Detecting Fraud with Machine Learning

Presented by:
Tri Le

Why credit card fraud?

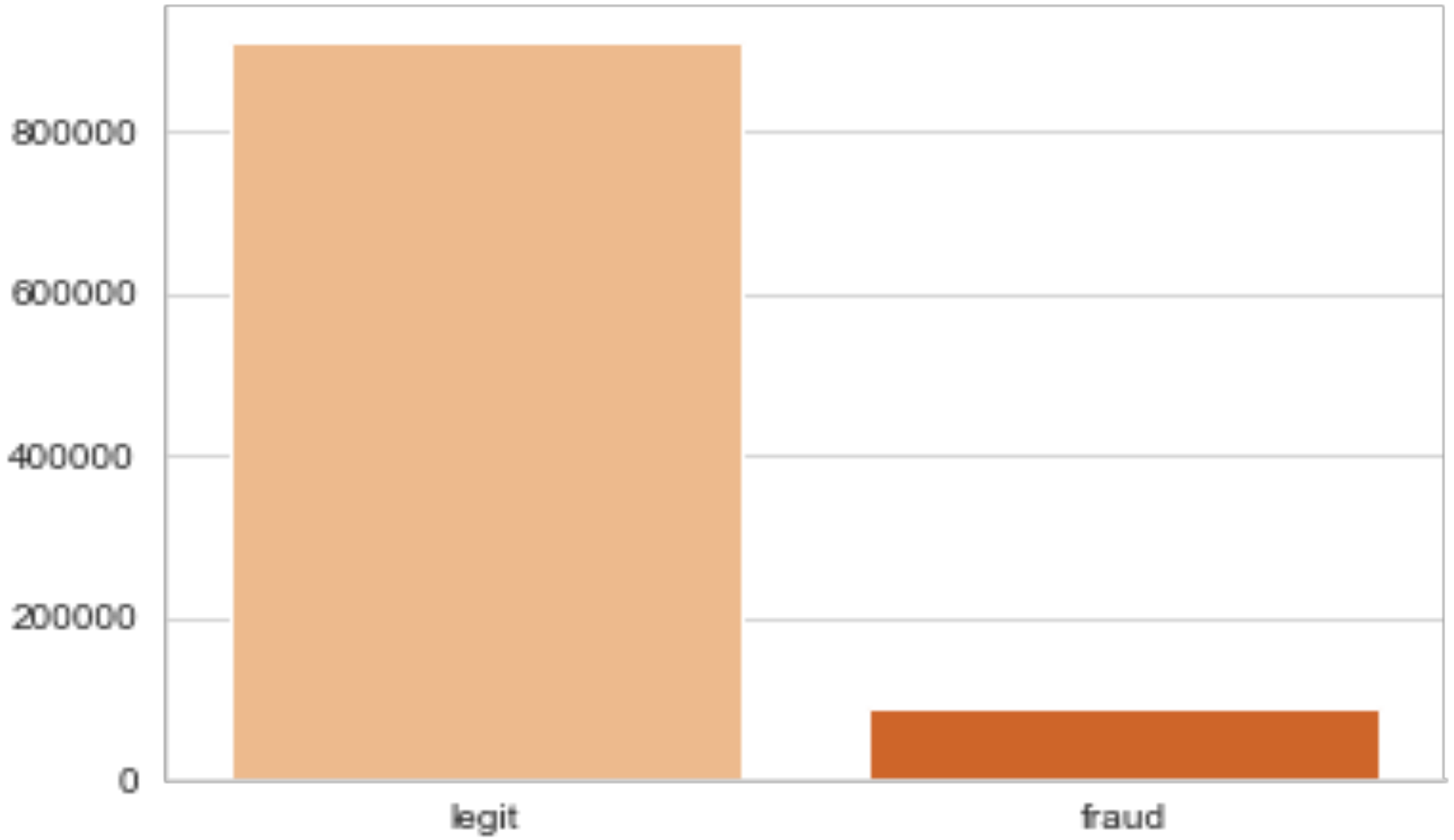
$27.85 billion in fraud loss in 2018
and projected to rise to $40.63 billion in 10 years[1]

U.S. companies incurred 33.99% of total fraud losses worldwide
despite generating only 21.54% of global card volume[1]

35% increase in fraudulent transactions
in April 2020 from the previous year and expected to rise

Sources:
1. https://www.prnewswire.com/news-releases/payment-card-fraud-losses-reach-27-85-billion-300963232.html
2. Fidelity National Information Services Inc.

Transactions: legit or fraud?



The data

- 1 million observable datapoints
- 3 numerical features:
  - Distance from home
  - Distance from last transaction
  - Ratio to median purchase
- 4 categorical features:
  - Repeat retailer
  - Used chip
  - Used pin number
  - Online order

Methodology

Credit Card Transactions

Split Data (Training & Testing)

