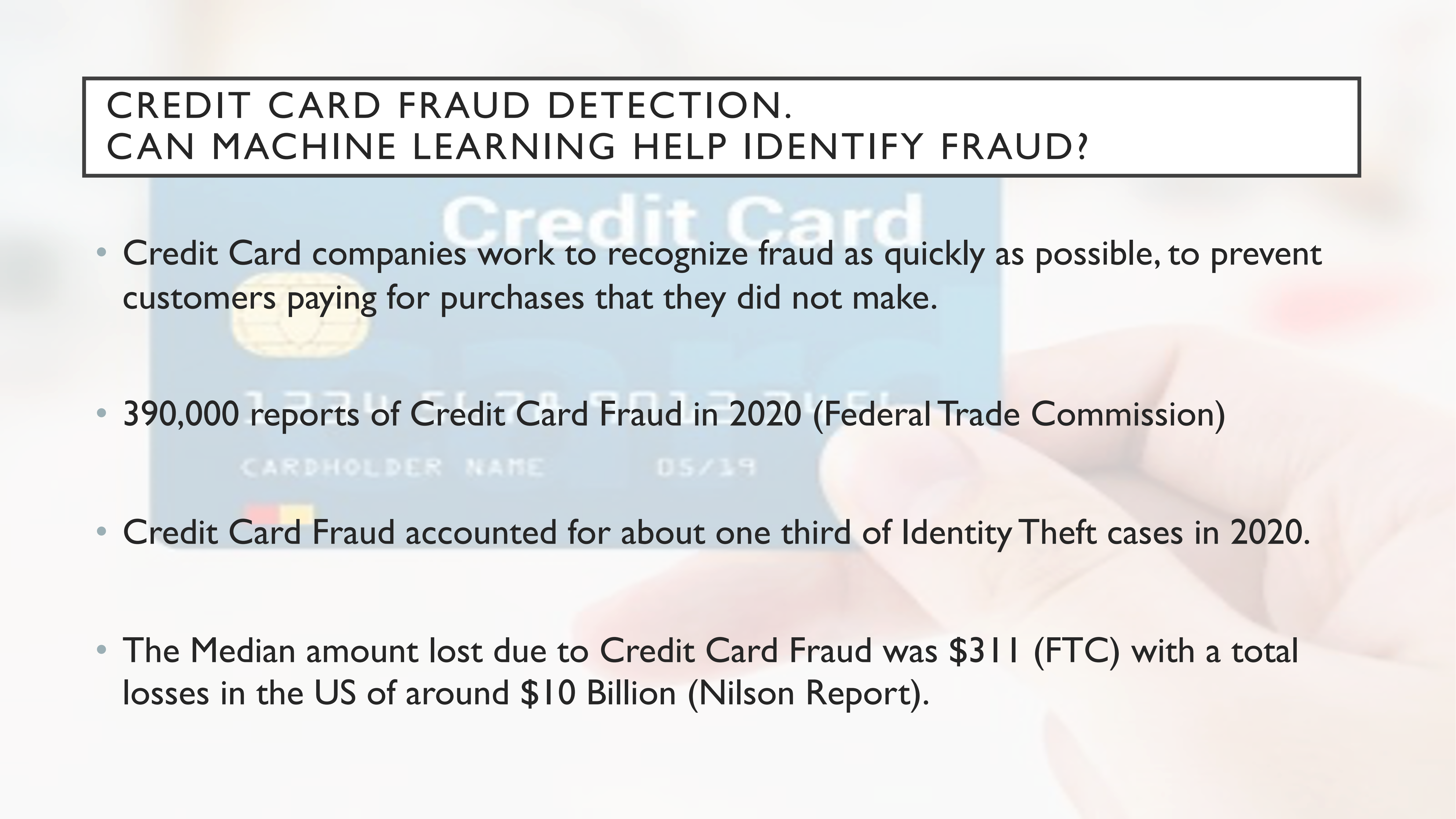
CREDIT CARD FRAUD DETECTION

Paul O’Leary

Comp 4449 Capstone

