterns. In particular, I count the occurrences of fraud and not fraud cases as well as identify the mean of all card transactions (88.00 USD). Before preprocessing the data, I split it into features and response variable. I use a test size of 20% and I stratify the split on the response variable due to the very few fraudulent transactions found in the dataset. I focused on each of the features, Time and Amount, only on the training set to get more insight on them.

Subsequently, I utilize a non-parametric method called Mutual Information to estimate the mutual dependance between Time and Amount variables. Mutual information of 0 represents no dependence, and higher values represent higher dependence. Due to the fact that I have 227,845 training samples and a target variable that is discrete, use mutual_info_classif.

Once the preprocessing phase is finished, I begin training my models. For Logistic Regression and Support Vector Classifier, I use first the class SGDClassifier to implement multiple linear classifiers with SGD training because it makes learning much faster on large datasets. The models also will be implemented as a machine learning pipeline that includes StandardScaler for data standardization. Then, I perform a grid search over several hyperparameter choices. The grid search, is implemented by GridSearchCV, uses StratifiedKFold with 5 folds for the train/validation splits. I utilize matthews_corrcoef as my scoring metric. For the other model, I implement RandomForestClassifier. Rescaling the data is not needed for tree-based models. Since the random forest takes longer to train on this large dataset, I don’t perform hyperparameter grid search, but I only specify only the number of estimators.

After obtaining the performance results of each model and seeing that Random Forest performed better than the other 2 models, I decide to evaluate its performance on the test set. According to the cross validated MCC scores, Random Forest performed better on the test set than on the training set due to the model being trained on the entire training dataset instead of smaller CV folds.

Section 3. Experiment, Results and Discussion

After discovering that most of the transactions are not fraud 99.83% of the time, while fraud transactions occur 017% of the time in the data frame (Figure 1), I was able to see that most of the transactions happened in daytime and that Mean of transaction amount is 88 USD and 75% quartile is 77 USD (Figure 2).Figure 1: Distribution of time