kNN

Logistic Regression

Decision Tree

Random Forest

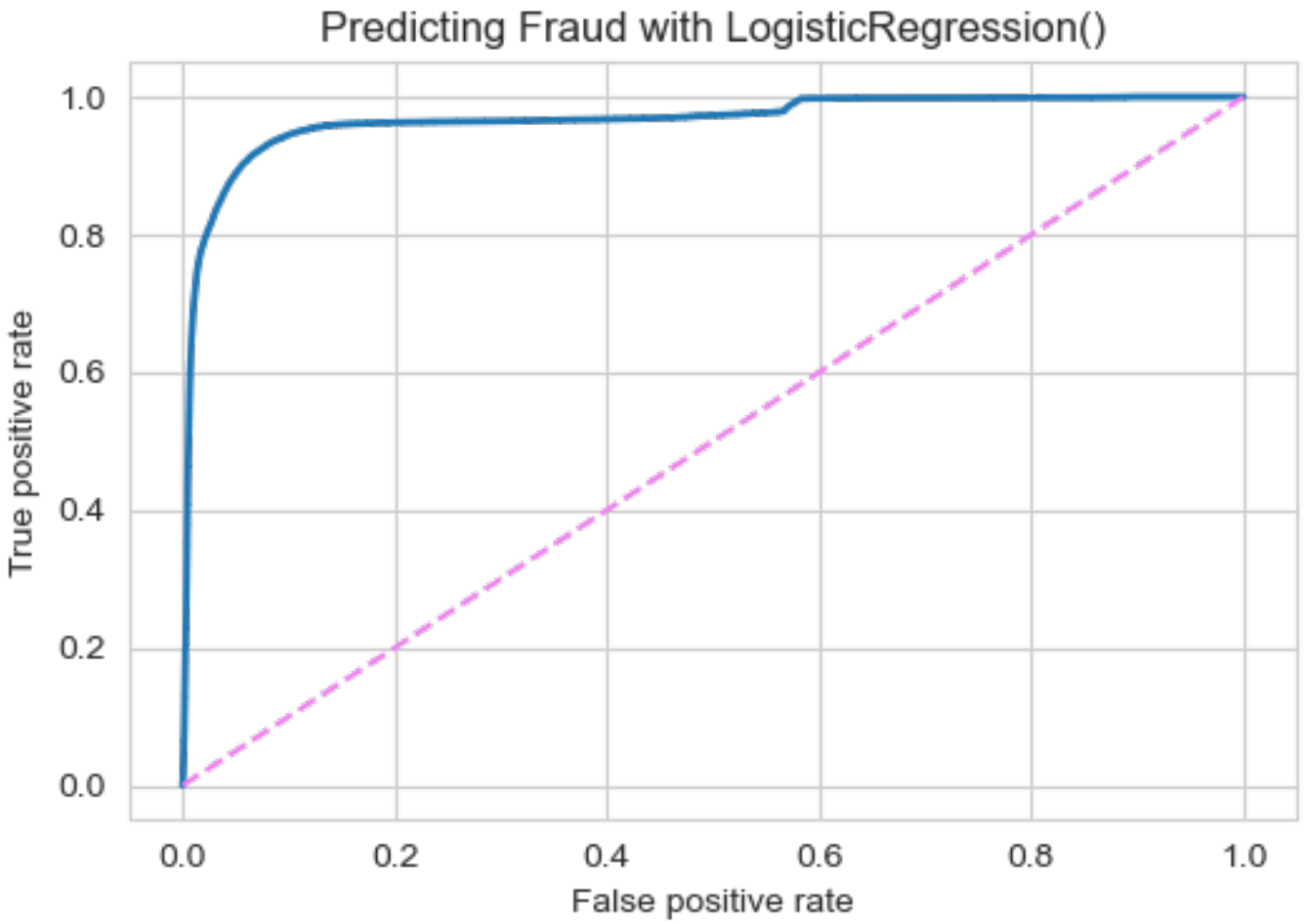
XGBoost

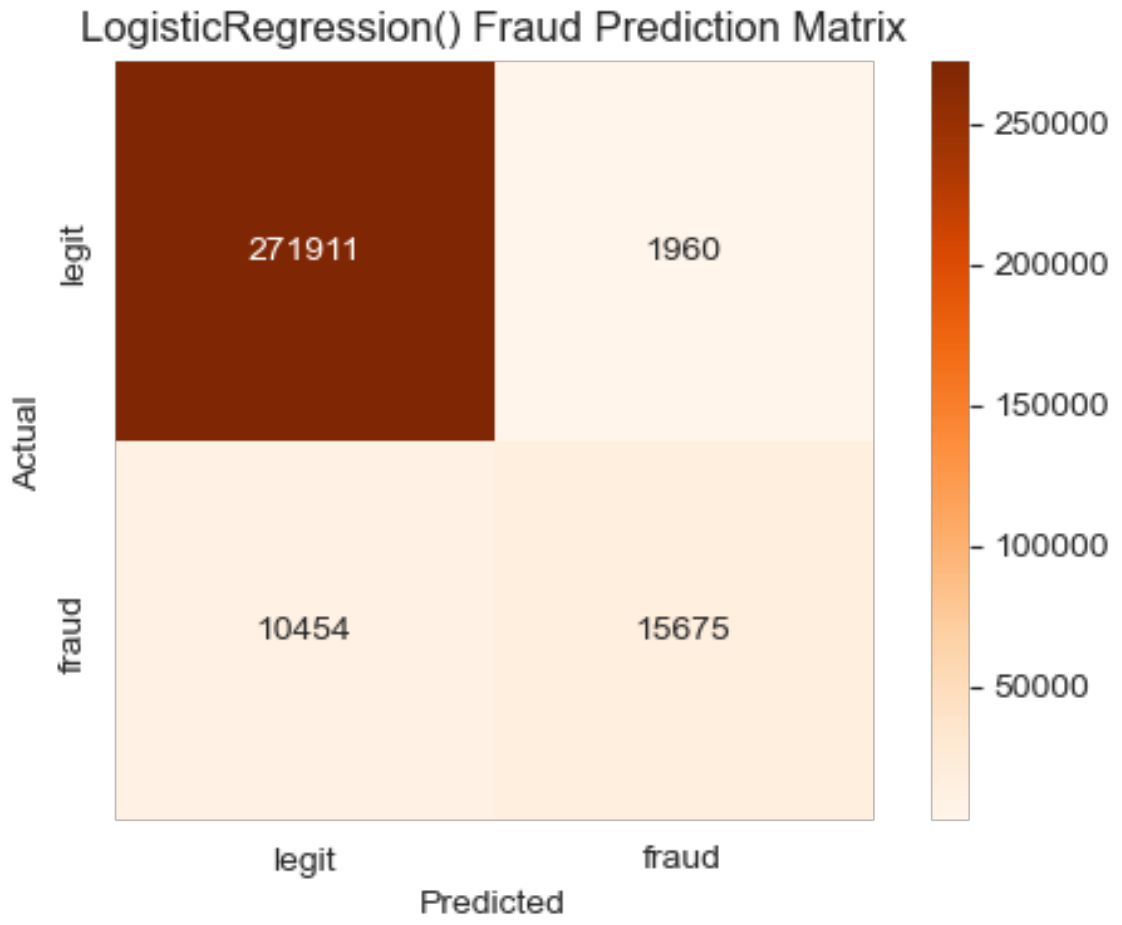
Select Ideal Model

Class Imbalance + Hyperparameter Tuning

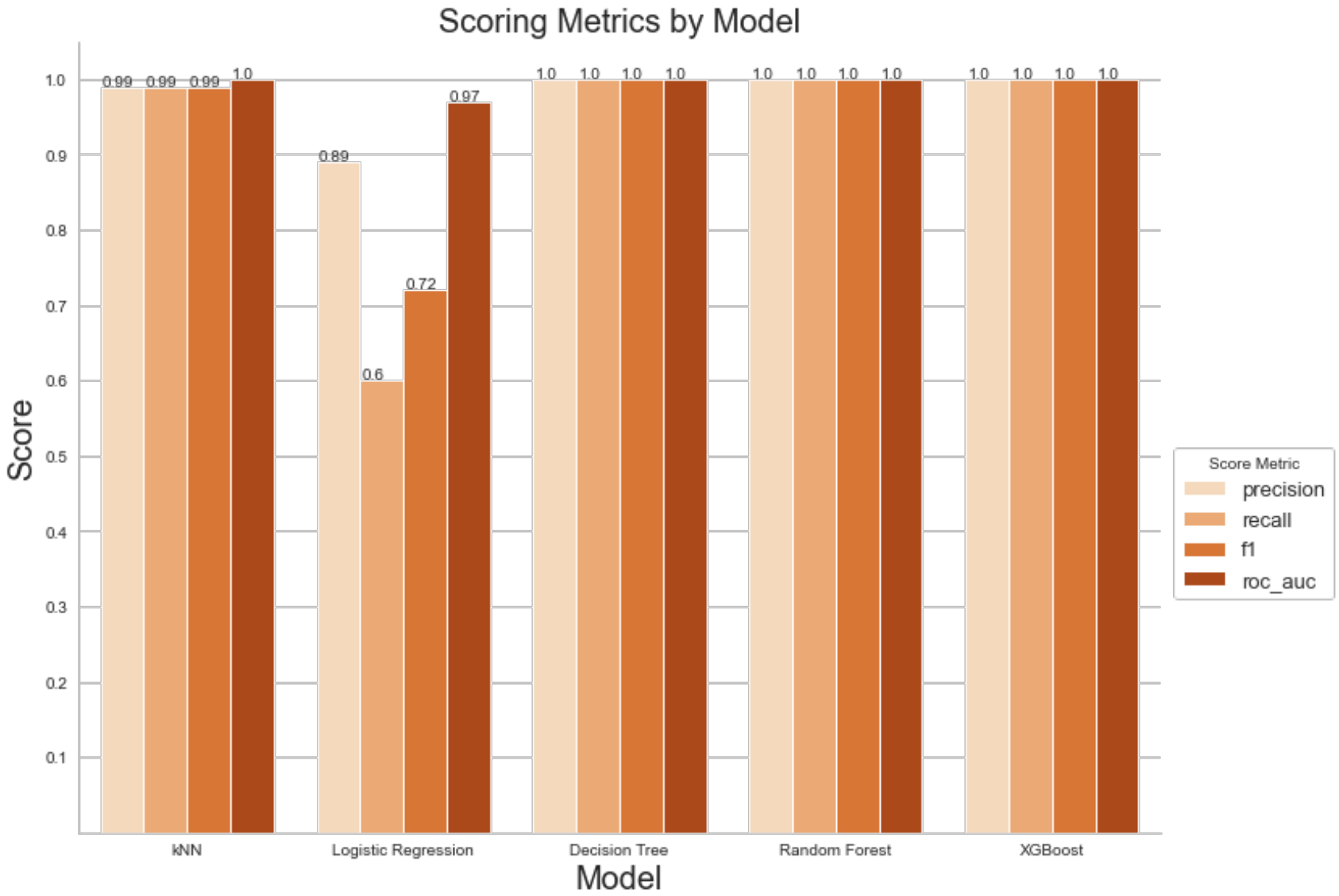
Evaluate Scoring Metrics

Findings: Baseline Metrics

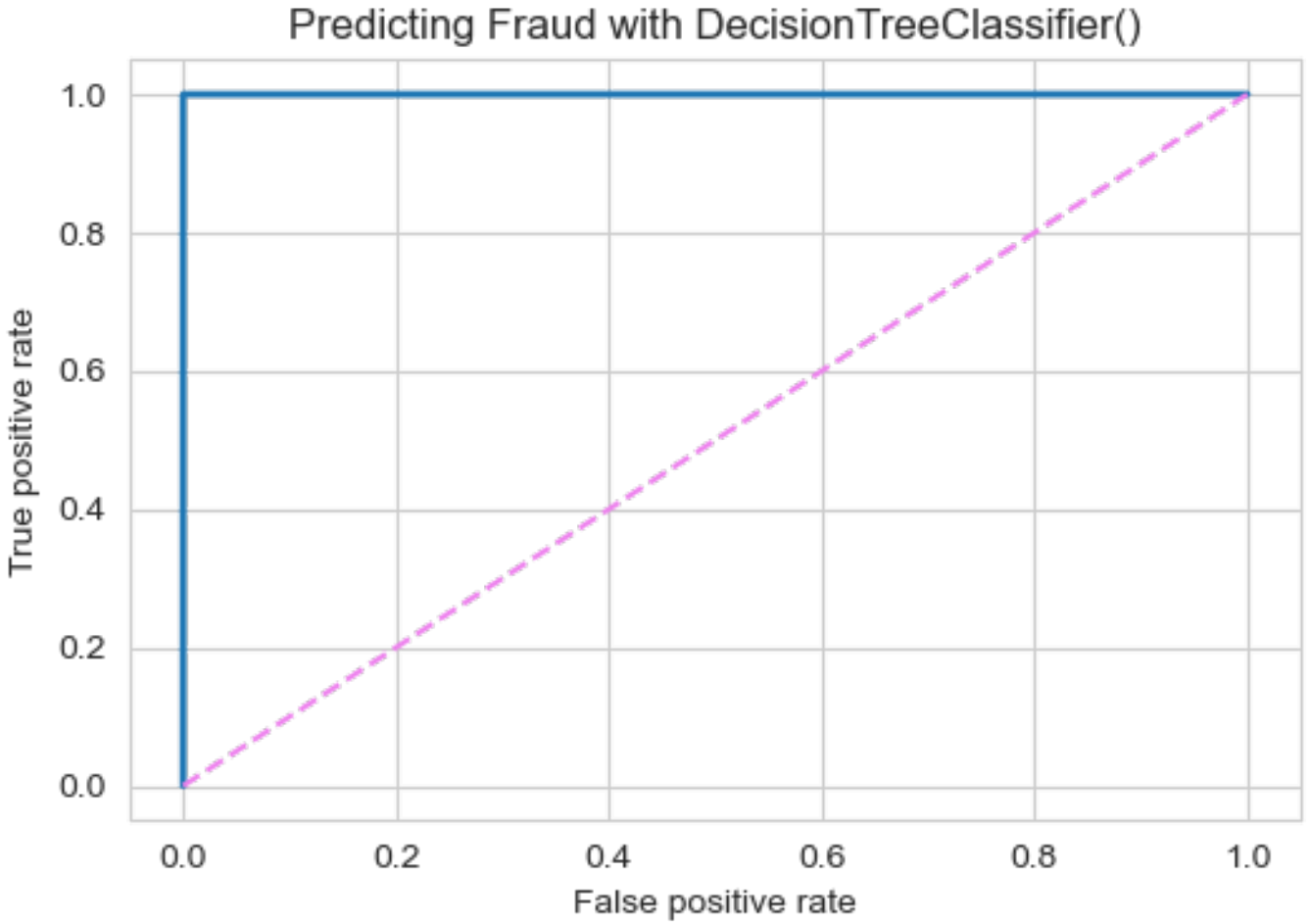


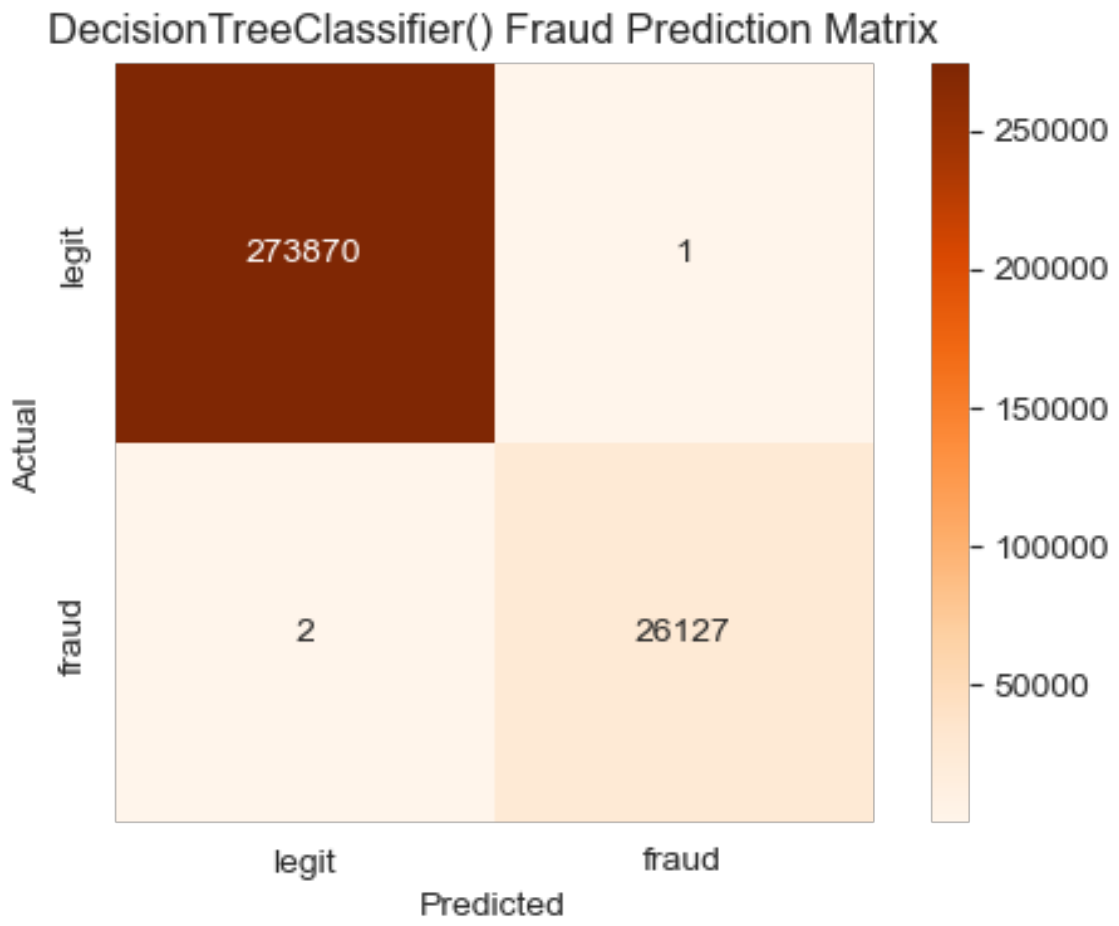


