Fraud Detection in Mobile Money Transactions

Nasim Shahbazian
Capstone Project 1
Springboard Data Science
May 2018

The Problem

- Detecting fraud in mobile money financial transactions.
- Fraud prevention is becoming a critical driver for the financial services industry.
- Financial services institutions are increasingly investing in algorithms and data analytics technology to spot and combat fraud.

The Clients

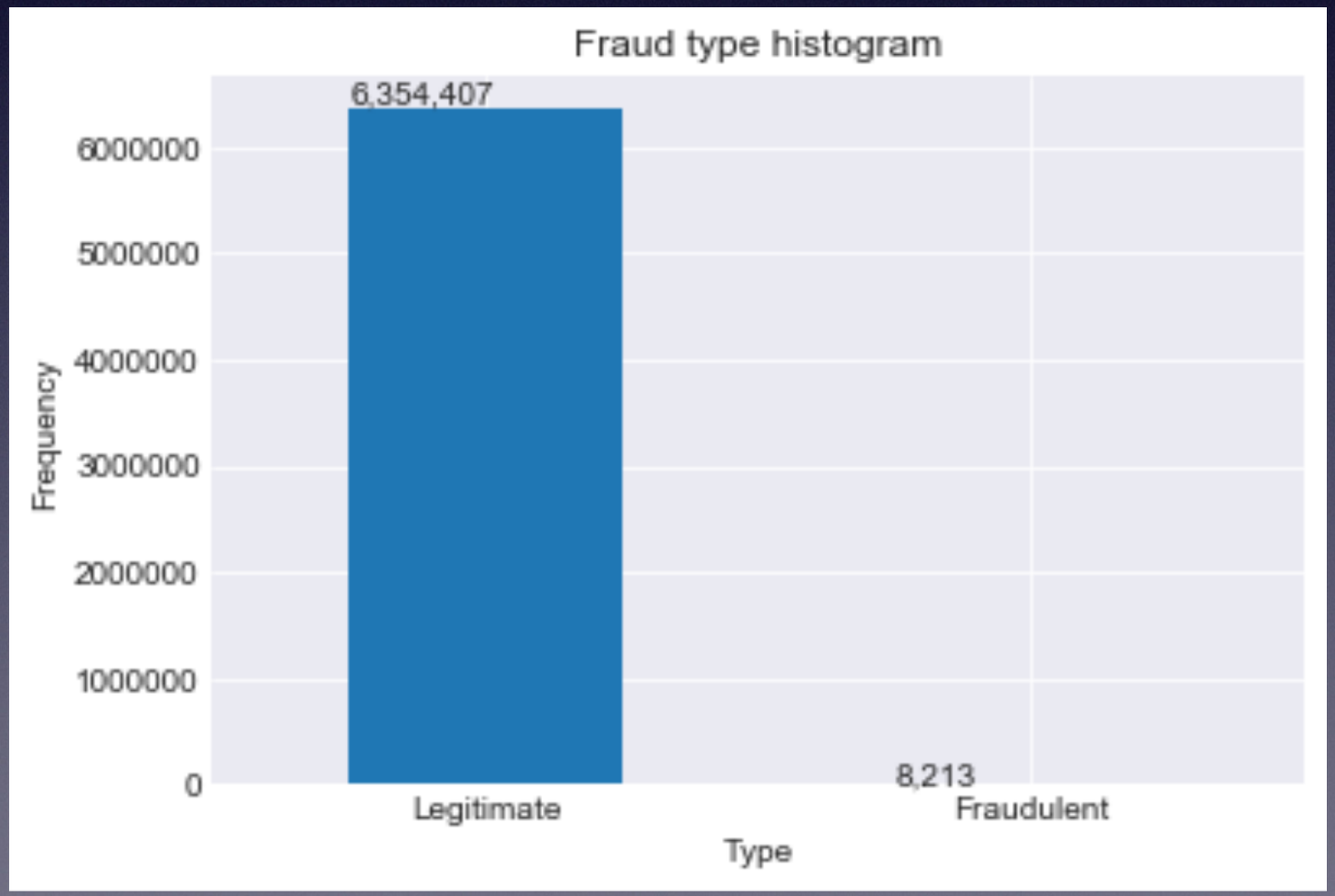
- Banks and financial services that provide mobile money transactions.
- Government
- Corporate sector

