Fraud Detection Strategy

Mitchell Seiter

Background

The Client:

- Any institution that handles credit card transactions (or any kind of financial transactions) and that would have some liability if fraudulent transactions happened with their service.

The Goal:

- Develop a fraud detection algorithm that ideally catches all cases of fraud without stopping any legitimate transactions from being blocked.

Data

- Data comes from a simulated data generator that mimics real life credit card transactions. For privacy purposes it was not possible to obtain “real “ credit card transaction data.
- Tools and models used were sklearn to generate and test various supervised machine learning models such as KNN, Logistic Regression, Decisions Trees, and Random Forests.
- Primary model we will look at is random forest as this provided the best metrics

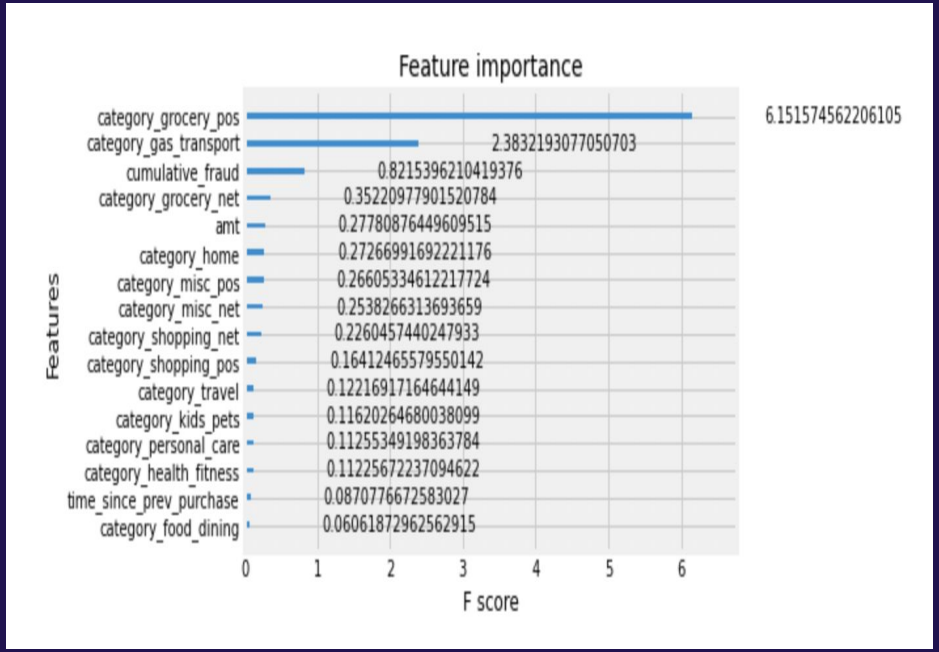
Preliminary Look at the Data

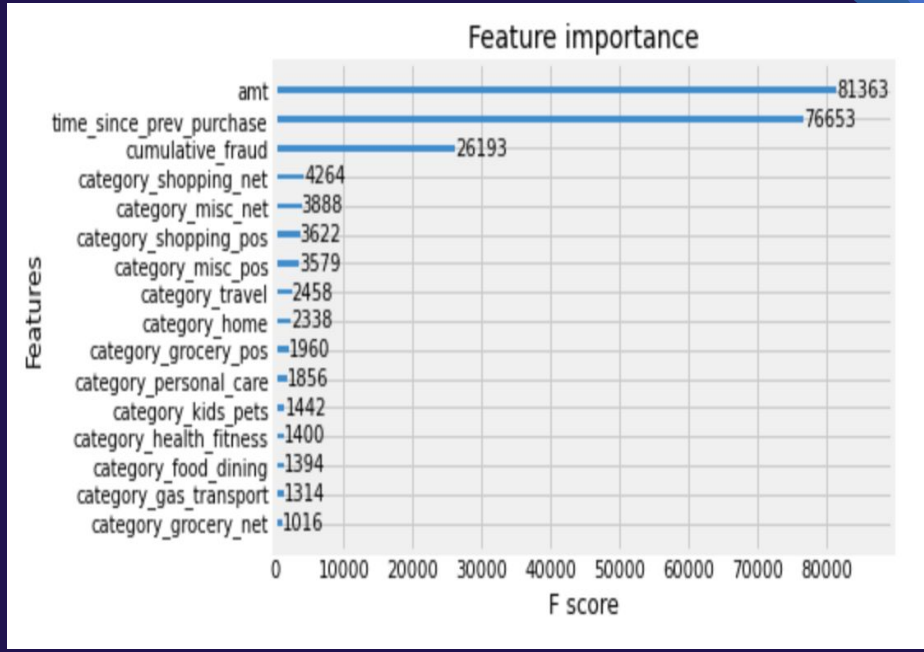
- **555,719** total transaction out of which **2,145** were fraud so ⅓ **of a percent** total fraud
- **218 out of 917** people responsible for all cases of fraud
- Fraudulent transactions have a mean value of **$529 compared to $69** for all transactions

