An Exploration of Classification Model Building:

Predicting Credit Card Fraud

Kayla Strunk

Department of Computer Science, Denver University

COMP4449: Data Science Capstone

Dr. Claudio Delrieux

February 3, 2023

Introduction

Research Question

How does the performance of a logistic regression compare to that of a random forest classifier on an unbalanced dataset using over sampling and under sampling techniques in predicting credit card fraud?

Description of Data set

The dataset for this investigation was obtained from Kaggle and contains 284,807 transactions made by European card holders in 2013. 491 of these transactions were classified as fraudulent accounting for less than 1% of the total transactions. Due to confidentially reasons, 28 of the original feature variables were transformed using principal component analysis (PCA). The amount spent in a transaction and the time in seconds from the first transaction in the dataset remained untransformed. The target variable is a binary variable which indicates if a transaction was non-fraudulent (0) or fraudulent (1).

The distributions of the data were identified through histogram plots (Appendix Figure A1). It was found that, pre-scaling, the amount variable was significantly skewed to the left indicating the presence of outliers. Although outliers can negatively influence a model, they were not removed as in the case of credit card transactions unusual spend amounts can be an important indicator of a suspicious transaction. Another distribution of significance, affecting model performance, is the distribution of the target variable. As previously mentioned, the dataset is highly imbalanced with fraudulent transactions accounting for less than 1% of the total transactions (Appendix Figure A2). To remedy this imbalance in variable distribution, a model was trained on the imbalanced dataset and balanced datasets created through under and over sampling utilizing a resampling technique.

Data Preprocessing

Data Preparation

Minimal data cleaning and preparation was needed due to the PCA pre-transformation of most of the original features. It was confirmed that there were no missing values, so no transactions or feature variables were removed from the dataset due to missingness. Because one of the models of interest is a logistic regression, the amount variable was scaled as to not provide un-do influence on the training of the model given its larger scale. The scaling used was a standard scaler to have the amount variable match the distributions of the other feature variables; after scaling, all the feature variables were roughly normally distributed. The time variable was removed as it did not provide any significant insight into the nature of the transactions themselves and would not aid in fraud detection.

Lastly, the dataset was reduced from 28 principal components to five principal components. Based on an analysis of the variability contained within the principal components, it was found that most of the variability in the dataset could be explained within the first five principal components (Appendix Figure A3). The final dataset to train and test the model was composed of the scaled amount variable, the first five principal components, and the target variable. The base model was trained on 188,261 transactions. The oversampled model dataset was created by resampling the fraudulent transactions in the dataset to match the number of non-fraudulent transactions resulting in a training set of 375,786 transactions. Inversely the under sampled model dataset was created by sampling the non-fraudulent transactions to match the number of fraudulent transactions resulting in a dataset containing 736 transactions. All models were tested using 96,546 transactions.

Data Analysis

Model Building and Comparison

In the initial analysis six models were built to detect credit card fraud: three logistic regression models and three random forest classifiers. The models were originally built using the default parameters. Due to the imbalanced nature of this dataset accuracy was determined to not be a good indicator of model performance. In the case of credit card fraud detection, if the model predicted all the transactions as non-fraudulent, it would be 99.9% accurate most of the time. As a result, the models are evaluated using f1 score and tradeoff between precision and recall. Of the three logistic regression models the model built on the base, imbalanced dataset preformed the best with an f1 score of 0.33 (Appendix Table B1 and Figure B2). The best performing model for the random forest classifier was the model trained on the oversampled dataset; this model had an f1 score of 0.49 (Appendix Table B3 and Figure B4). The random forest classifier trained on the oversampled dataset was found to have performed the best overall with the default parameters.

In terms of credit card fraud detection, it would be more cost effective for the model to classify a transaction to be fraudulent when it is not (false positive) versus classifying a fraudulent transaction as non-fraudulent (false negative). The administrative costs of investigating a false positive are less than what it would cost an institution to reconcile a false negative. Given the cost factor, an additional model was built to give more weight to false positives. The model built was a random forest classifier that was trained on the oversampled dataset. The class weights and max depth parameters were selected to bias the model into predicting more transactions as fraudulent. The f1 score of this model was 0.25 (Appendix Table B5 and Figure B6); this low f1 score can be the result of sacrificing the precision of the model to increase recall. The biased model had the least number of false negatives predicted (0.05%).