Current Dataset

- A synthetic dataset generated using the simulator called PaySim (on Kaggle)
- PaySim uses aggregated data from the private dataset to generate a synthetic dataset for both normal and fraudulent operations

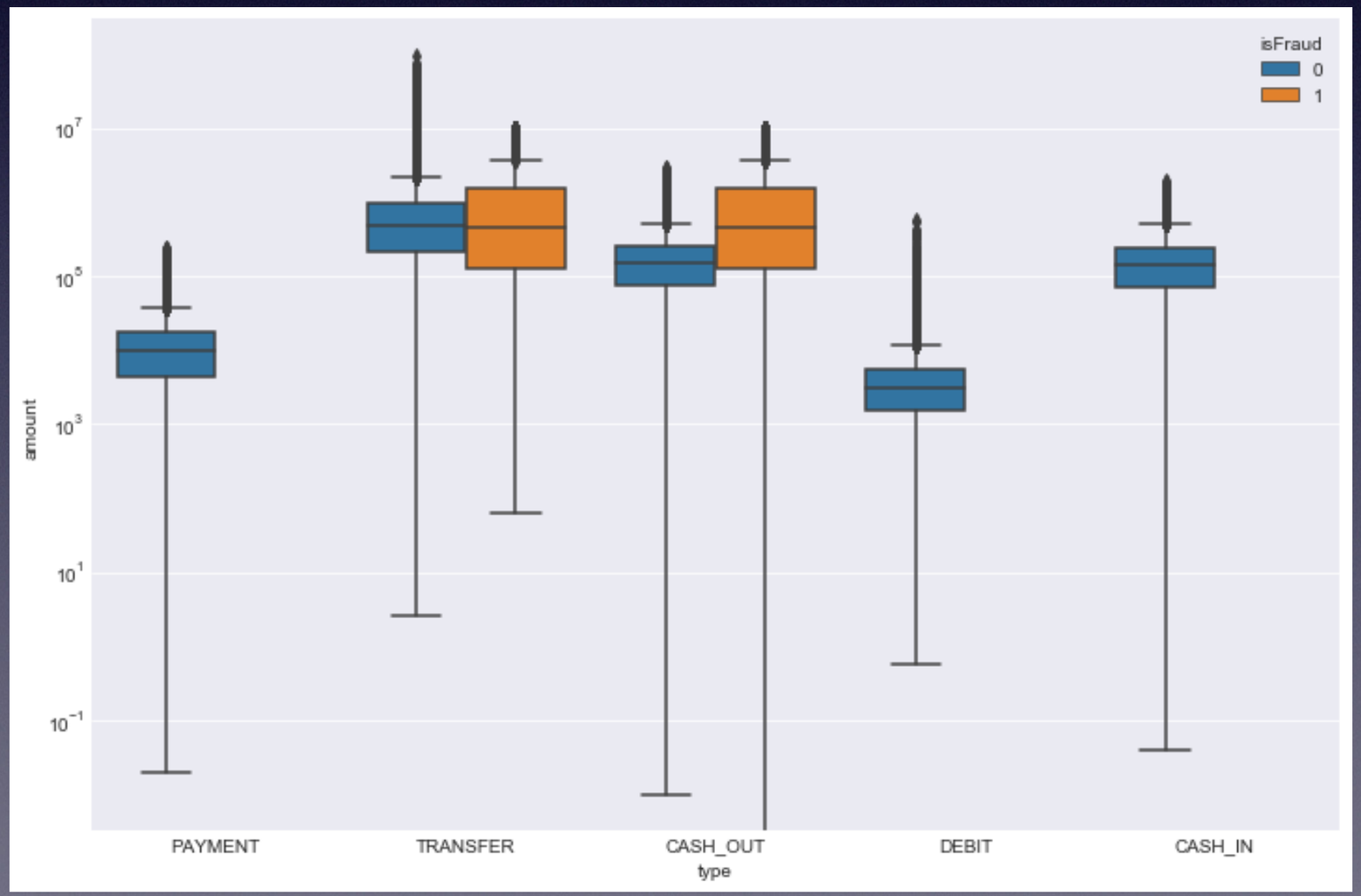
Imbalanced Dataset

The number of fraudulent transactions are much lower than the legitimate ones. The data is highly imbalanced.