Findings: Baseline Metrics



Ideal Model: Decision Tree





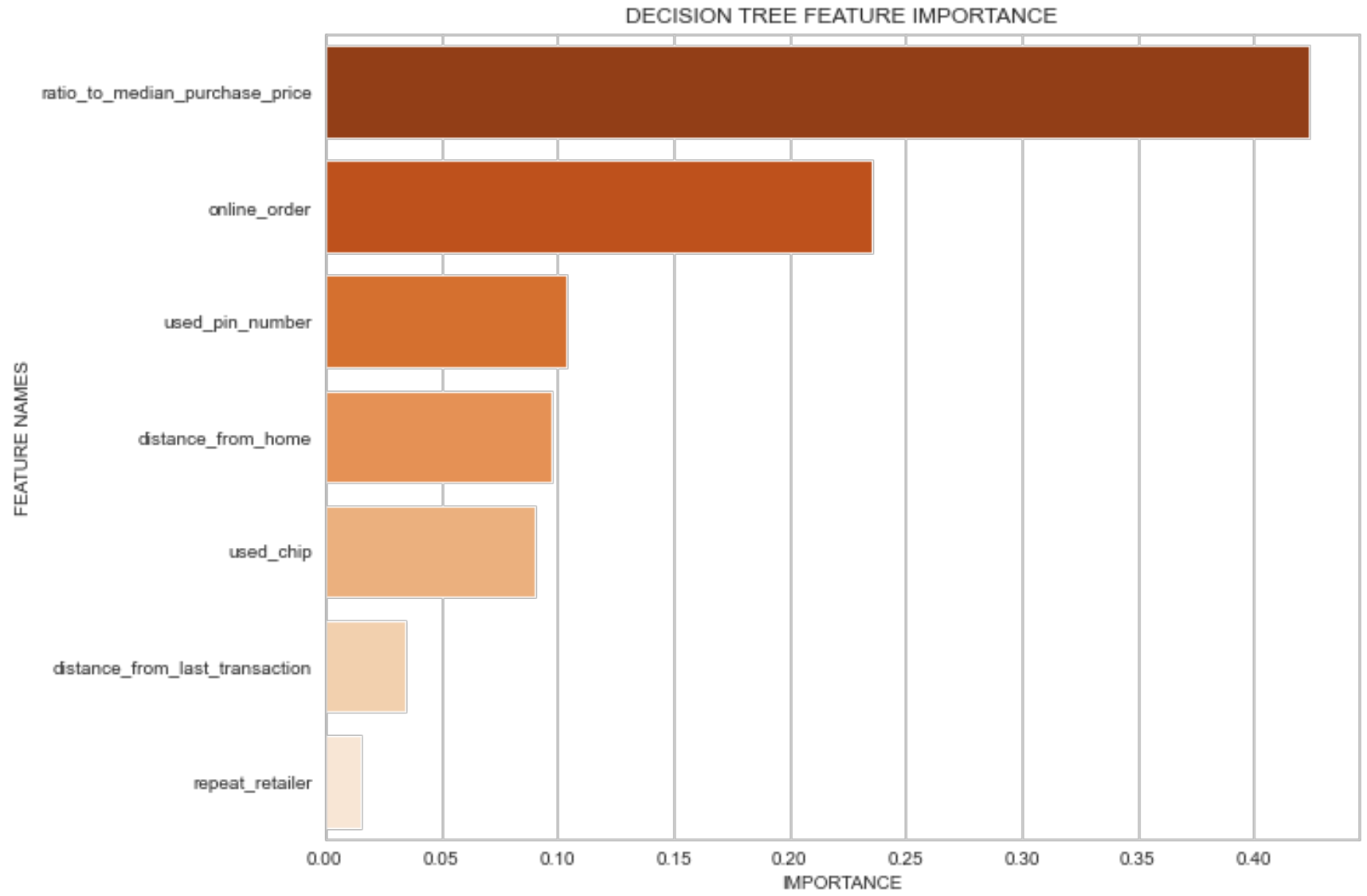
Decision Tree Evaluation

Optimal hyperparameters:

- max_leaf_nodes = 39
- min_samples_split = 3

Recall, Precision, F1 Score: 100%

Final Model: Decision Tree



Conclusion

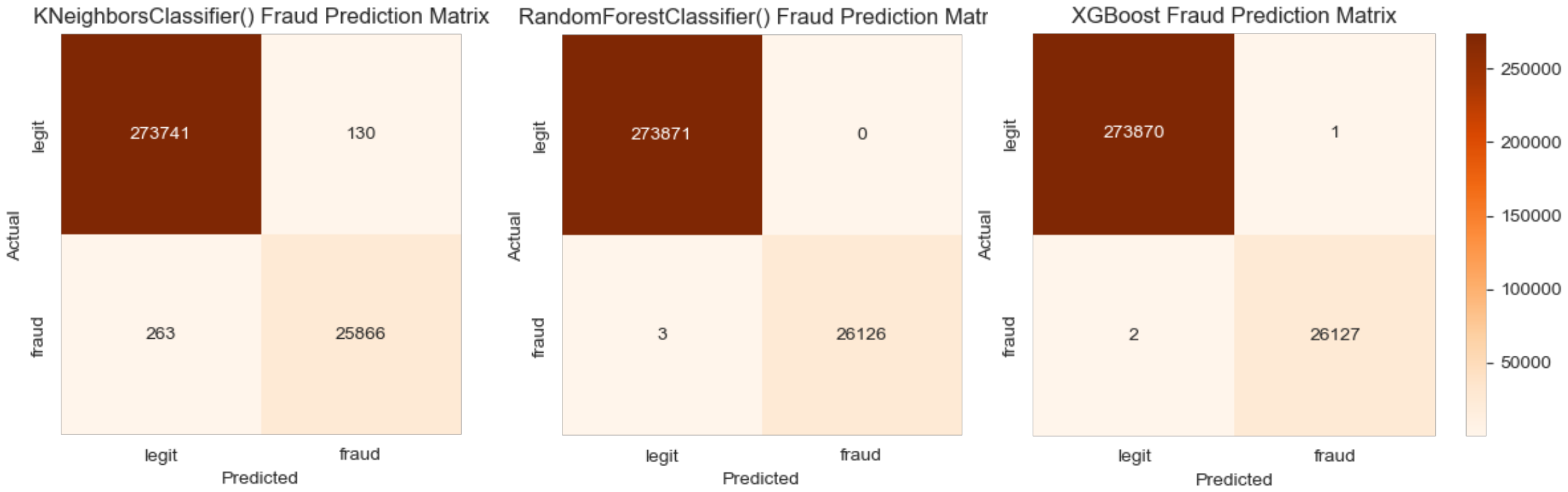
- We can predict which transaction will be fraudulent with exceptionally perfect accuracy, precision, and recall at 100%
- Because the model performed so well, there was no need to adjust class imbalance via over/under-sampling, weighting, or threshold adjustments
- Which begs the question: Does this represent real-world data? Will our analysis benefit from the inclusion of additional data for further modeling?