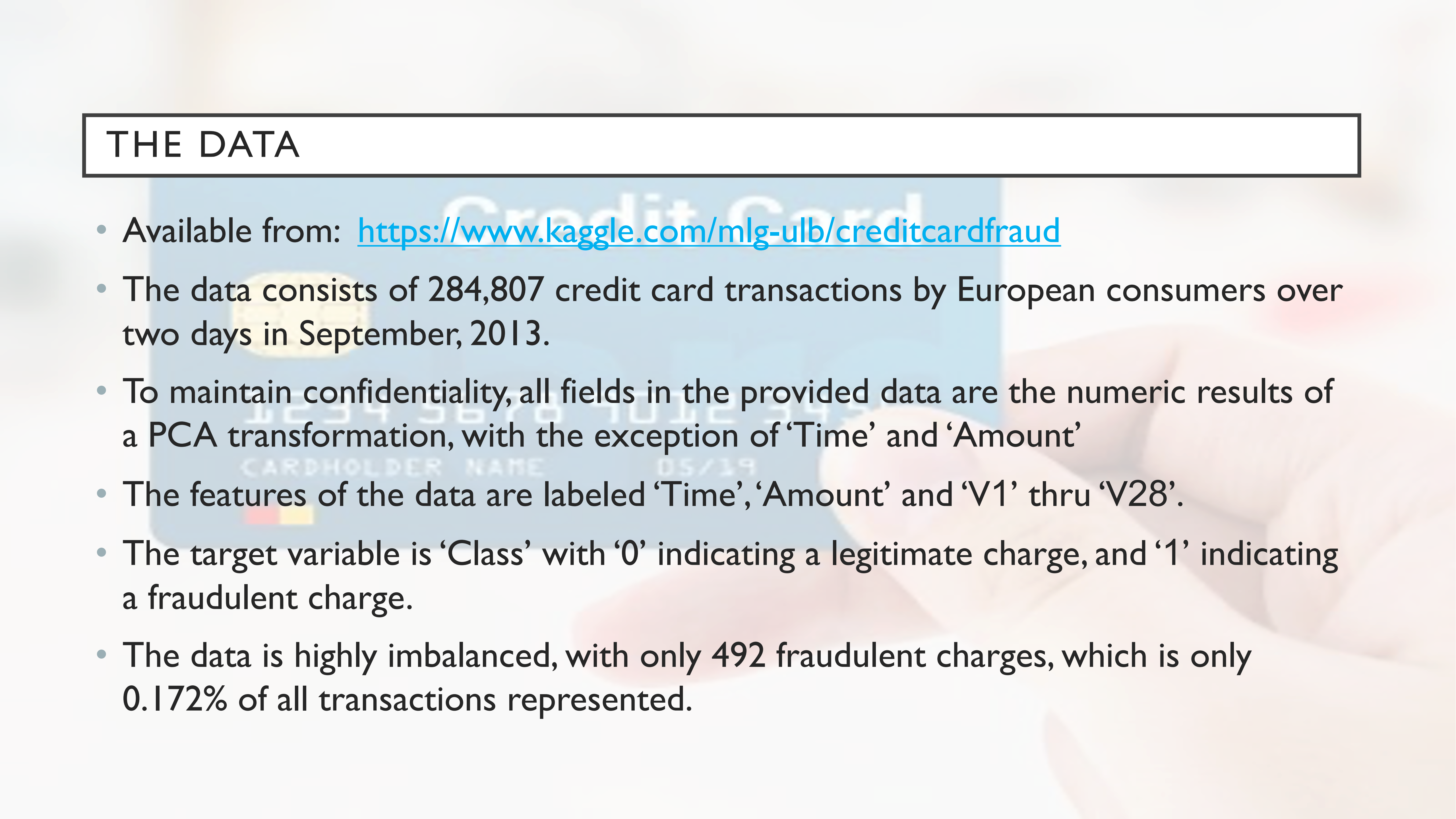
October, 2021

CREDIT CARD FRAUD DETECTION.
CAN MACHINE LEARNING HELP IDENTIFY FRAUD?