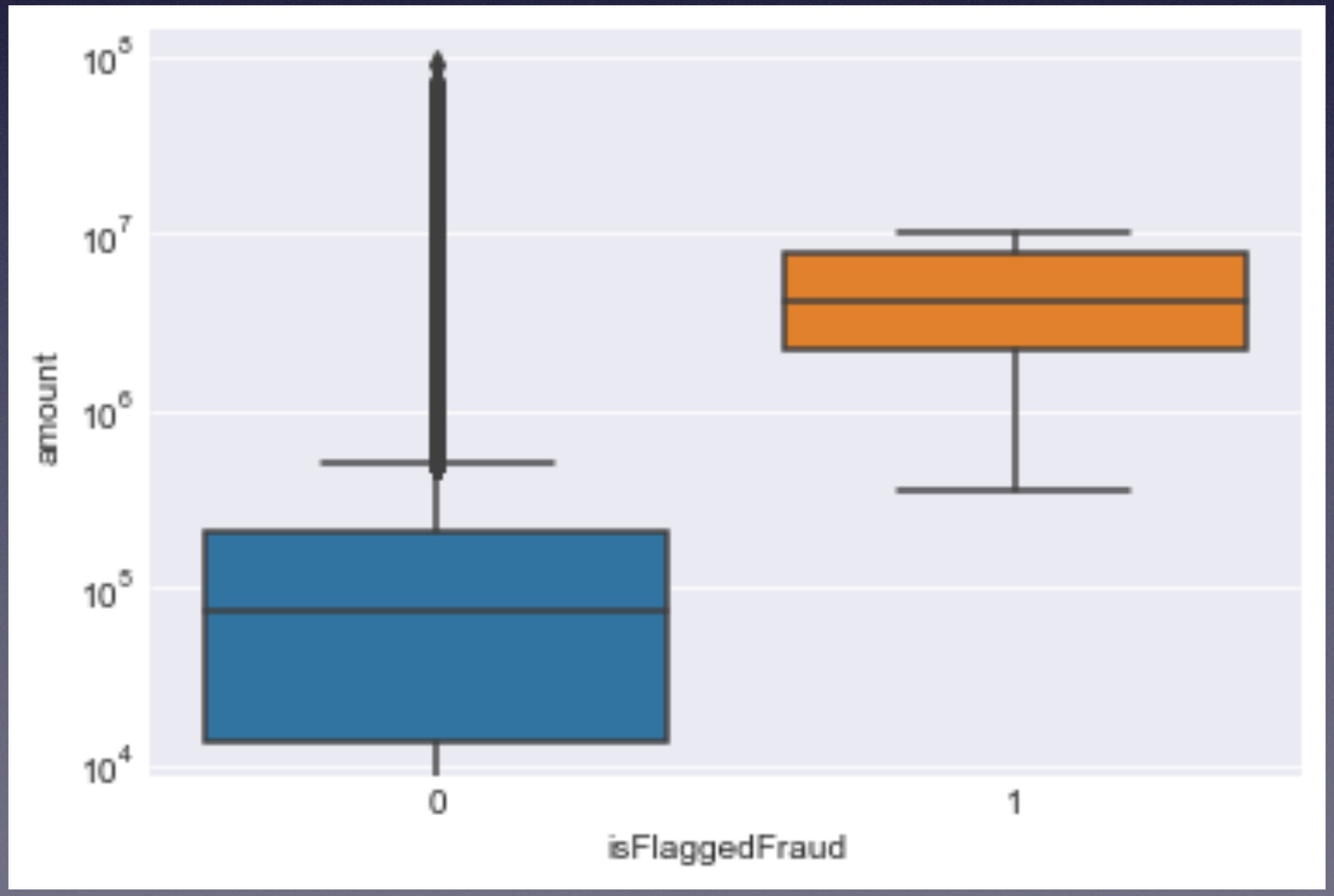
Imbalanced Dataset

Fraud only happens in TRANSFER and CASH-OUT transactions.



Imbalanced Dataset

- Any TRANSFER over 200,000 is not necessarily flagged as isFlaggedFraud=1
- Number of cases where isFraud=1 but isFlaggedFraud=0: 8197
- isFlaggedFraud does not play any significant meaningful role in setting Fraud cases and will be removed.

