The Difference Between Precision-recall and ROC Curves for Evaluating the Performance of Credit Card Fraud Detection Models

Rustam Fayzrakhmanov, Alexandr Kulikov and Polina Repp
Information Technologies and Computer-Based System Department,
Perm National Research Polytechnic University,
Komsomolsky Prospekt 29, 614990, Perm, Perm Krai, Russia
fayzrakhmanov@gmail.com, thewato@gmail.com, polina.repp@gmail.com

Keywords: Credit Card Fraud Detection, Weighted Logistic Regression, Random Undersampling, Precision-Recall curve, ROC Curve

Abstract: The study is devoted to the actual problem of fraudulent transactions detecting with use of machine learning. Presently the receiver Operator Characteristic (ROC) curves are commonly used to present results for binary decision problems in machine learning. However, for a skewed dataset ROC curves don't reflect the difference between classifiers and depend on the largest value of precision or recall metrics. So the financial companies are interested in high values of both precision and recall. For solving this problem the precision-recall curves are described as an approach. Weighted logistic regression is used as an algorithm-level technique and random undersampling is proposed as data-level technique to build credit card fraud classifier. To perform computations a logistic regression as a model for prediction of fraud and Python with sklearn, pandas and numpy libraries has been used. As a result of this research it is determined that precision-recall curves have more advantages than ROC curves in dealing with credit card fraud detection. The proposed method can be effectively used in the banking sector.

1 INTRODUCTION

Fraud detection is generally considered as a data mining classification problem, where the objective is to classify the credit card transactions as legitimate or fraudulent correctly. Detection of fraudulent transactions combined with machine learning has become an exciting subject of research over the last years [1].

The credit card fraud exhibits unique characteristics which render the task extremely challenging for any machine learning technique. The most common characteristic is that the credit card datasets are highly unbalanced, which means they admit and uneven distribution of class transactions. The fraud class is represented by only a small number of examples (minority class) while the legal class makes up the rest (majority class). The ratio from legal class size to fraud class size can vary up to hundred fold [2]. Using these datasets as training sets in the learning process can bias the learning algorithm resulting in poor accuracy on the minority class but high accuracy on the majority class [3].

Approaches of solving the problem of unbalanced classes are divided into data-level methods and algorithm-level methods (or combinations of these techniques). Data-level methods are focused on modifying the training set to make it suitable for a standard learning algorithm. There are distinguish approaches which generate new objects for minority groups (oversampling) and which remove examples from majority groups (undersampling). Algorithm-level methods are focused on modifying existing learners to alleviate their bias towards majority groups. This requires a good insight into the modified learning algorithm and a precise identification of reasons for its failure in mining skewed distributions. The most popular branch is cost-sensitive approaches, such as weighted logistic regression [4].

To evaluate the performance these approaches [5][6] use Receiver Operator Characteristic (ROC) curves, which show how the number of correctly classified positive examples varies with the number

of incorrectly classified negative examples. However, ROC curves can present an overly optimistic view of an algorithm's performance if there is a large skew.

Precision-Recall (PR) curve is an alternative to ROC curves for tasks with a large skew in the class distribution, such as a credit card fraud. Precision-recall curves are highly informative about the performance of binary classifiers, and the area under these curves is a popular scalar performance measure for comparing different classifiers [7].

In this article, a model for detecting a credit card fraud using weighted logistic regression and random undersampling techniques was built and ROC and PR curves for them were analysed.

2 EVALUATION OF A CLASSIFICATION MODEL

The aim of detection a credit card fraud is to design a binary classifier with a highest possible accuracy of fraud transactions. To design it many different machine learning technics are used; the most wide-spread of them are logistic regression, decision trees, support vector machine, its varieties and assembles. In this case the sets of data (containing dozens and hundreds of features) have become online payment transactions belonged to financial companies. Features are different information about an online purchase, such as the transaction's amount, IP-address, payment card data, etc. Since fraud transactions usually present less than 1% of the total number of transactions, the process of a classifier design is called imbalance learning, and the data is called imbalance dataset.