Finally, a root cause analysis was conducted on the unbiased oversampled random forest classification model to investigate what may be causing the model to mis-classify transactions. The predictions of the transactions were added alongside their true classification to identify which transactions were mis-classified. It was found that 7 transactions were false positives, and 81 transactions were false negatives. These transactions were plotted using a histogram of their spend amounts (Appendix Figure B7). It was found the transactions that were mis-classified as fraudulent had a higher than average spend amount whereas the inverse is true for the transactions mis-classified as non-fraudulent. The transactions incorrectly classified as non-fraudulent had a lower than average spend amount. Having knowledge of common credit card fraud schemes would help justify why that may be. For instance, some fraudsters may spend a small amount at first to see if they can get away with it; on the other side fraudsters may spend a large amount to get as much as possible before being caught. A model to detect anomalies like extremely small or extremely large transactions may lend better to accurately classifying these transactions.

Conclusion

The oversampled datasets outperformed the under sampled datasets for both the logistic regression and random forest classifier; the random forest classifier outperformed the logistic regression model for all datasets. Overall, it was determined that the best model to be put into production would be the biased random forest regressor. This model would reduce the cost of a fraudulent transaction being mis-classified while at the same time being able to identify more fraudulent transactions correctly.

Finally, the performance of over sampling with the random forest regressor could be further investigated using synthetic minority oversampling technique (SMOTE). The resampling

technique used in this research did not add any additional information when training the model; resampling simply replicated the fraudulent transactions already in the dataset. With SMOTE new information could be fed to the model which may improve its performance as it is seeing more unique fraudulent and non-fraudulent transactions.