Some Key Features

- Amount of transaction
- Category of merchant for the purchase such as: groceries, gas, entertainment, etc.
- Time since last transaction (added feature)
- Does this person have a history of fraud (added feature)

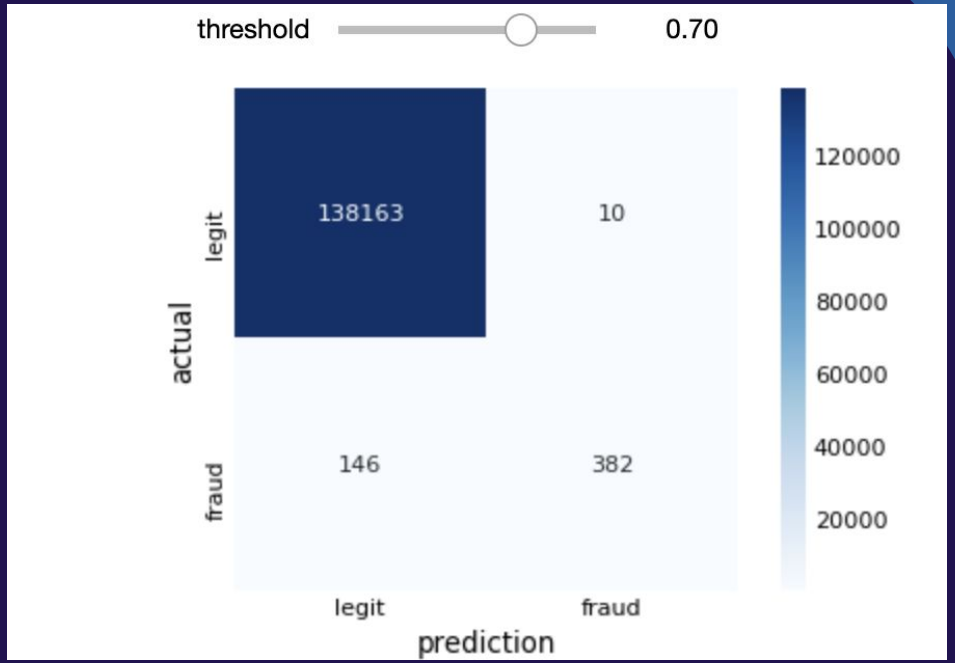
Importance of Features





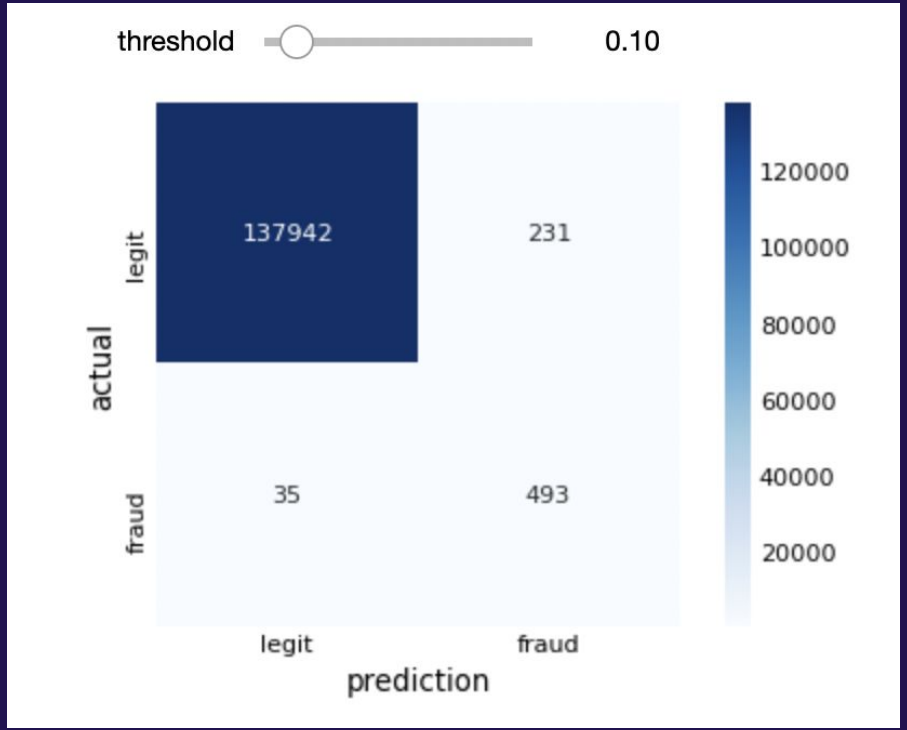
Focus on not alerting customers needlessly

- Set a threshold of 70%
- Catches 72% of all fraud
- Virtually no customers will receive false alerts



Focus on catching as much fraud as possible

- Set a threshold of 10%
- Catches 93% of all fraud
- 23x more customers impacted but still less than 1%