Since the credit card fraud task is binary a confusion matrix to evaluate a performance of approaches is used. The confusion matrix summarizes information about actual and predicted classifications performed by a classifier. Confusion matrix for binary classifiers is shown in Table 1. The table shows that four different forecast results are possible. Really positive and really negative outcomes are the correct classification, while the false positive and false negative outcomes are two possible types of errors [8].

Table 1: Confusion matrix.

Actual	Predicted	
	Positive class	Negative class
Positive class	True Positive (TP)	False Negative (FN)
Negative class	False Positive (FP)	True Negative (TN)

A false positive example is a negative example class that is wrongly classified as a positive one (legitimate transactions as fraudulent in context of the paper) and a false negative example is a positive example of the class that is wrongly classified as a negative (fraudulent as legitimate) one.

Standard performance metrics such as predictive accuracy and error rate can be derived from the confusion matrix:

$$Predictive\ Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Error\ Rate = \frac{FP + FN}{TP + FP + TN + FN}$$

The usage of a predictive accuracy and error rate leads to a poor performance for the minority class [9]. For that reason, a variety of common evaluation metrics based on confusion matrix are developed to assess the performance of classifiers for imbalanced data sets:

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$True\ Positive\ Rate = \frac{TP}{TP + FN}$$

$$False\ Positive\ Rate = \frac{FP}{FP + TN}$$

These metrics are developed from the fields of information retrieval. They are used in situations when performance for the positive class (the minority class) is preferred, since both precision and recall are defined with respect to the positive class.

Alternatively, the Receiver Operating Characteristic (ROC) can be employed to evaluate the overall classification performance. The ROC is a graphical representation that plots the relationship between the benefits (TPR) and costs (FPR) as the decision threshold varies. The ROC curve provides

evidence that the true positive rate is directly proportional to the false positive rate [10].

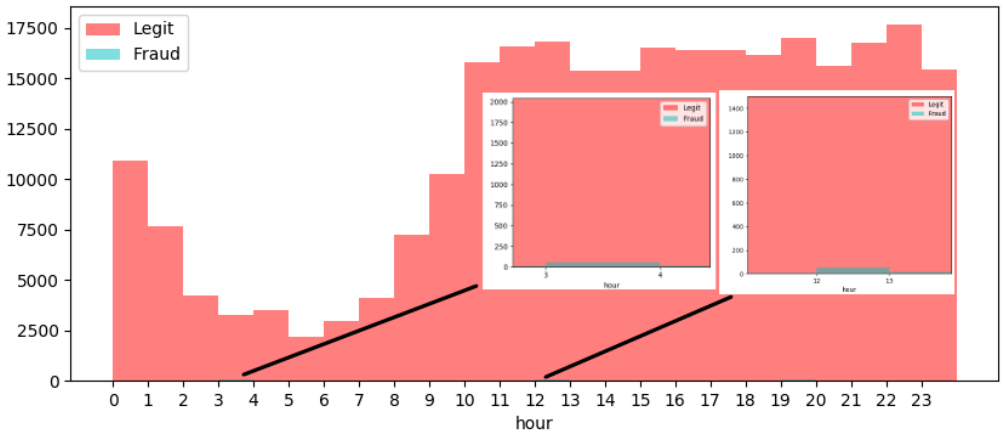

Figure 1: Distribution of the dataset.

Precision-recall (PR) curves, like the ROC curves, are an evaluation tool for binary classification that allows performance visualization. PR curves are increasingly used in the machine learning community, particularly for imbalanced datasets. On these imbalanced or skewed data sets, PR curves are a useful alternative to ROC curves that can highlight performance differences that are lost in ROC curves.