Appendix A

Appendix A contains figures relevant to exploratory data analysis

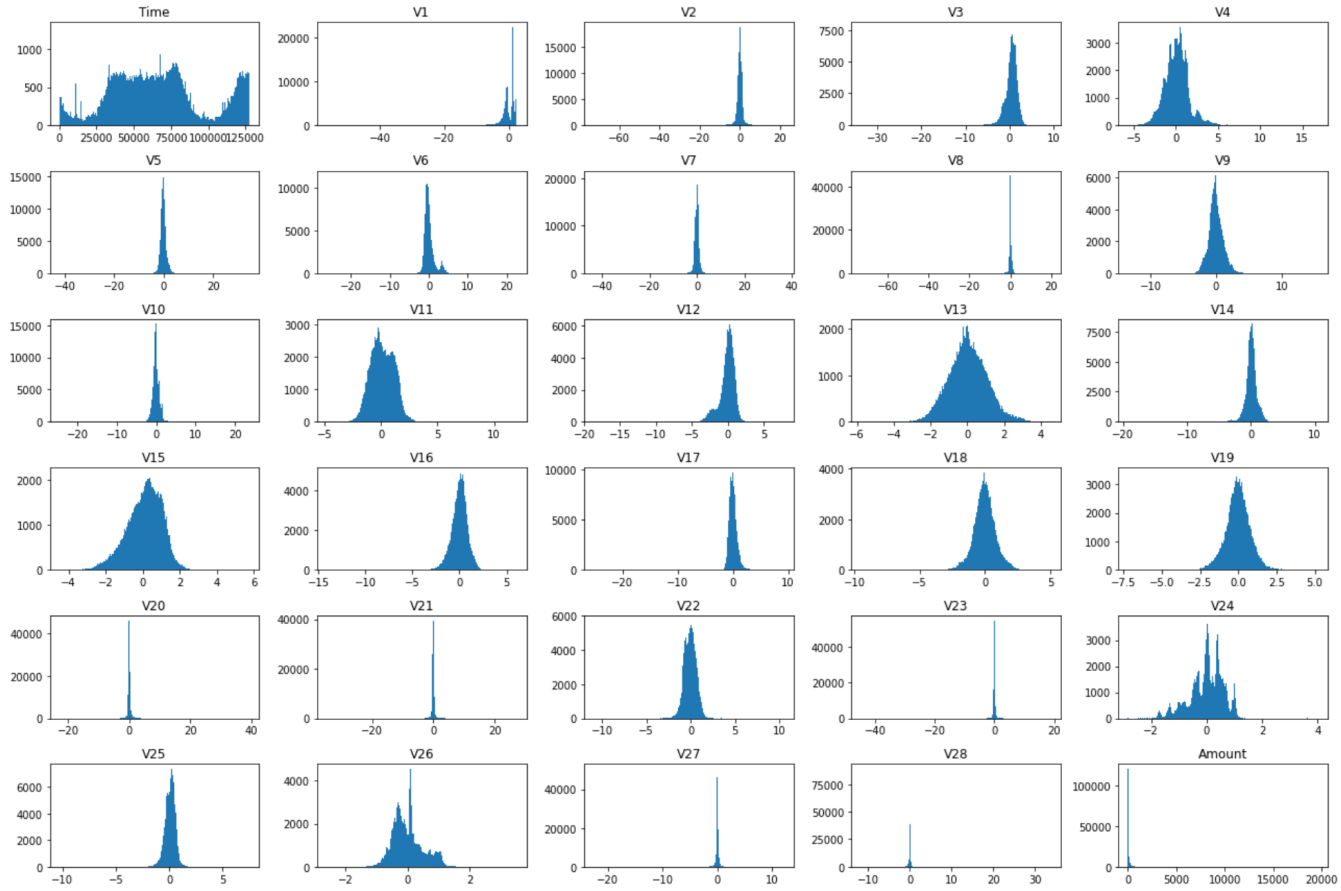

Figure A1. Distributions of the feature variables

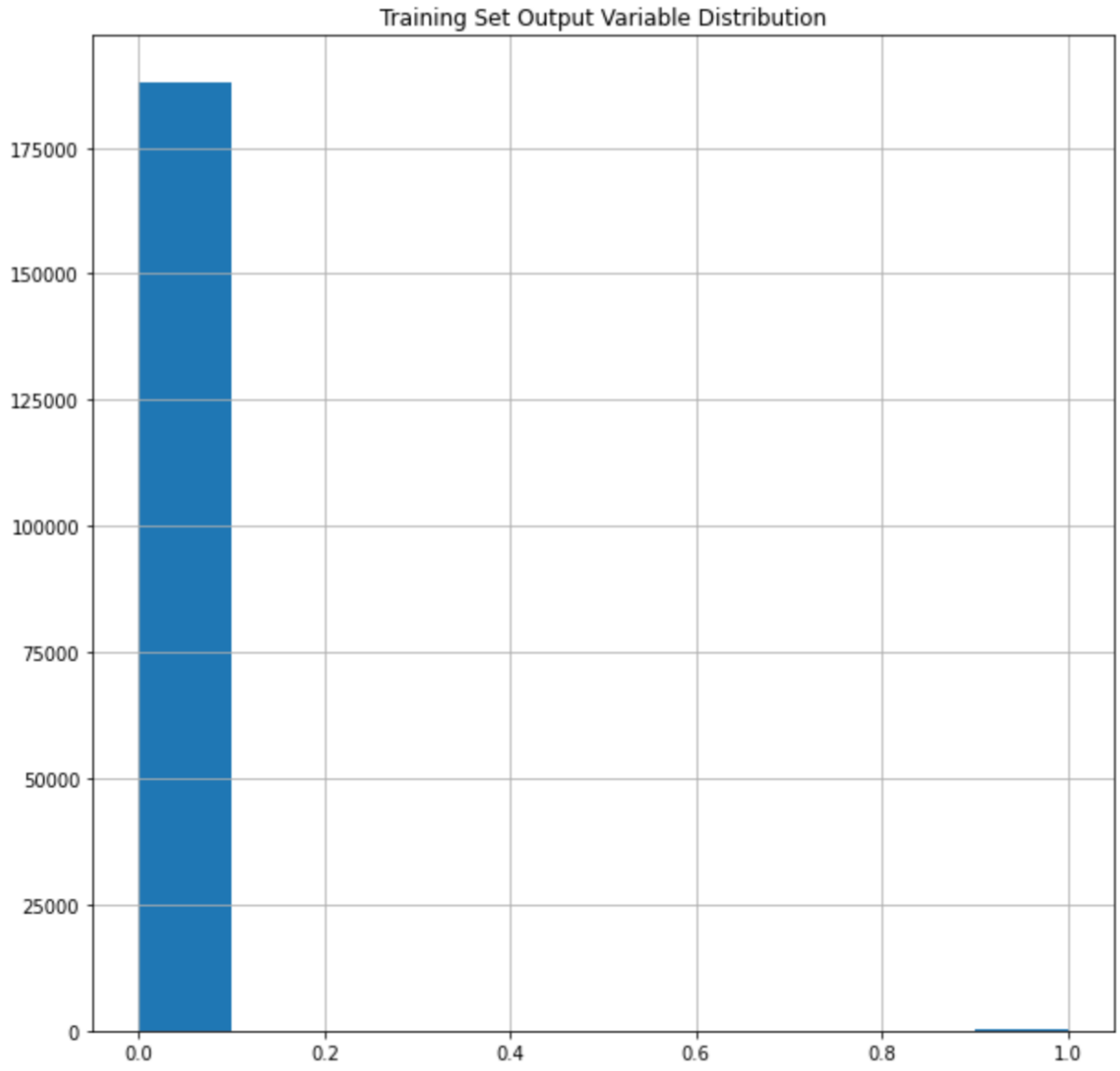

Figure A2. Distribution of the target variable