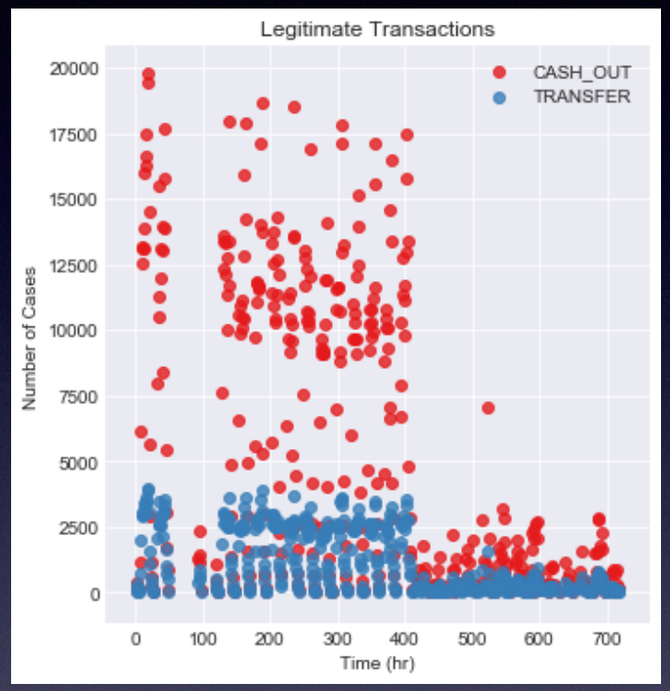


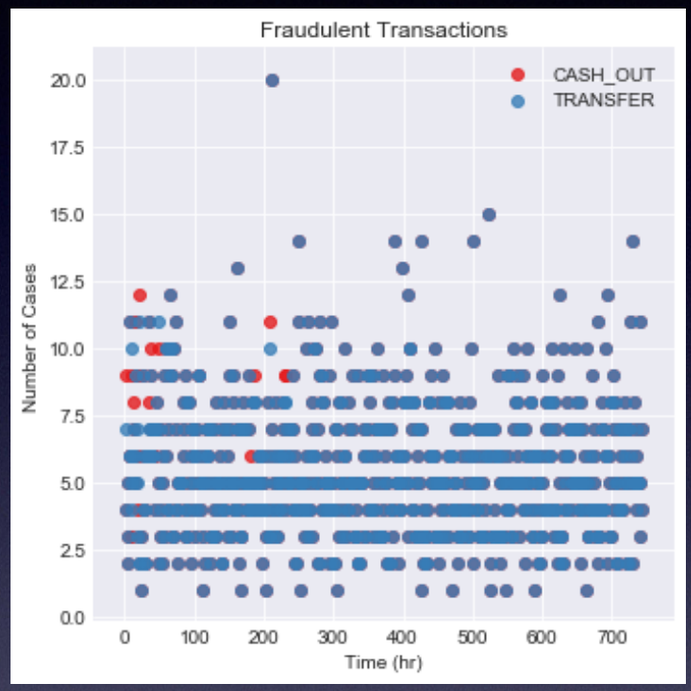
New Features

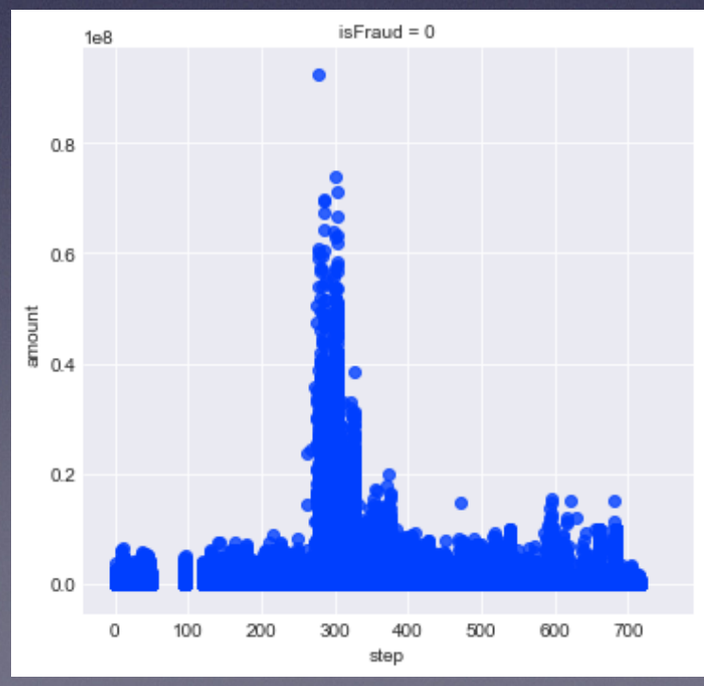
Two new features are defined:

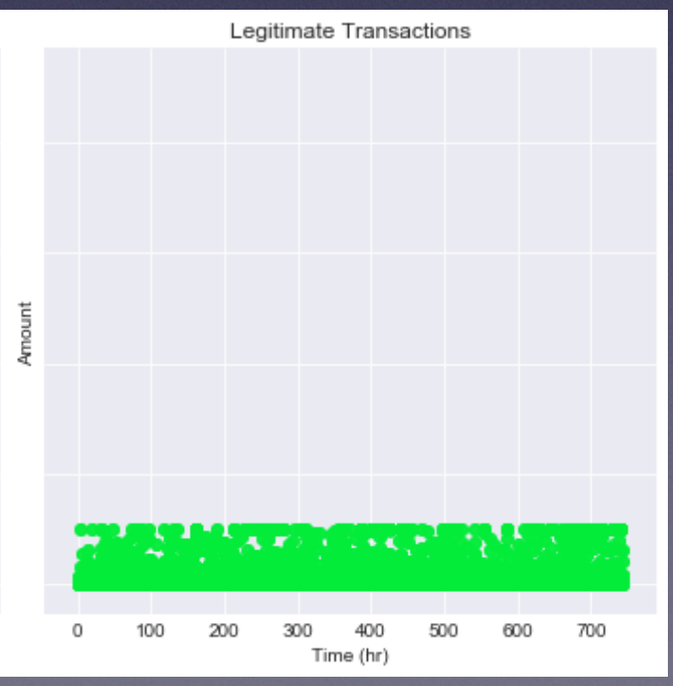
- X[‘errorBalanceOrig’]=X[‘newBalanceOrig']+X['amount']-X['oldBalanceOrig']
- X['errorBalanceDest']=X['newBalanceDest']+X['amount']-X['oldBalanceDest']