The area under curve (AUC) measure summarizes the performance of the classifier into a single quantitative measure, usually to determining what classifier is more superior. Generally, a better performing classifier has a larger AUC than that of an inferior one.

ROC and PR curves facilitate clear visualization comparisons between two or more classifiers over a large span of operating points.

Financial companies don't want to miss catching fraud (FN), therefore recall is important. However, it is necessary to consider that an accuracy lost (FP) is also money lost for companies, because they have to call the customer and verify that the purchase was authentic indeed which takes resources. Therefore, it is important to obtain high precision and recall values for the classifier.

3 EXPERIMENTS

Consider the dataset that contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where there are 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) amounts 0.172% of all transactions. It contains only numerical input variables which are the result of a principal component analysis (PCA) transformation. Due to confidentiality issues, there is no possibility to obtain the original features and more background information about the data. Features V1-V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise. Dataset is illustrated in Figure 1.

	Time	V1	V2	...	V28	Amount	Class
0	0.0	-1.359807	-0.072781	...	-0.021053	149.62	0
1	0.0	1.191857	0.266151	...	0.014724	2.69	0
2	1.0	-1.358354	-1.340163	...	-0.059752	378.66	0
3	1.0	-0.966272	-0.185226	...	0.061458	123.50	0
4	2.0	-1.158233	0.877737	...	0.215153	69.99	0

Figure 2: Dataset example.

The distribution of the dataset is illustrated in Figure 2. The data is totally unbalanced. This is a clear example where a typical accuracy score to evaluate our classification algorithm is used. For example, in case having just used a majority class to assign values to all records, a high accuracy still will be had, but all fraudulent transactions would be classified incorrectly.

To perform computations a logistic regression as a model for prediction of fraud and Python with

sklearn, pandas and numpy libraries has been chosen. Consider the confusion matrix, precision and recall metrics on the raw dataset. The matrix is illustrated in Figure 3.

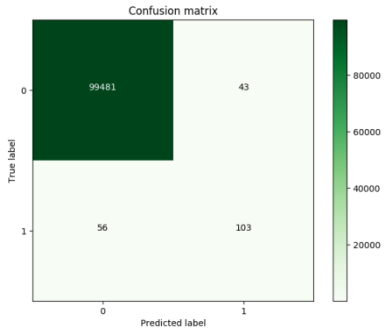

Figure 3: Confusion matrix of a model.

The recall of this model is 0.64 and precision is 0.71. These are fairly low scores. This is due to the fact that logistic regression as a standard classifier algorithm has bias to classes that have a number of instances. They tend only to predict data of most classes. The characteristics of the minority class are considered as noise and are often ignored. Thus, there is a high probability of mistaken classification of the minority class in comparison with the majority class [11]. This problem can be solved by algorithms of a family of decision trees, such as a random forest, but such algorithms are not stable to high overfitting [12].

To solve the unbalanced problem, a weighted logistic regression as an algorithm-level method and random undersampling as a data-level method was used.

Weighting is a procedure that weights the data to compensate the differences in a sample and population. In rare events, such as a credit card fraud, we tend to sample all the 1's (rare events) and a fraction of 0's (non-events). In such cases the observations have to be weighed accordingly.

Some arbitrary weights for a model to illustrate the tradeoff between precision and recall are specified. The weights to n = {1, 5, 10, 25, 50, 100, 500, 1000, 10000} are set. The results are shown in Table 2.

Table 2: Results of the model with different weight parameters.

Weight	Precision	Recall
1	0.65	0.71
5	0.68	0.71
10	0.77	0.65
25	0.81	0.41
50	0.84	0.46
100	0.85	0.27
500	0.90	0.08
1,000	0.94	0.04
10,000	0.97	0.005

Clustering, as an effective data-level technique [13], can be used. However, since the dataset has anonymous data, random undersampling is a better choice. Undersampling is one of the techniques used for handling class imbalance. In this technique, we under sample the majority class to match the minority class. So in our case, a random sample of non-fraud class to match number of fraud samples is taken. This makes sure that the training data has equal amount of fraud and non-fraud samples [14]. And then the model to the whole dataset is applied.