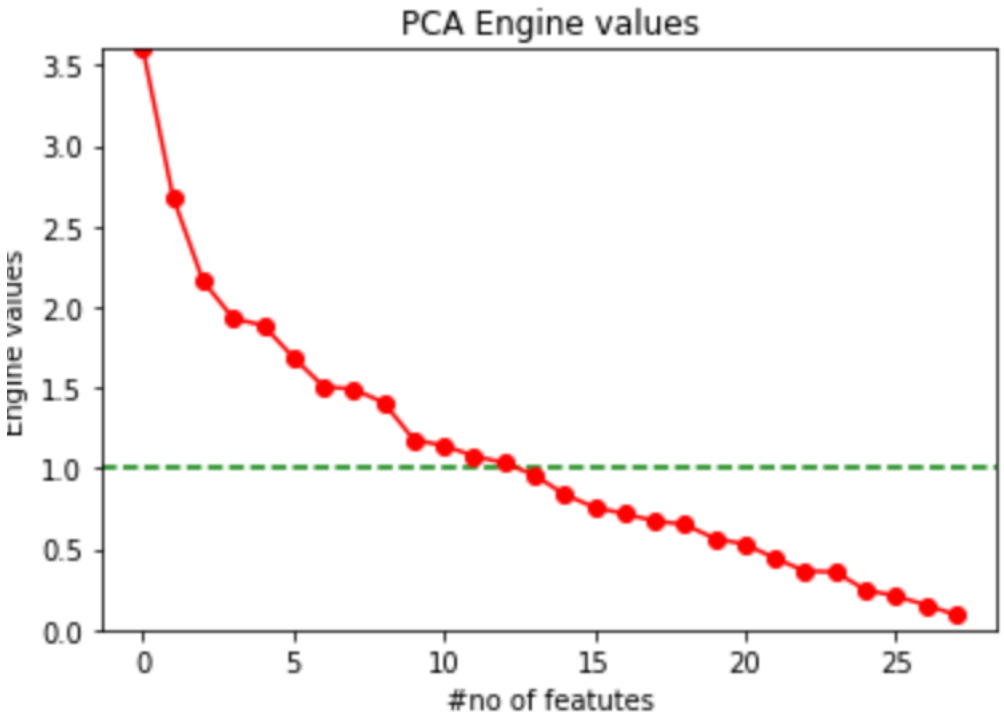

Figure A3. Scree plot to explain variance of dataset contained within each principal component

Appendix B

Appendix B contains figures and tables relevant to the performance of the models discussed in this paper