Future Work

Step 1

Re-classify some feature variable

Step 2

Try out more gradient boosting techniques

Step 3

Check if other metrics can build better models

Step 4

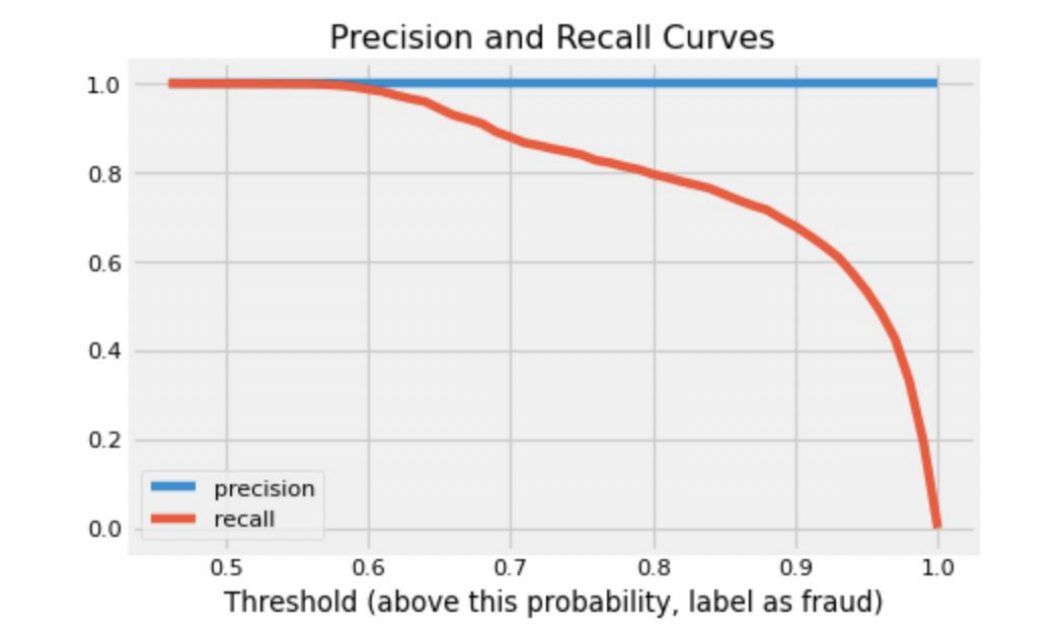
Gather more data and repeat steps 1-3

Conclusions

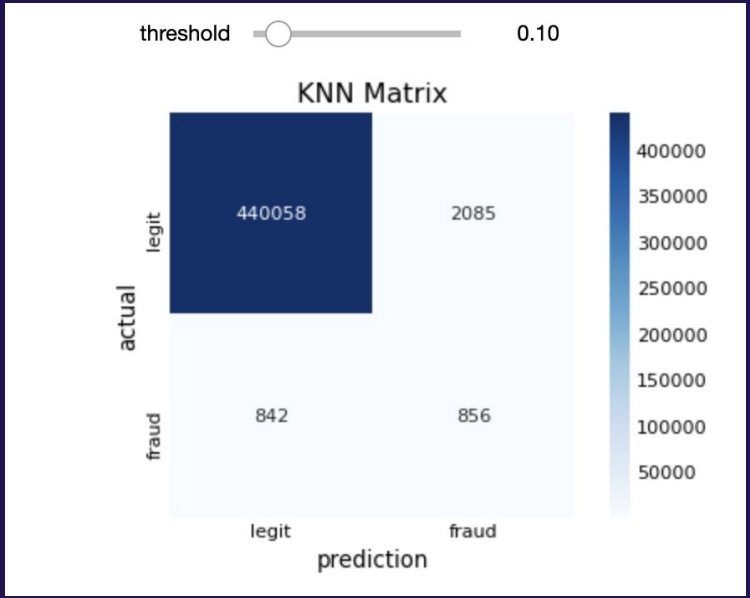
- Random forest provided the best model fit.
- 4 key features with purchase categories having some of the strongest influence but appeared less frequently
- Deciding on business need will help dictate which model to choose, but regardless we can catch a high amount of fraud along with having minimal unnecessary contacts