For undersampling random 25%, 10% and 1% legitimate samples of dataset are taken as well as random 492, 984 and 1476 legitimate samples (1x, 2x and 3x of fraud samples). The results are shown in Table 3.

Table 3: Results of the model with different random legitimate samples.

Samples	Precision	Recall
56862	0.81	0.81
28431	0.74	0.82
2843	0.28	0.88
1476	0.12	0.90
984	0.11	0.90
492	0.04	0.93

Due to the manually selecting a range of weights to boost the minority class and undersampling minority class our model has been improved to have a better recall, and in some cases, a better precision also. Recall and precision are usually tradeoffs of each other, so when both are improved at the same time, our model's overall performance is undeniably improved.

4 ANALYSIS OF PR AND ROC CURVES

For financial companies, as it has earlier been mentioned, both the high accuracy and the high completeness are important. To calculate the specific values of these metrics, different companies develop their own evaluation algorithms based on their financial strategy or use universal ones like Economic Efficiency [15]. Thus, for our

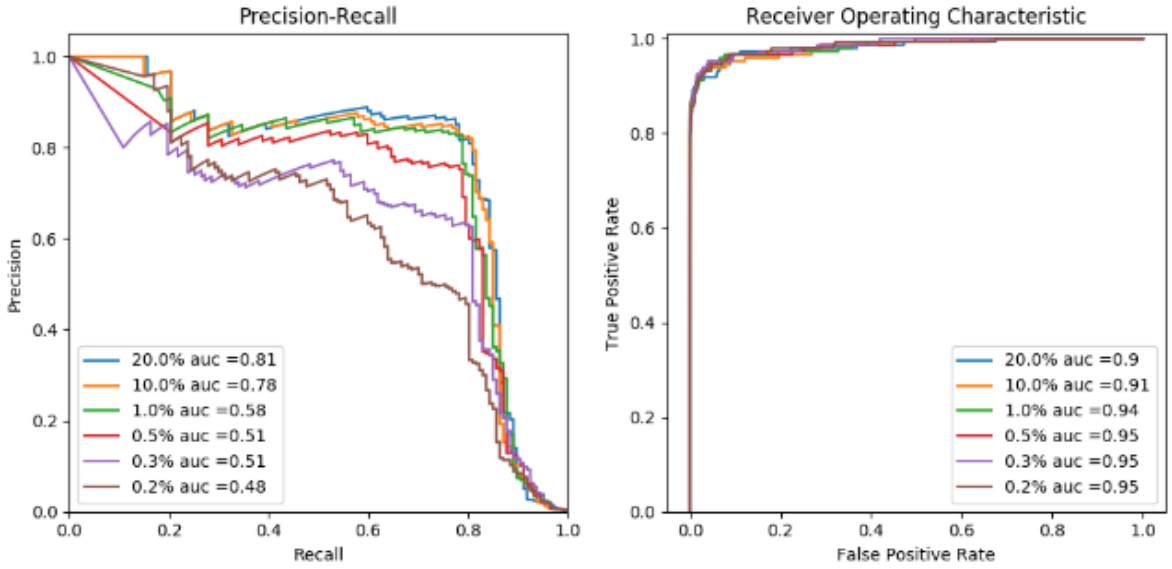

Figure 4: PR and ROC curves for random undersampling technique.