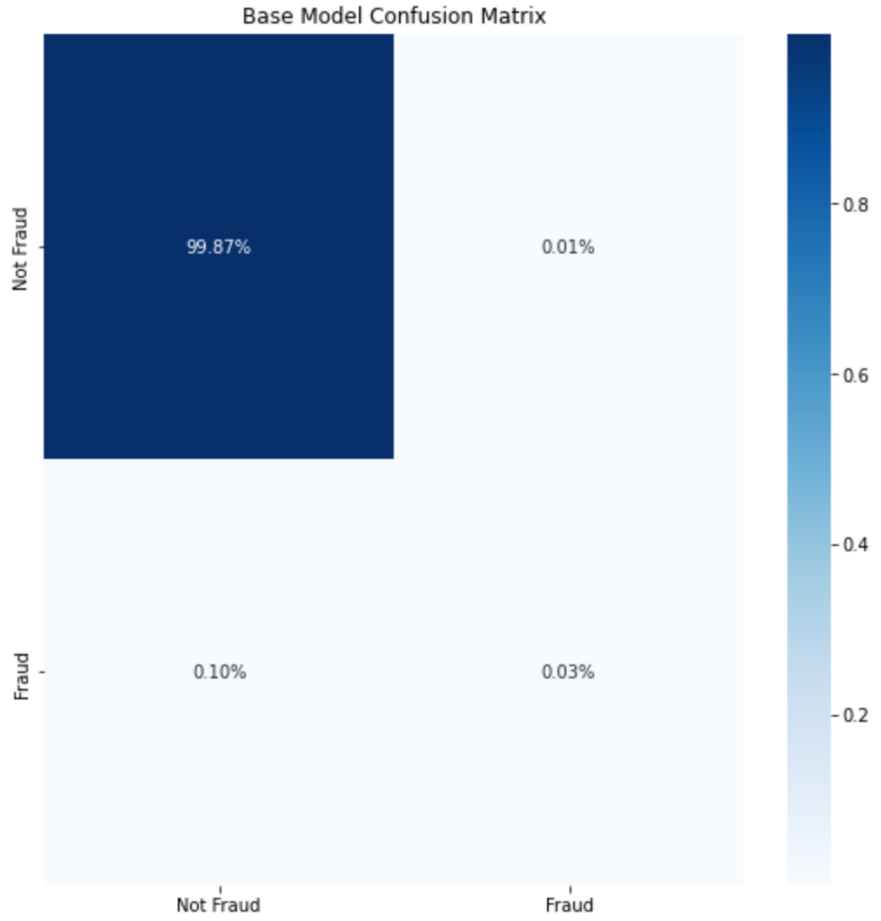

Figure B1. Confusion matrix for the best performing Logistic Regression Model

	precision	recall	f1-score	support
0	1.00	1.00	1.00	96422
1	0.81	0.21	0.33	124
accuracy			1.00	96546
macro avg	0.91	0.60	0.67	96546
weighted avg	1.00	1.00	1.00	96546

Table B2. Classification report for the best performing Logistic Regression Model