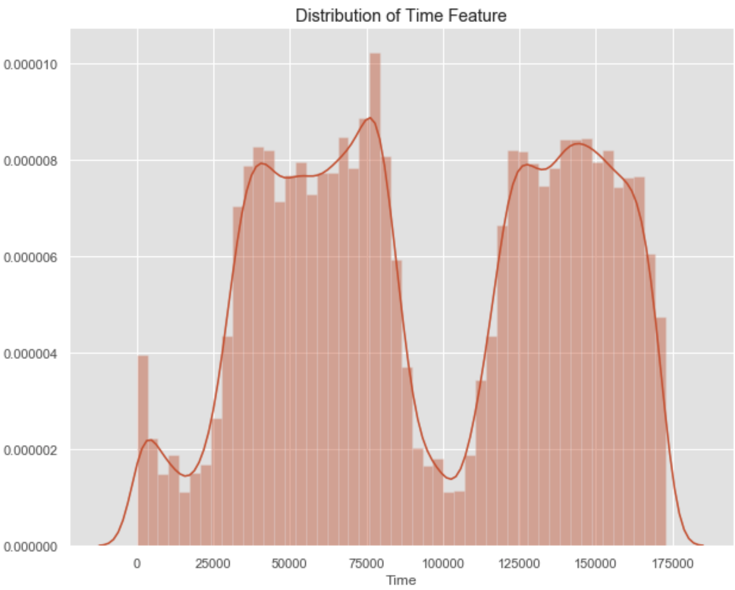


Figure 1: Distribution of Time

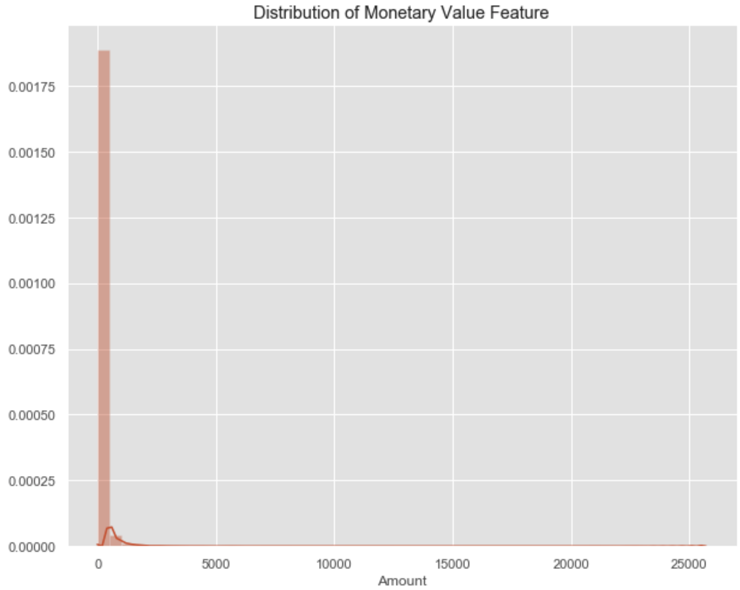
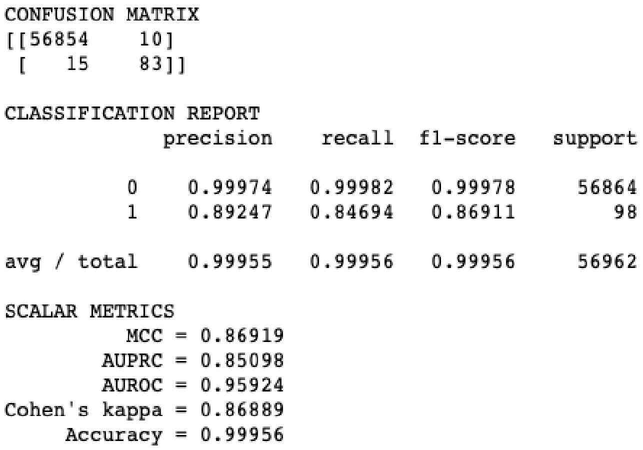


Figure 2: Distribution of Amount

When building the models, I found that linear SVC performed better than logistic regression, and with a high level of regularization and that, without any hyperparameter tweaking, the Random Forest performed much better than the linear SVC on the training set.

Since cross validated MCC scores displayed that random forest is the best-performing model, I decided to evaluate its performance on the test set. The result of the confusion matrix proved that this model performs even better on the test set with AUROC of 0.95924. (Figure 3) This could be due to the refit model being trained on the entire training data set, and not on the smaller CV folds.



```
CONFUSION MATRIX
[[56854    10]
 [   15    83]]

CLASSIFICATION REPORT
             precision    recall  f1-score   support

          0    0.99974   0.99982   0.99978     56864
          1    0.89247   0.84694   0.86911        98

avg / total    0.99955   0.99956   0.99956     56962

SCALAR METRICS
          MCC = 0.86919
        AUPRC = 0.85098
        AUROC = 0.95924
 Cohen's kappa = 0.86889
     Accuracy = 0.99956
```

Figure 3: Confusion Matrix

Section 4. Conclusion

Thus far in the project, I was able to identify fraudulent credit card transactions utilizing a random forest model. In addition, I was able to find that the five variables most correlated with fraud are V17, V14, V10, V12, and V11. The following preprocessing steps were adopted to construct the predictive models:

- Split the data using a random and stratified train/test split
- Box-Cox power to transform transaction amounts and remove skewness in the data
- Mean and variance standardization of all features as part of a machine learning pipeline

In cross validation, the best linear model (logistic regression, linear SVC) achieved a cross-validated MCC score of 0.807, and a random forest achieved a cross-validated MCC score of 0.856. With additional time and computational power, I would like to use LOF and Isolation Forest Algorithm to detect outiliers.

Bibliography

Choudhary, P. (n.d.). Introduction to Anomaly Detection. Retrieved December 03, 2020, from https://www.datascience.com/blog/python-anomaly-detection

Lewinson, E. (2019, September 26). Outlier Detection with Isolation Forest. Retrieved from https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e

Maniraj, S & Saini, Aditya & Ahmed, Shadab & Sarkar, Swarna. (2019). Credit Card Fraud Detection using Machine Learning and Data Science. International Journal of Engineering Research and. 08. 10.17577/IJERTV8IS090031.

D. S. (2020, July 21). Supervised vs. Unsupervised Learning. Retrieved from https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d