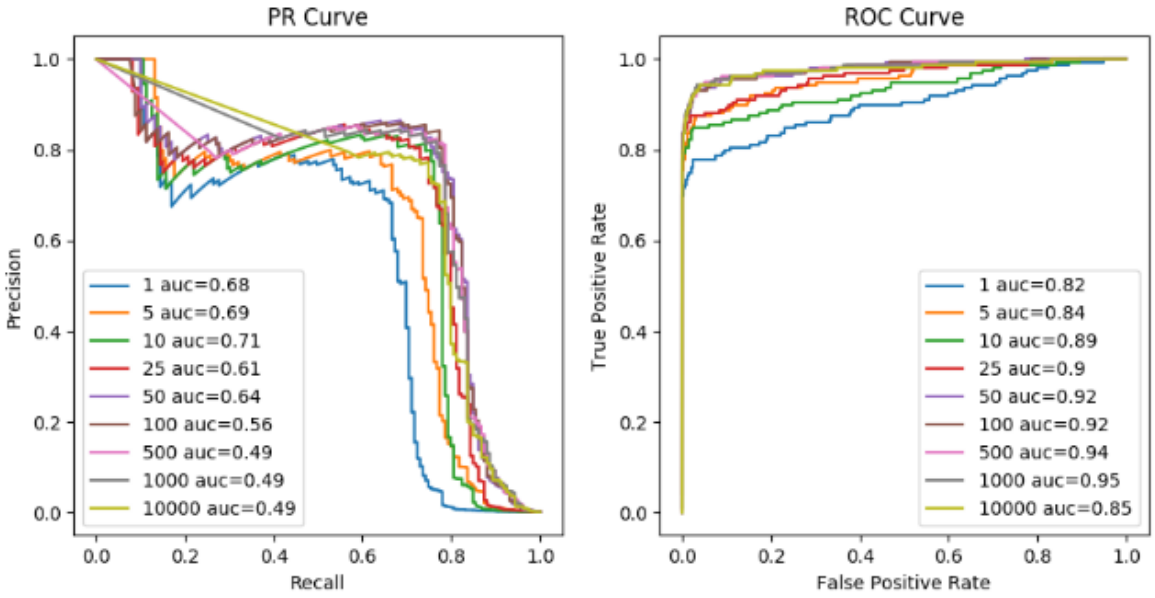

Figure 5: PR and ROC curves for weighted logistic regression.

calculations, a combination of the most possible values of precision and recall is used. To do this PR and ROC curves for both techniques (weighted logistic regression and random undersampling) are built and the area under curves (AUC) as a metric to evaluate both precision and recall is calculated. Plots of curves are illustrated on Figure 4 and Figure 5

For a PR curve, a good classifier aims at the upper right corner of the chart but the upper left corner aims at the ROC curve.

While PR and ROC curves use the same data, i.e. the real class labels and predicted probability for the class labels, different behaviour is observed, with some weights and samples seem to perform better in ROC than in the PR curve. This difference exists because the number of legitimate transactions greatly exceeds the number of fraud transactions in this domain. Consequently, a large change in the number of false positives can lead to a small change in the false positive rate used in ROC analysis. Precision, on the other hand, by comparing false

positives to true positives rather than true negatives, captures the effect of the large number of negative examples on the techniques performance [16].

Such a difference can lead to wrong conclusions. In case to evaluate using only ROC curves for undersampling technique, it is seen that reducing the examples of the majority class to the size of the minority class leads to a better performance of the model. But according to Section 3, if the kind of the transformation is made the maximum recall will be really obtained, at the same time a low precision (20 false positives occur for each fraud) will be received. PR curves reflect the real picture: 20% of the majority class leads to the maximum possible values of recall and precision.

A similar situation is observed for the weighing technique. Using ROC curves, we can see that the maximum efficiency is achieved with weights of 500-1000, whereas PR curves show the maximum efficiency for 5-10. Compared these values with the obtained values of precision and recall, it is valid at weights 5-10 that precision and recall have the most effective values.

Precision is directly influenced by class imbalance since FP is affected, whereas TPR only depends on positives. That is why ROC curves do not capture such effects. Therefore, for cases where both precision and recall are important, in skewed data, such as credit card fraud detection, PR curves have a greater advantage over ROC curves.