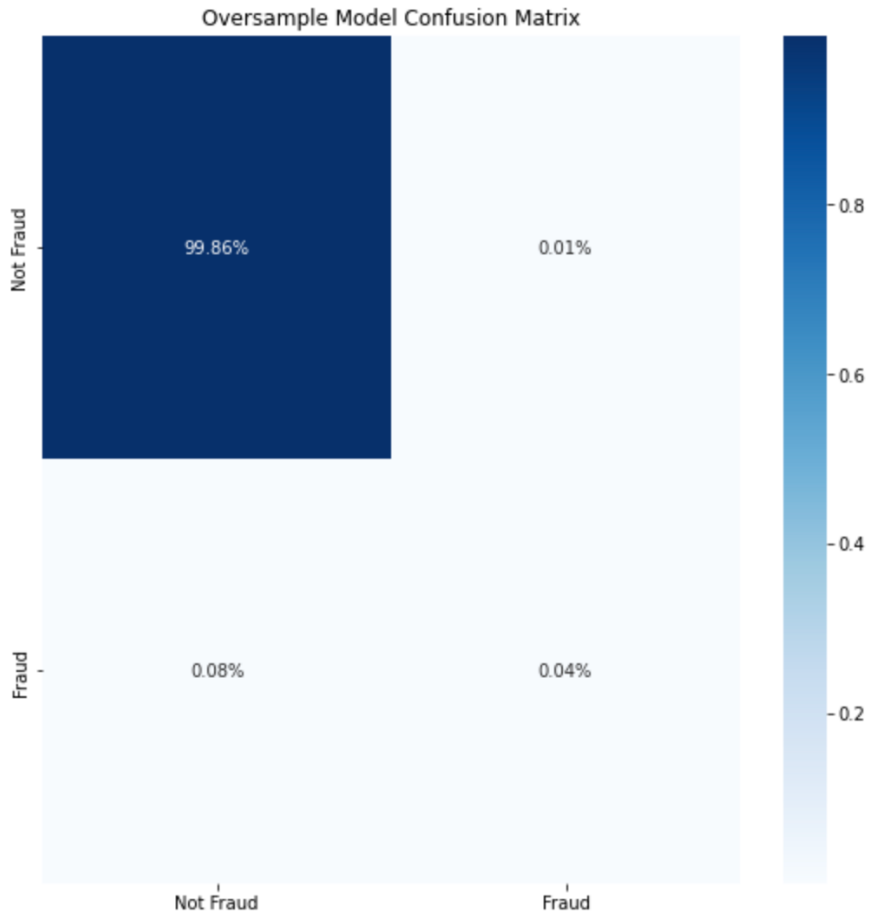

Figure B3. Confusion matrix for the best performing Random Forest Classification Model

	precision	recall	f1-score	support
0	1.00	1.00	1.00	96422
1	0.86	0.35	0.49	124
accuracy			1.00	96546
macro avg	0.93	0.67	0.75	96546
weighted avg	1.00	1.00	1.00	96546

Table B4. Classification report for the best performing Random Forest Classification Model

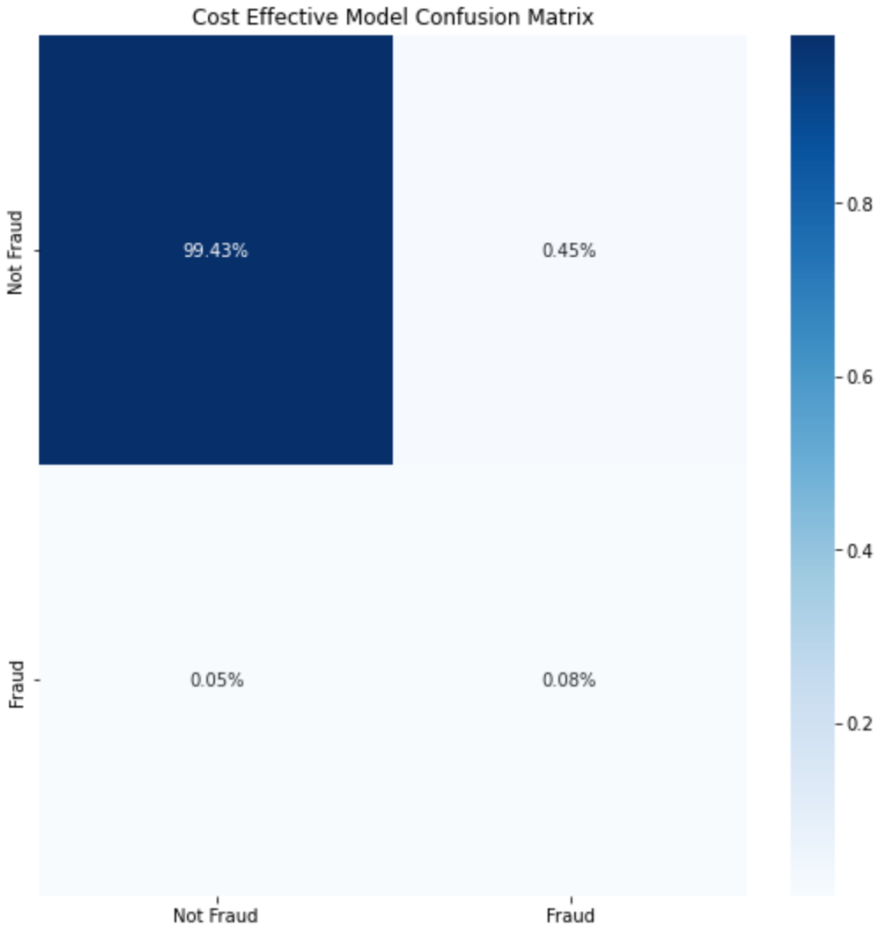

Figure B5. Confusion matrix for the biased model

	precision	recall	f1-score	support
0	1.00	1.00	1.00	96422
1	0.16	0.65	0.25	124
accuracy			1.00	96546
macro avg	0.58	0.82	0.62	96546
weighted avg	1.00	1.00	1.00	96546