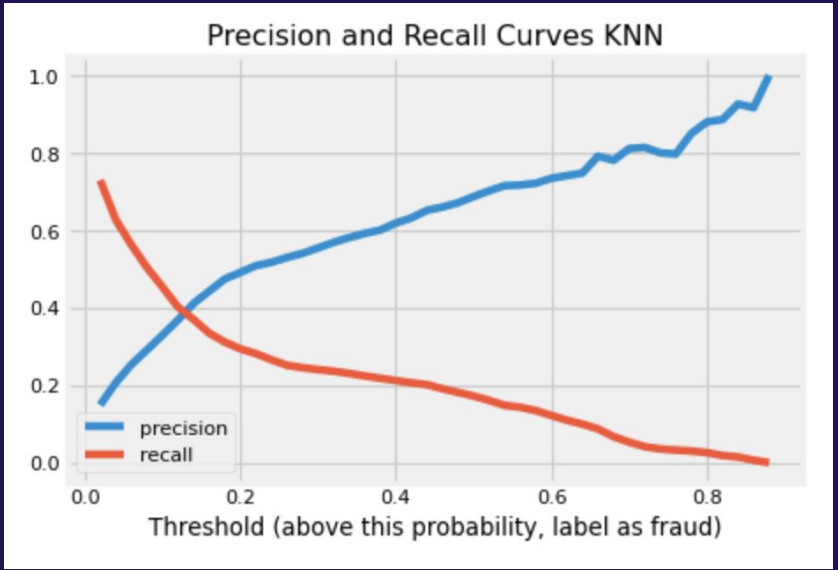
Appendix

Random Forest Curves

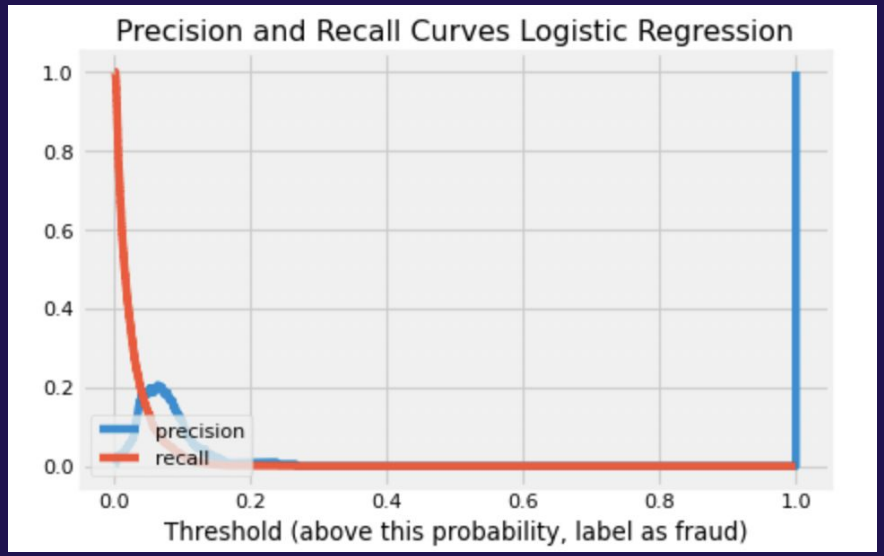


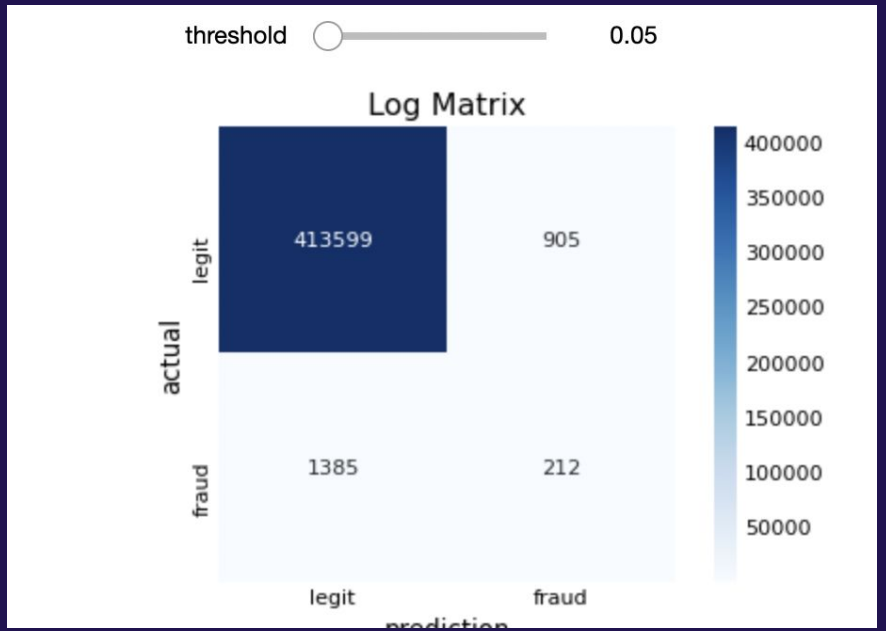
Appendix


threshold
0.10
KNN Matrix
440058
2085
842
856
legit
fraud
actual
legit
fraud
prediction
400000
350000
300000
250000
200000
150000
100000
50000