5 CONCLUSIONS

As a result of this research it is determined that precision-recall curves have more advantages than ROC curves in dealing with credit card fraud detection. Area under precision-recall curve correlates with actual values precision and recall for both algorithm-level and data-level techniques. Since the credit card fraud is an unbalanced task, where the ratio of classes is less than 1%, ROC curve doesn't capture the effect of improving both precision and recall metrics.

Future research will be concentrated on improving machine learning methods to detecting credit card fraud using precision-recall curve as a main metric to evaluating performance of classifiers.

REFERENCES

[1] A. C. Bahnsen, D. Aouada, A. Stojanovic, and B. Ottersten, “Feature engineering strategies for credit card fraud detection,” Expert Systems with Applications, vol. 51, pp. 134–142, 2016.

[2] Lu Q., Ju C., “Research on credit card fraud detection model based on class weighted support vector machine,” Journal of Convergence Information Technology, vol. 6, no. 1, Jan. 2011.

[3] Monard, M.C., Batista, G.E., “Learning with skewed class distributions,” Advances in Logic, Artificial Intelligence and Robotics (LAPTEC'02), pp. 173-180, 2002.

[4] Zhou, Zhi-Hua, and Xu-Ying Liu, " On multi-class cost-sensitive learning," Computational Intelligence, vol. 26, no. 3, pp. 232-257, 2010.

[5] Dal Pozzolo, Andrea, et al, "Learned lessons in credit card fraud detection from a practitioner perspective," Expert systems with applications, vol .41, no. 10, pp. 4915-4928, 2014.

[6] Anis, Maira, Mohsin Ali, and Amit Yadav, "A comparative study of decision tree algorithms for class imbalanced learning in credit card fraud detection," International Journal of Economics, Commerce and Management, vol. 3, no. 12, 2015.

[7] Keilwagen, Jens, Ivo Grosse, and Jan Grau, "Area under precision-recall curves for weighted and unweighted data," PLoS One, vol. 9, no. 3, 2014.

[8] Novaković, Jasmina Dj, et al, "Evaluation of Classification Models in Machine Learning," Theory and Applications of Mathematics & Computer Science, vol. 1, no. 1, pp. 39-46, 2017.

[9] Fawcett, Tom, "An introduction to ROC analysis," Pattern recognition letters, vol. 27, no. 8, pp. 861-874, 2006.

[10] Yong, Terence Koon Beh, Chuan Tan Swee, and Theng Yeo Hwee, "Building classification models from imbalanced fraud detection data," Malaysian Journal of Computing, vol. 2, no. 2, pp. 1-21, 2014.

[11] King, Gary, and Langche Zeng, "Logistic regression in rare events data," Political analysis, vol. 9, no. 2, pp. 137-163, 2001.

[12] Fayzrakhmanov, R. A., Kulikov, A.S., et al, "Prediction of the need for narcotic and psychotropic drugs in the region using random forest model," artificial intelligence in solving actual social and economic problems of the XXI century, pp. 108-111, 2016.

[13] Fayzrakhmanov, R. A., et al, "Aplplication of cluster analysis in developing approaches to the selection and designation of treatment regimens for HIV-infected patients," Bulletin of Siberian Medicine, vol. 16, no. 3, pp. 52-60, 2017.

[14] Bhattacharyya, Siddhartha, et al, "Data mining for credit card fraud: A comparative study," Decision Support Systems, vol. 50, no.3. pp. 602-613, 2011.

[15] Lima, Rafael Franca, and Adriano CM Pereira, "Feature Selection Approaches to Fraud Detection in e-Payment Systems," International Conference on Electronic Commerce and Web Technologies, Springer, Cham, 2016.

[16] Davis, Jesse, and Mark Goadrich, "The relationship between Precision-Recall and ROC curves," Proceedings of the 23rd international conference on Machine learning. ACM, 2006.