Questions?

Metis
Classification
July 2020

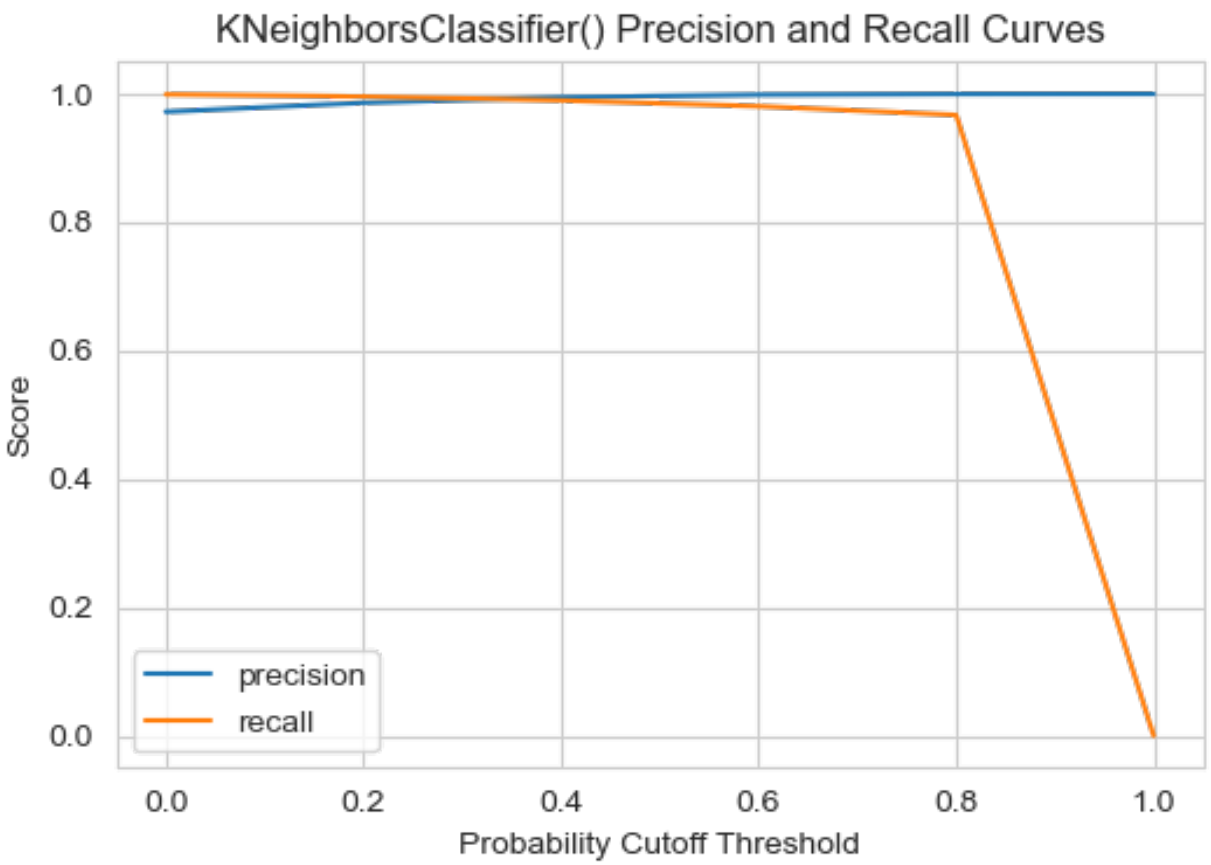

CARD

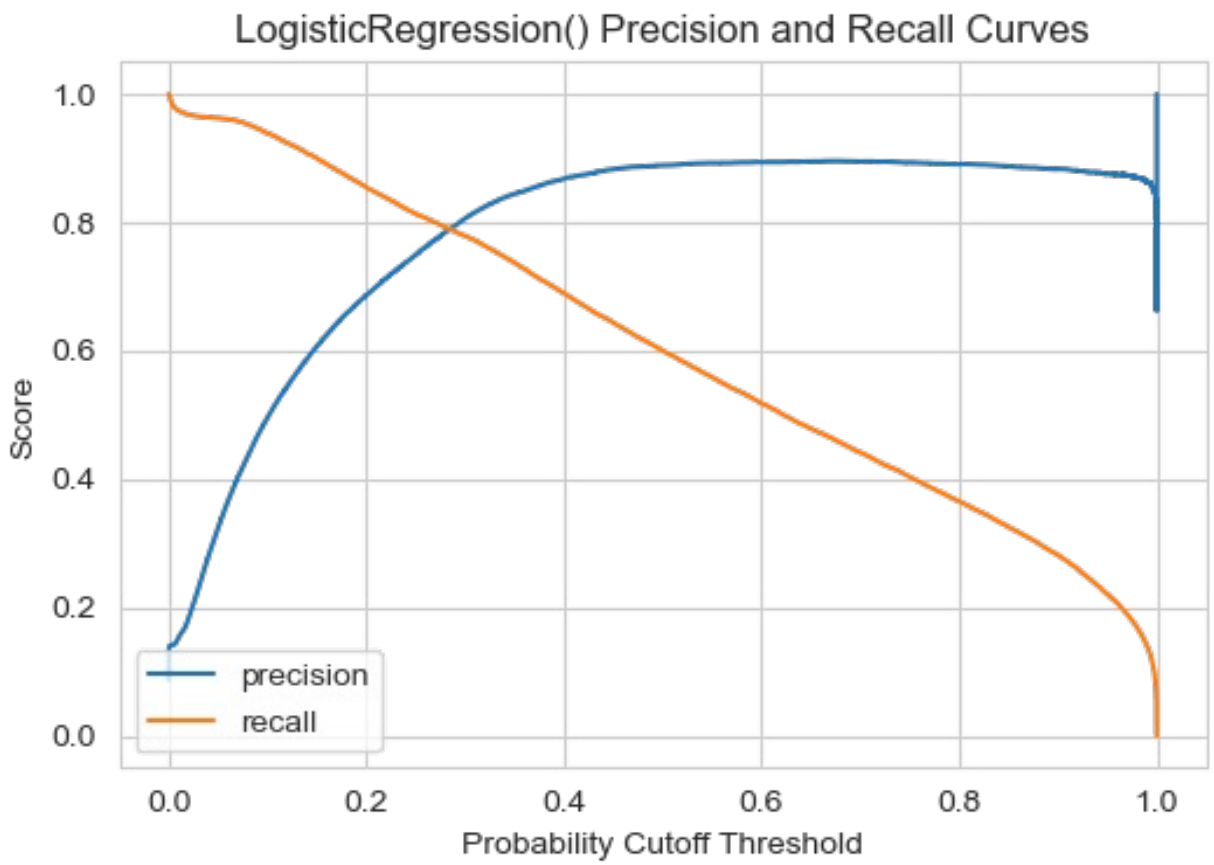
Appendix


KNeighborsClassifier() Fraud Prediction Matrix
Actual
legit
fraud
273741
130
263
25866
legit
fraud
Predicted
RandomForestClassifier() Fraud Prediction Matr
Actual
legit
fraud
273871
0
3
26126
legit
fraud
Predicted
XGBoost Fraud Prediction Matrix
Actual
legit
fraud
273870
1
2
26127
legit
fraud
Predicted
250000
200000
150000
100000
50000