Time Variations

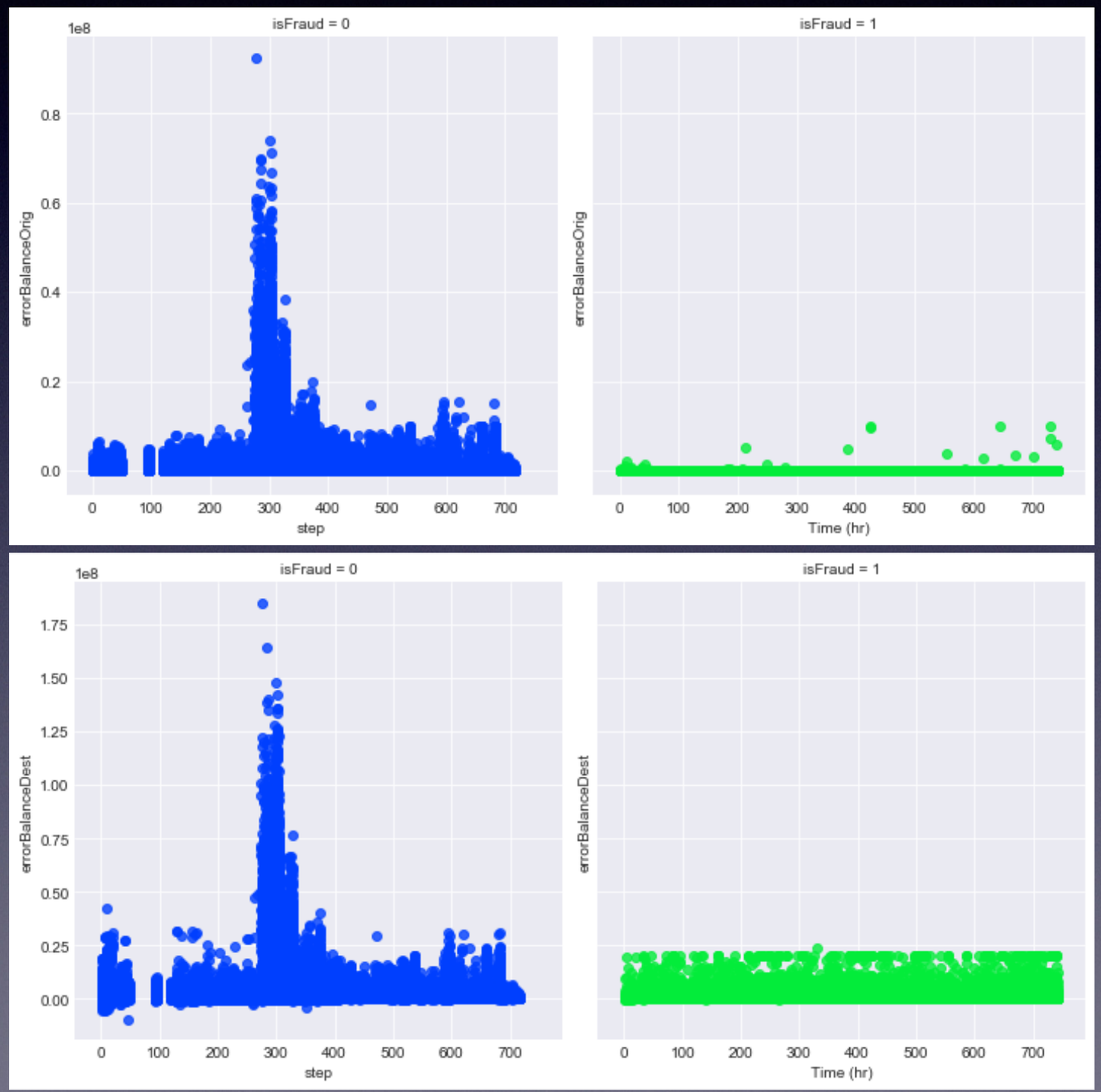




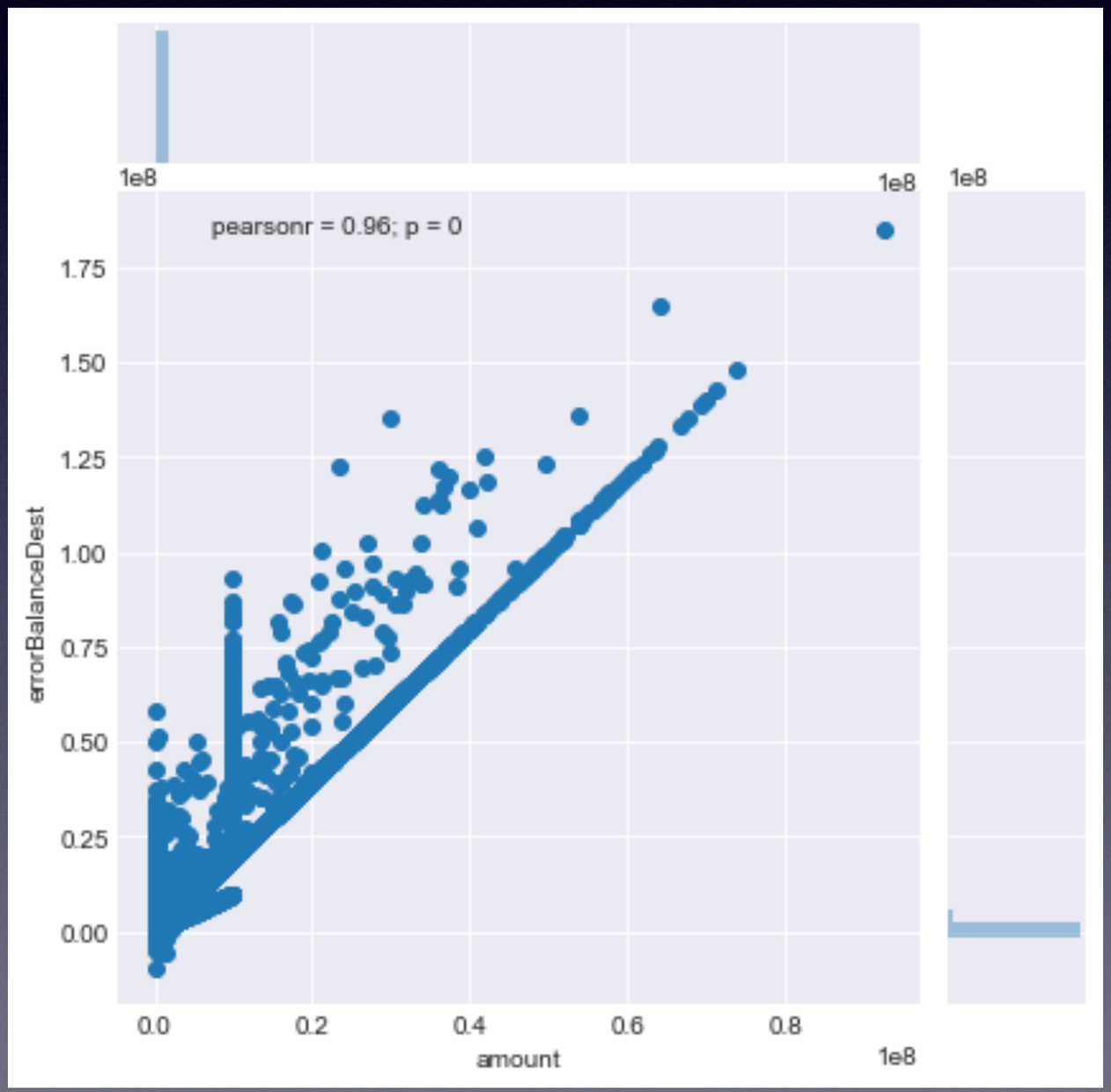




Time Variations



Correlation Plot



Heat Map


Genuine transactions
Fraudulent transactions
amount
oldBalanceOrig
newBalanceOrig
oldBalanceDest
newBalanceDest
errorBalanceOrig
errorBalanceDest
-0.0028
-0.031
0.9
0.31
-0.028
-0.015
0.5
-0.025
-0.02
0.97
1
-0.048
-0.048
0.32
0.5
0.96
-0.0015
-0.027
0.29
0.51
0.96
0.86
0.34
0.77
-0.007
-0.013
-0.016
0.3
0.19
-0.032
0.88
0.086
0.1
0.21
-0.0057
-0.011
0.93
0.73
0.19
0.023
0.44
0.047
1.0
0.5
0.0
−0.5
−1.0

Models Used for Machine Learning

- Random Forest Classifier
- Support Vector Machine (SVM) Classifier
- Logistic Regression Classifier without class_weight
- Logistic Regression Classifier with class_weight

Over-sampling using SMOTE

- Since the current dataset is highly imbalanced, in order to get equal weight on both legitimate and fraudulent transactions, we need to use a method that processes the data to have an approximate 50-50 ratio.
- Over-sampling is performed using Synthetic Minority Over-sampling Technique (SMOTE).

Accuracy, Precision and Recall

- Accuracy = (TP+TN)/total
- Precision = TP/(TP+FP)
- Recall = TP/(TP+FN)

Feature Importances for Random Forest Classifier

	feature_importances_
errorBalanceOrig	0.458438
oldBalanceOrig	0.182455
newBalanceOrig	0.114431
newBalanceDest	0.100220
oldBalanceDest	0.039651
step	0.034901
amount	0.031889
errorBalanceDest	0.023274
type_TRANSFER	0.007971
type_CASH_OUT	0.006769

Algorithms and Hyperparameters

Hyperparameters are tuned using GridSearchCV 5-fold cross validation

Classifier	Hyperparameters
Random Forest	n_estimators = 100
SVM	C = 10
Logistic Regression (no class-weight)	C = 1e-08
Logistic Regression (with class-weight)	C = 0.1