Table B6. Classification report for the biased model

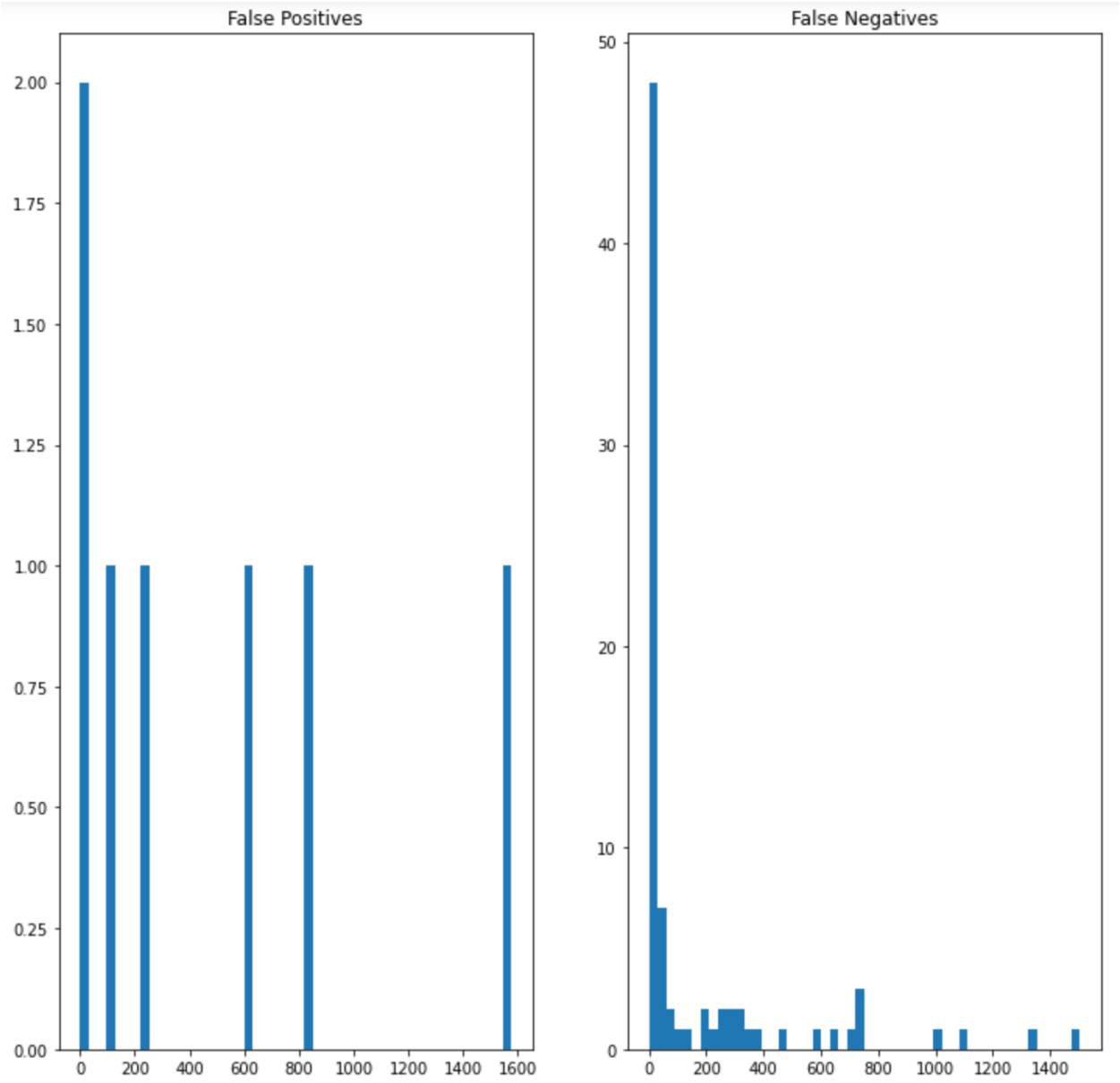

Figure B7. Distributions of the Transactions for False Negatives and False Positives