CARD

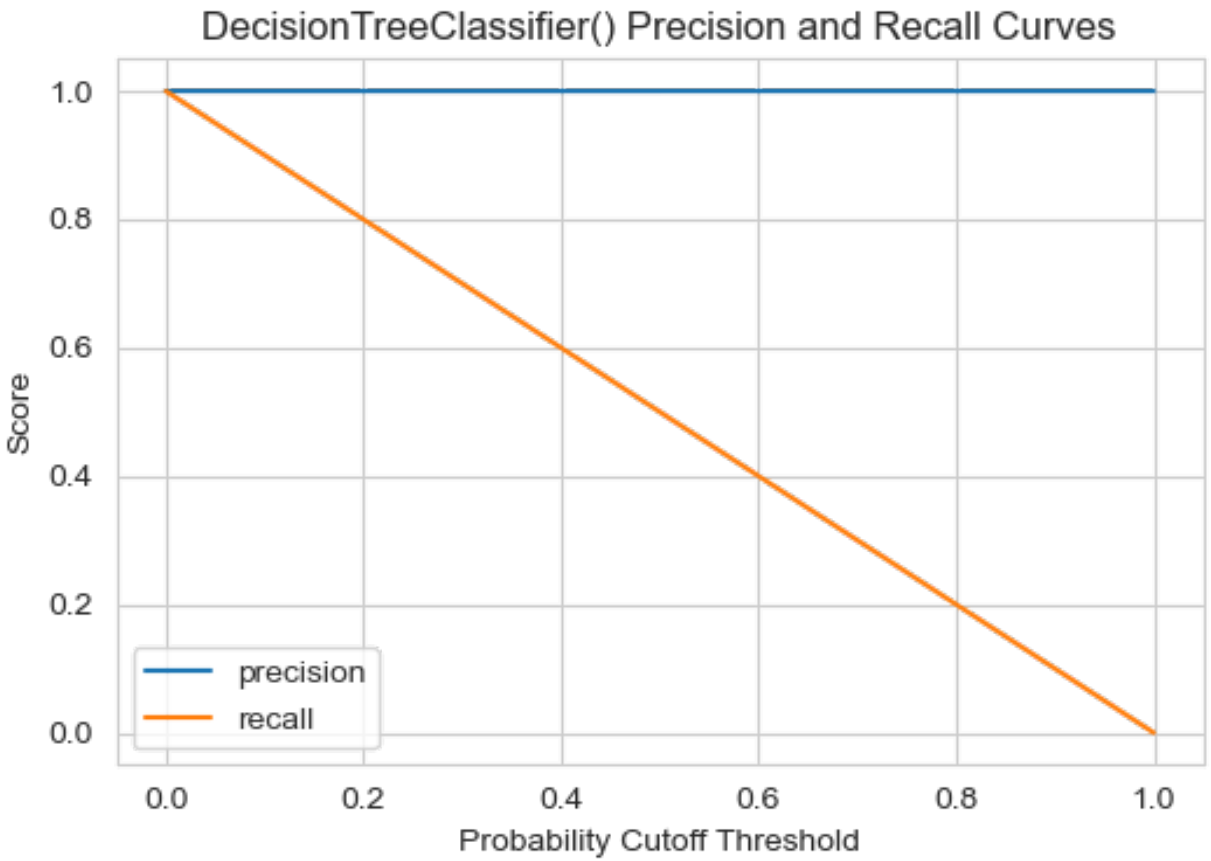
Appendix

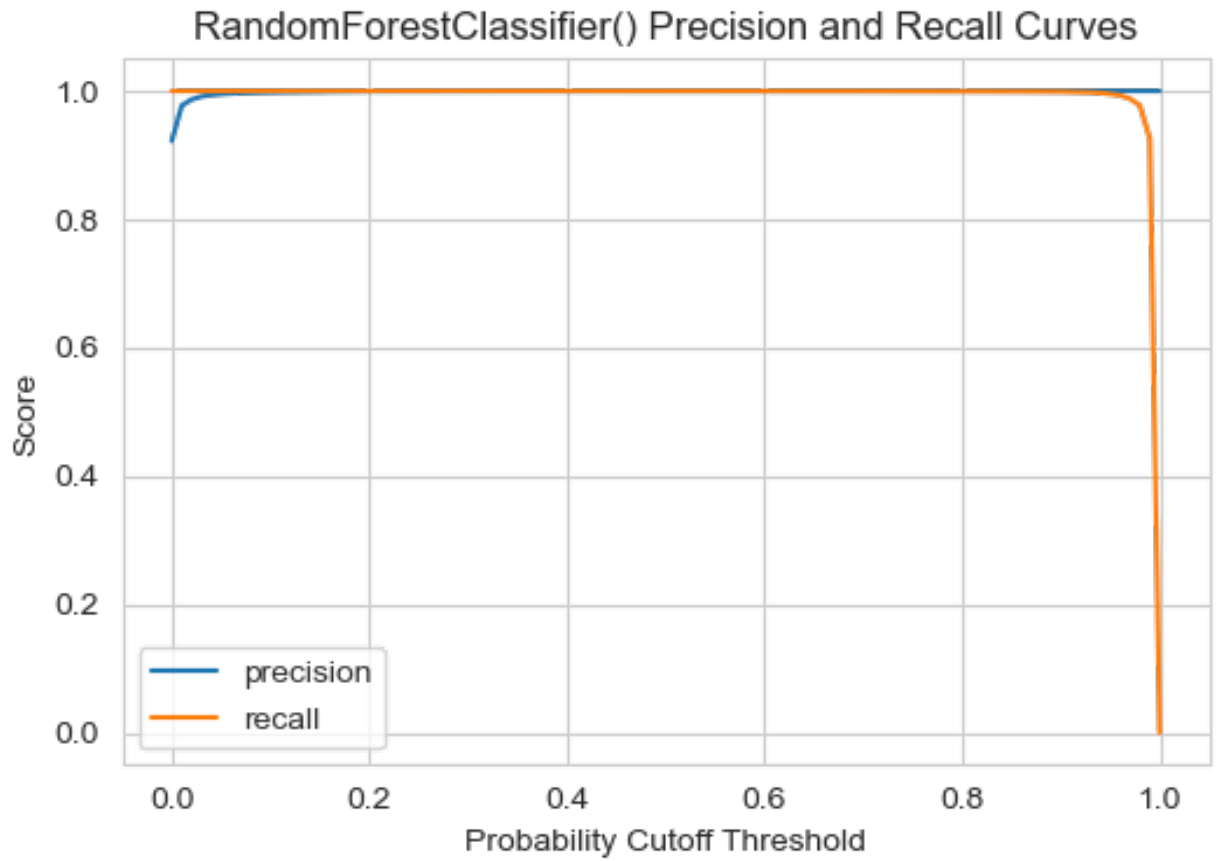

KNeighborsClassifier() Precision and Recall Curves
Score
Probability Cutoff Threshold
precision
recall
0.0
0.2
0.4
0.6
0.8
1.0