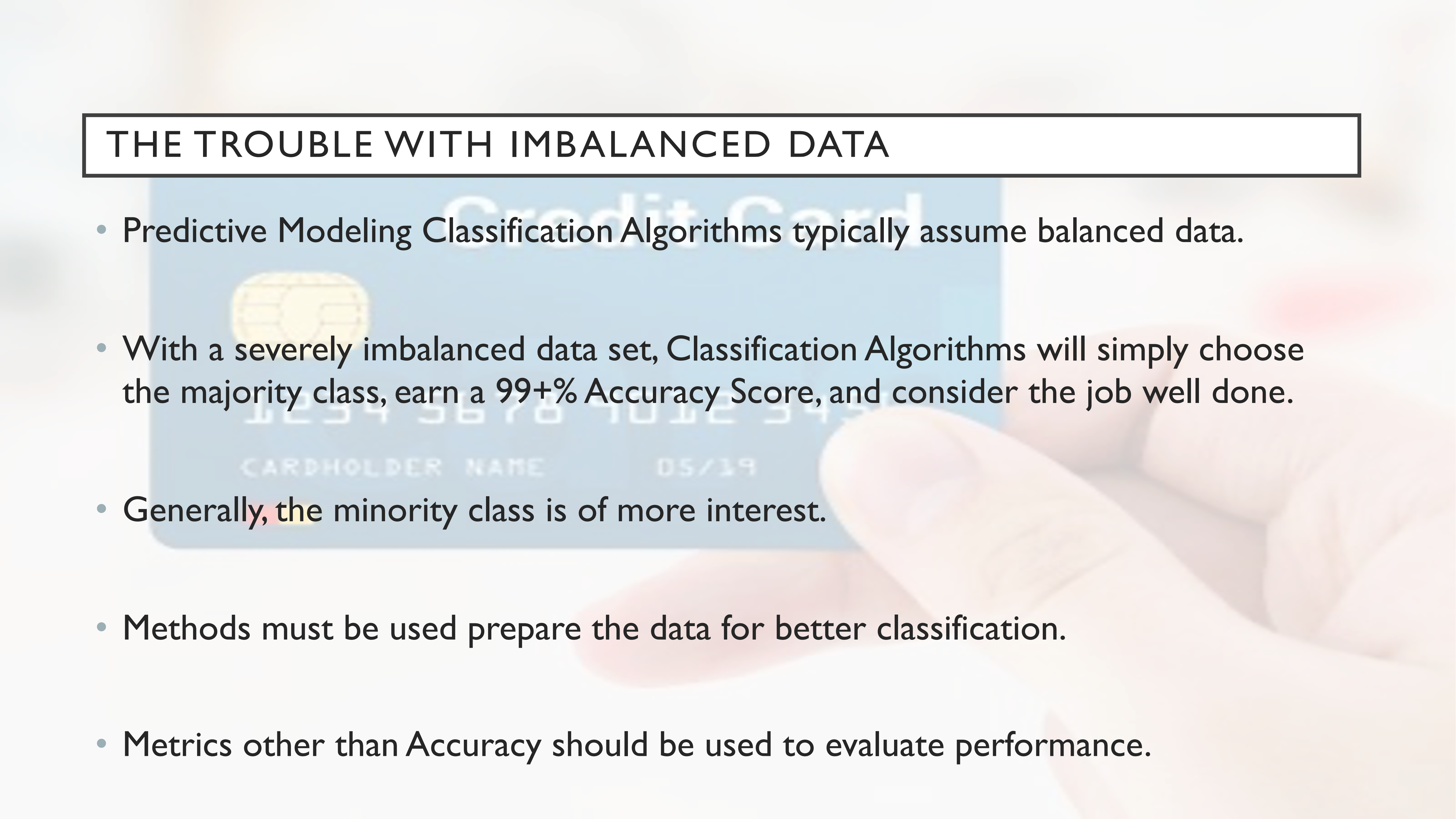
- Credit Card companies work to recognize fraud as quickly as possible, to prevent customers paying for purchases that they did not make.
- 390,000 reports of Credit Card Fraud in 2020 (Federal Trade Commission)
- Credit Card Fraud accounted for about one third of Identity Theft cases in 2020.
- The Median amount lost due to Credit Card Fraud was $311 (FTC) with a total losses in the US of around $10 Billion (Nilson Report).

THE DATA



- Available from: https://www.kaggle.com/mlg-ulb/creditcardfraud
- The data consists of 284,807 credit card transactions by European consumers over two days in September, 2013.
- To maintain confidentiality, all fields in the provided data are the numeric results of a PCA transformation, with the exception of 'Time' and 'Amount'
- The features of the data are labeled 'Time', 'Amount' and 'V1' thru 'V28'.
- The target variable is 'Class' with '0' indicating a legitimate charge, and '1' indicating a fraudulent charge.
- The data is highly imbalanced, with only 492 fraudulent charges, which is only 0.172% of all transactions represented.