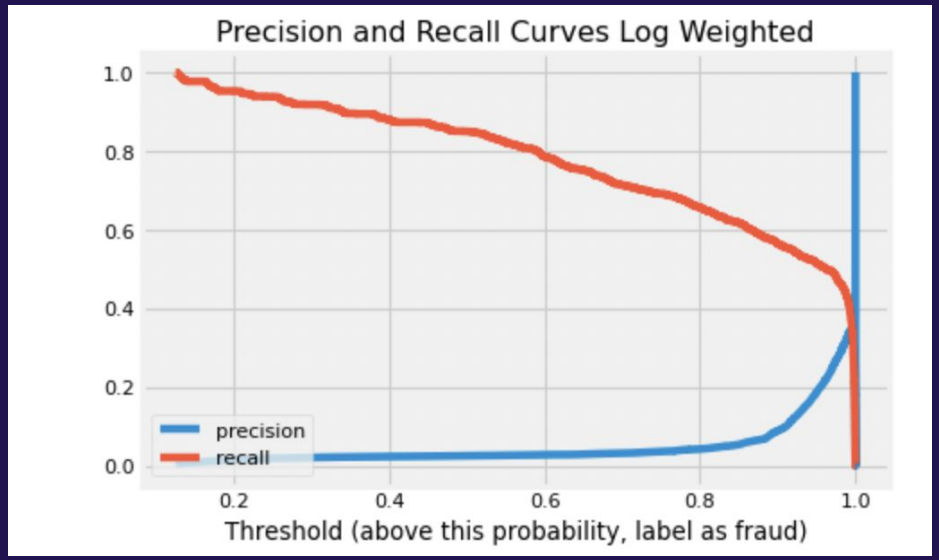

Precision and Recall Curves KNN
1.0
0.8
0.6
0.4
0.2
0.0
precision
recall
0.0
0.2
0.4
0.6
0.8
Threshold (above this probability, label as fraud)

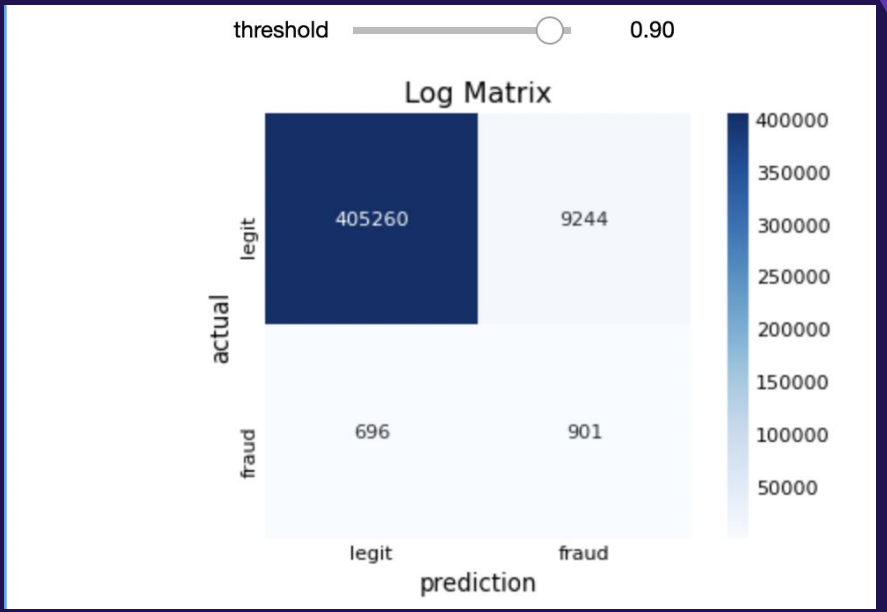
Appendix


Precision and Recall Curves Logistic Regression
1.0
0.8
0.6
0.4
0.2
0.0
precision
recall
0.0
0.2
0.4
0.6
0.8
1.0
Threshold (above this probability, label as fraud)