LogisticRegression() Precision and Recall Curves
Score
Probability Cutoff Threshold
precision
recall
0.0
0.2
0.4
0.6
0.8
1.0


CARD

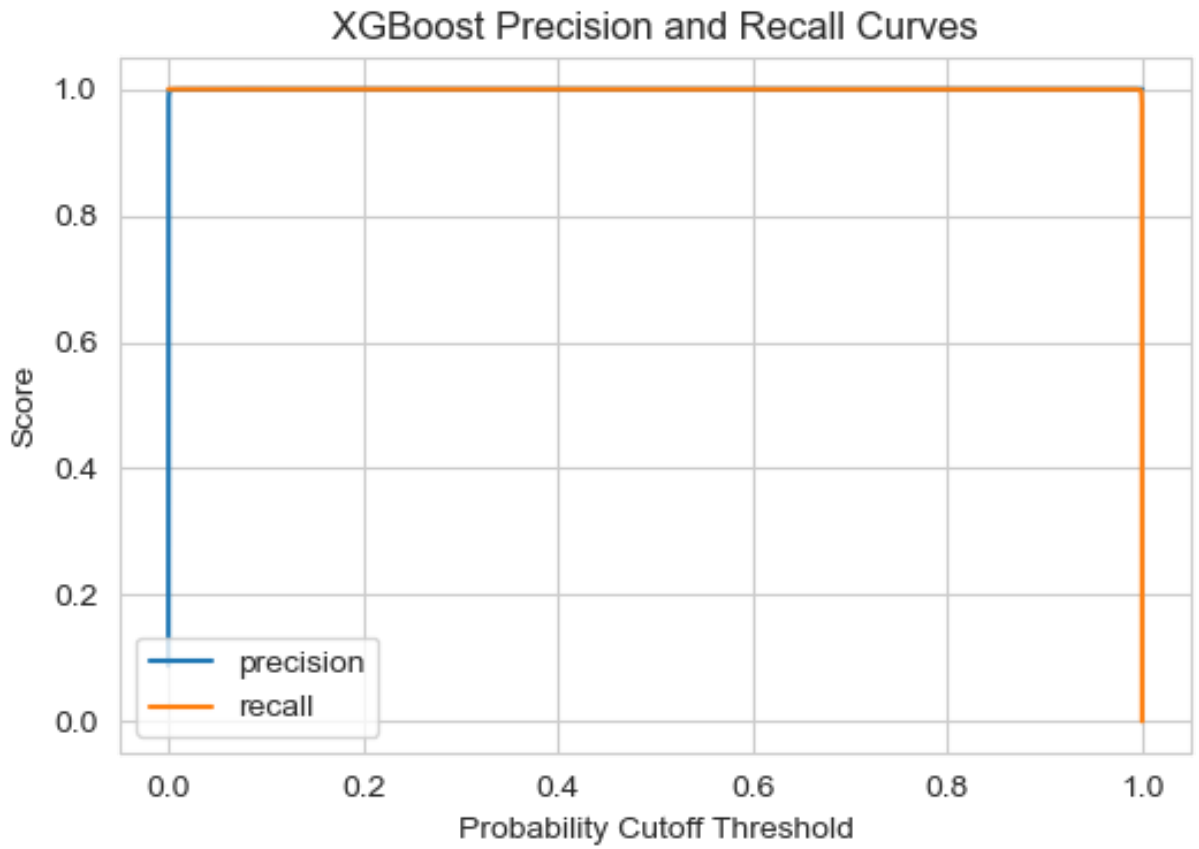
Appendix


DecisionTreeClassifier() Precision and Recall Curves
Score
Probability Cutoff Threshold
precision
recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


RandomForestClassifier() Precision and Recall Curves
Score
Probability Cutoff Threshold
precision
recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

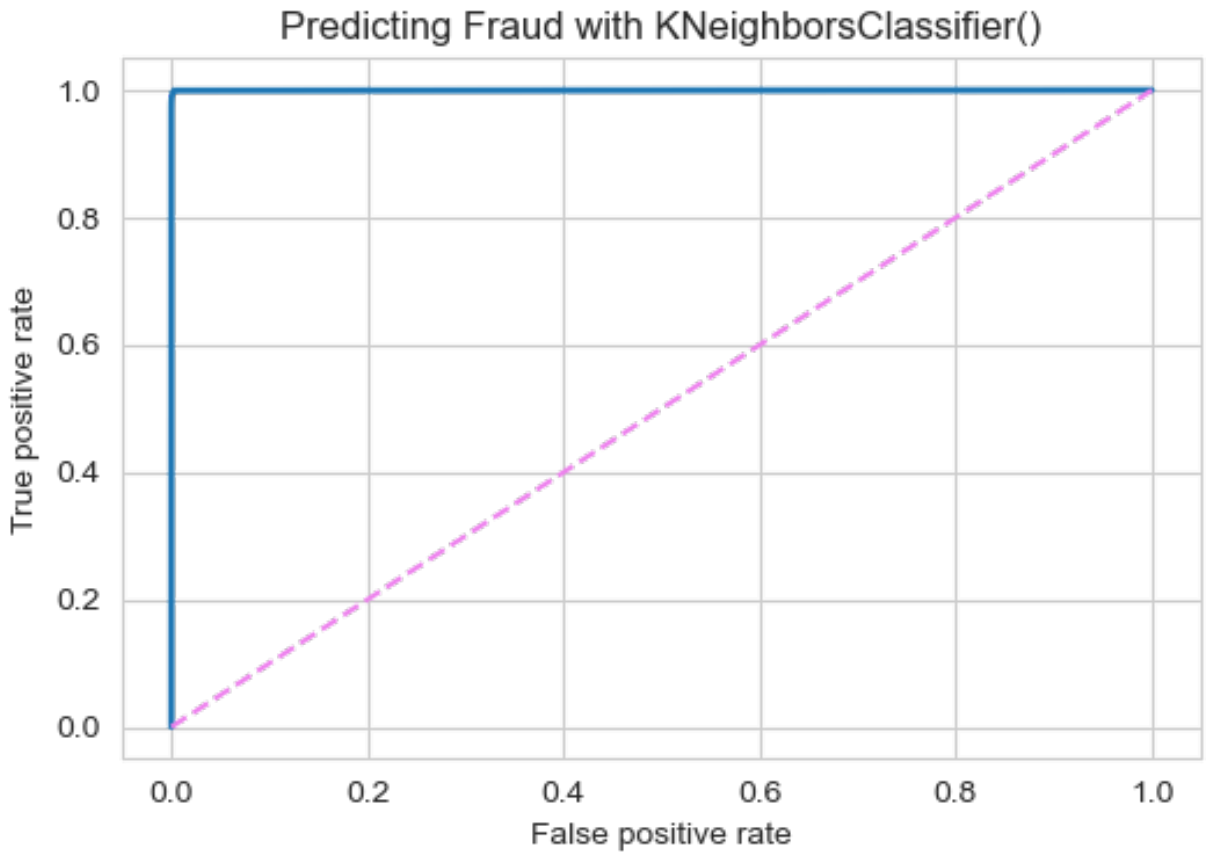

CARD

Appendix


XGBoost Precision and Recall Curves
Score
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
Probability Cutoff Threshold