Model Evaluation

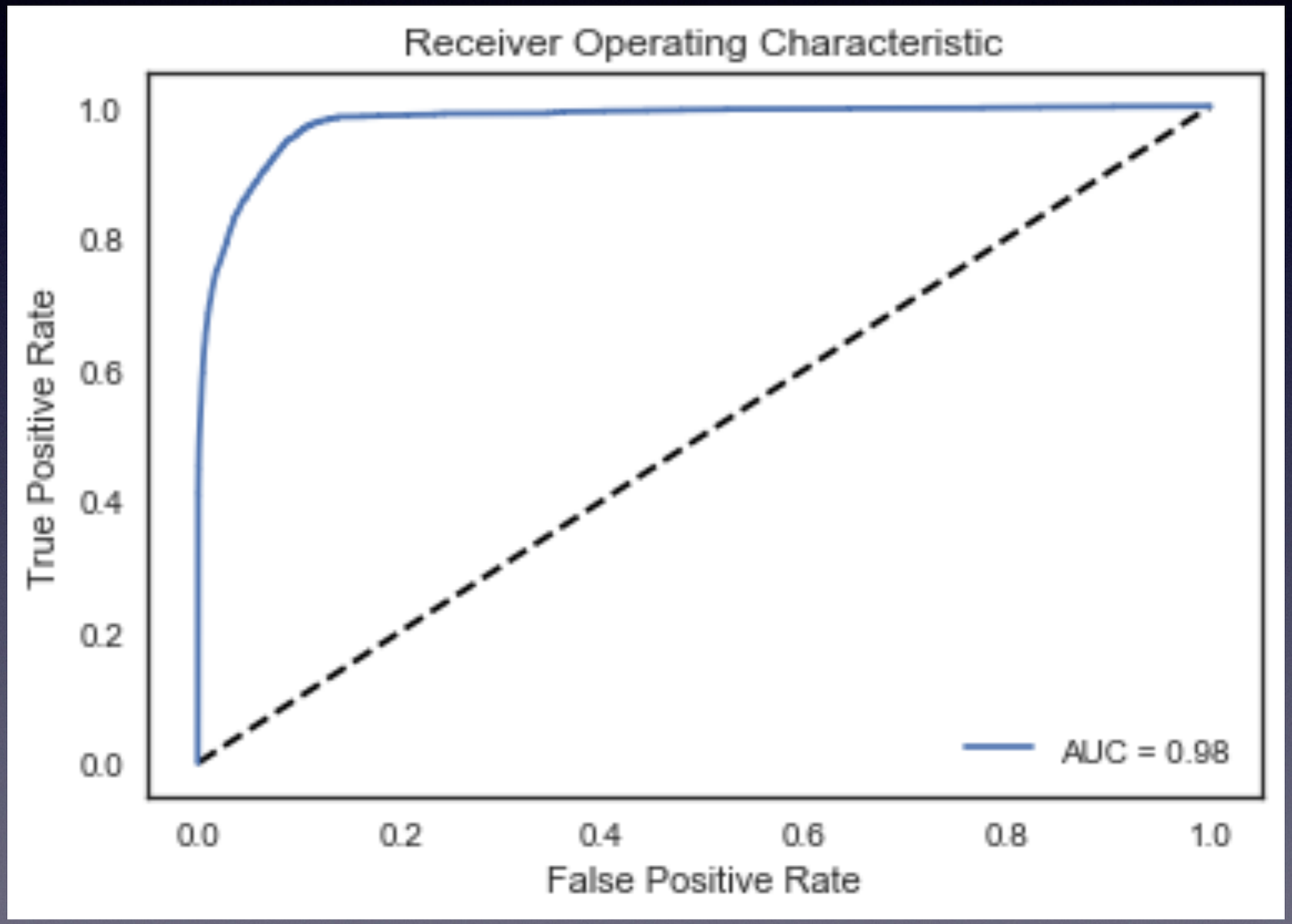
class 0: Legitimate

class 1: Fraudulent

Classifier	Class 0 Precision	Class 1 Precision	Class 0 Recall	Class 1 Recall
Random Forest	1	0.97	1	1
SVM	1	0.03	0.90	0.90
Logistic Regression (no class-weight)	1	0.03	0.89	0.97
Logistic Regression (with class-weight)	1	1	1	0.26

Receiver Operating Characteristic (ROC) Curve

(For Logistic Regression)



Conclusions

- The Paysim synthetic dataset of mobile money transactions has been analyzed.
- To deal with the large skew in the data, we over-sampled the minority class.
- For the current data, Random Forest classifier performs the best. While both SVM and Logistic Regression result in good recall scores, the precision scores for these models are low.
- Using class_weight for linear regression on the original imbalanced data lead to a low recall score.

Acknowledgements

- My mentor, Dipanjan Sarkar
- Springboard
- Edgar Alonso Lopez-Rojas, Ph.D. Thesis (http://urn.kb.se/resolve?urn=urn:nbn:se:bth-12932)
- Kaggle
- Course teaching assistant, Jenny Hung