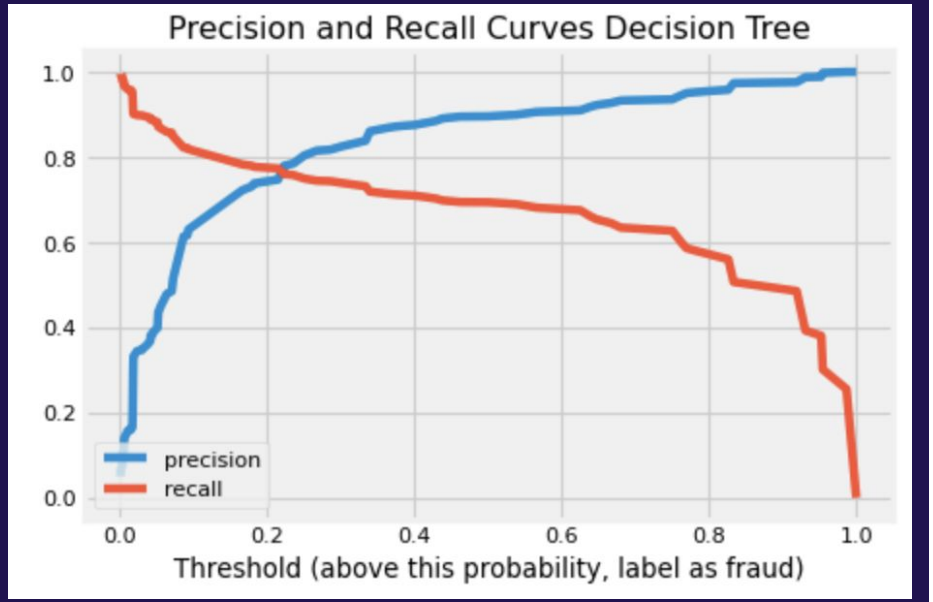

threshold
0.05
Log Matrix
413599
905
1385
212
legit
fraud
actual
legit
fraud
400000
350000
300000
250000
200000
150000
100000
50000

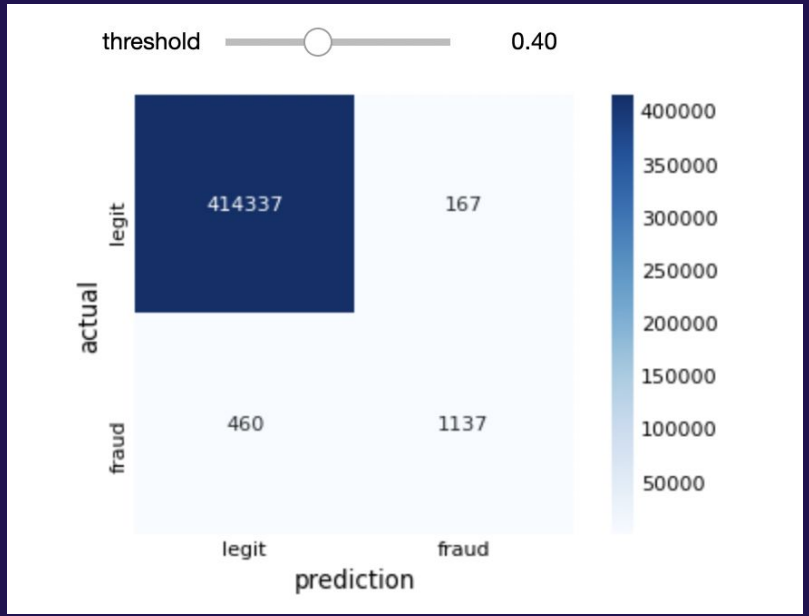
Appendix


Precision and Recall Curves Log Weighted
1.0
0.8
0.6
0.4
0.2
0.0
precision
recall
0.2
0.4
0.6
0.8
1.0
Threshold (above this probability, label as fraud)


threshold
0.90
Log Matrix
405260
9244
696
901
legit
fraud
actual
legit
fraud
prediction
400000
350000
300000
250000
200000
150000
100000
50000

Appendix


Precision and Recall Curves Decision Tree
1.0
0.8
0.6
0.4
0.2
0.0
precision
recall
0.0
0.2
0.4
0.6
0.8
1.0
Threshold (above this probability, label as fraud)


threshold
0.40
414337
167
460
1137
legit
fraud
actual
legit
fraud
prediction
400000
350000
300000
250000
200000
150000
100000
50000