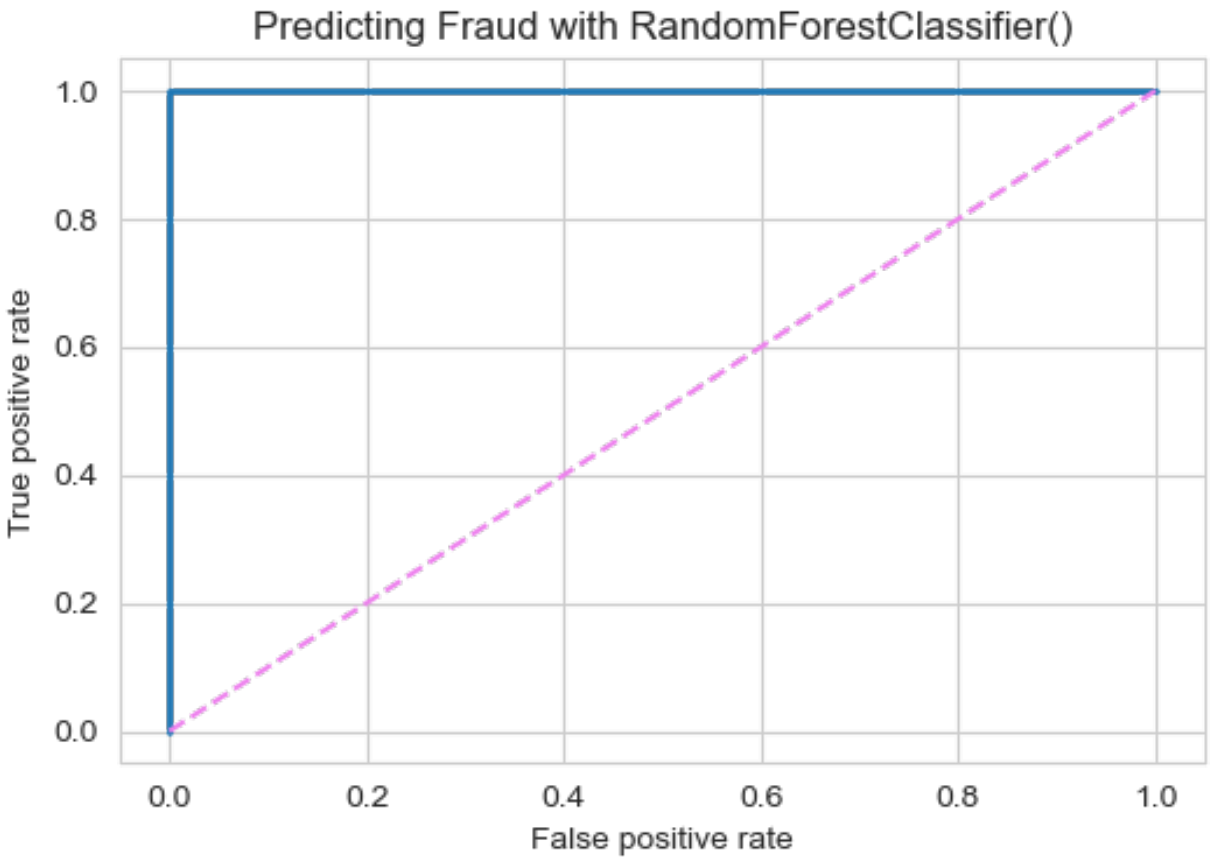

CARD

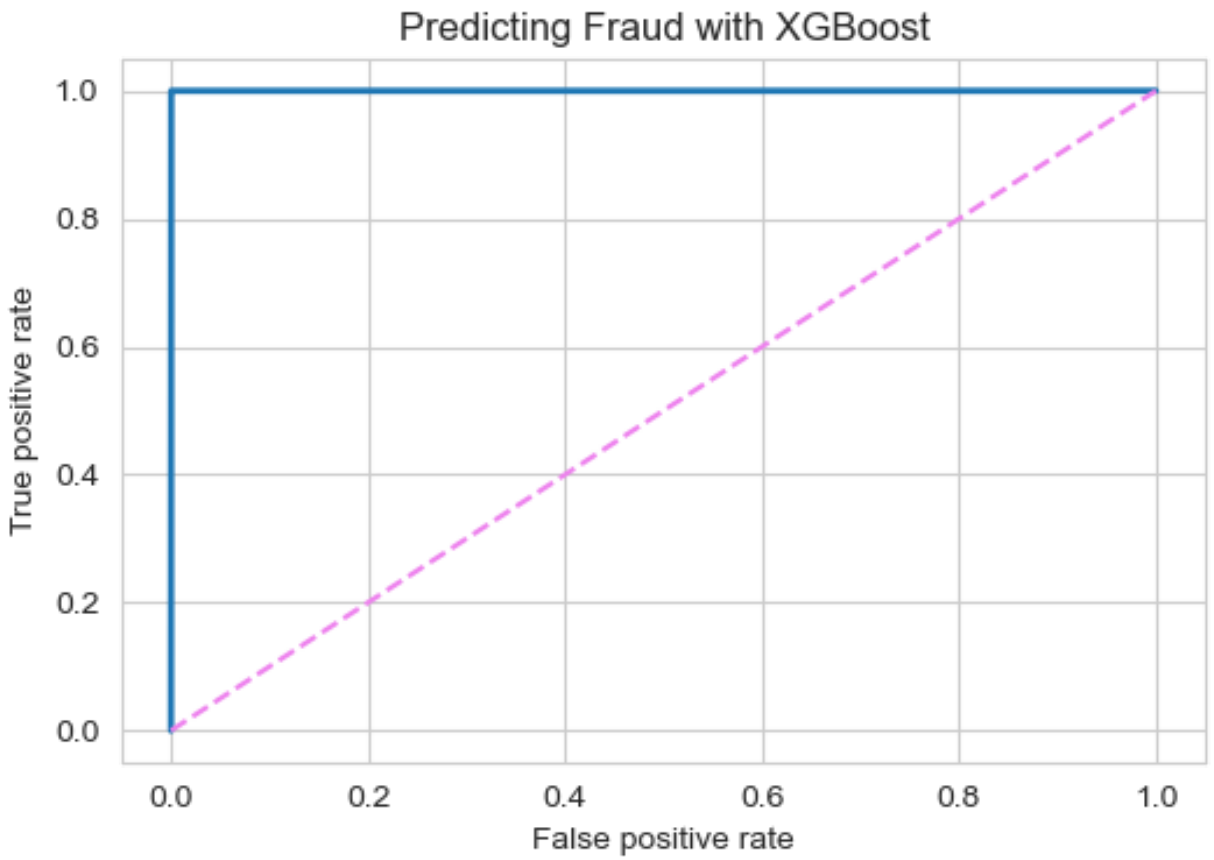
Appendix


Predicting Fraud with KNeighborsClassifier()
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0


CARD

Appendix


Predicting Fraud with RandomForestClassifier()
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0


Predicting Fraud with XGBoost
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0