THE TROUBLE WITH IMBALANCED DATA



- Predictive Modeling Classification Algorithms typically assume balanced data.

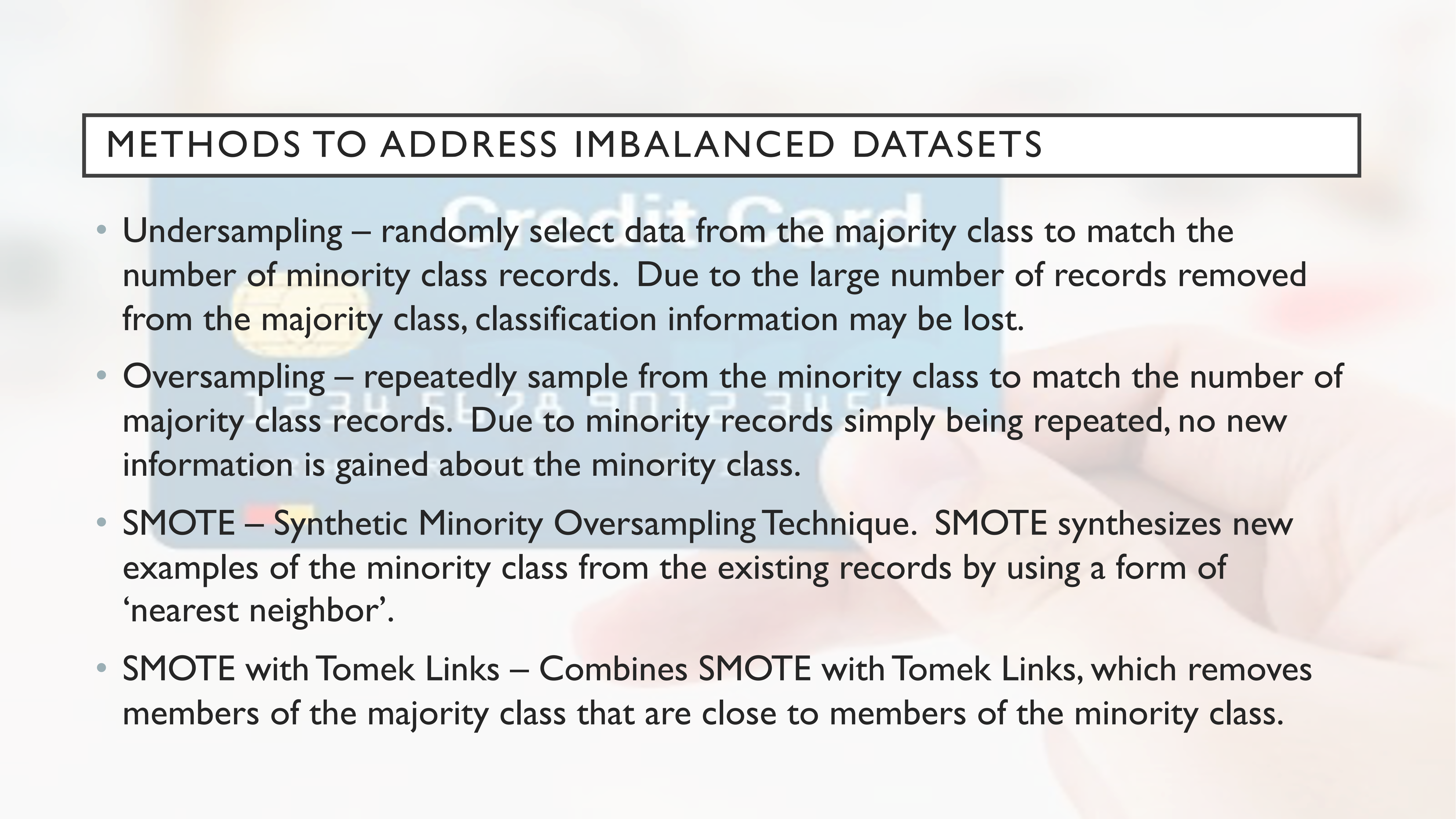
- With a severely imbalanced data set, Classification Algorithms will simply choose the majority class, earn a 99+% Accuracy Score, and consider the job well done.

- Generally, the minority class is of more interest.

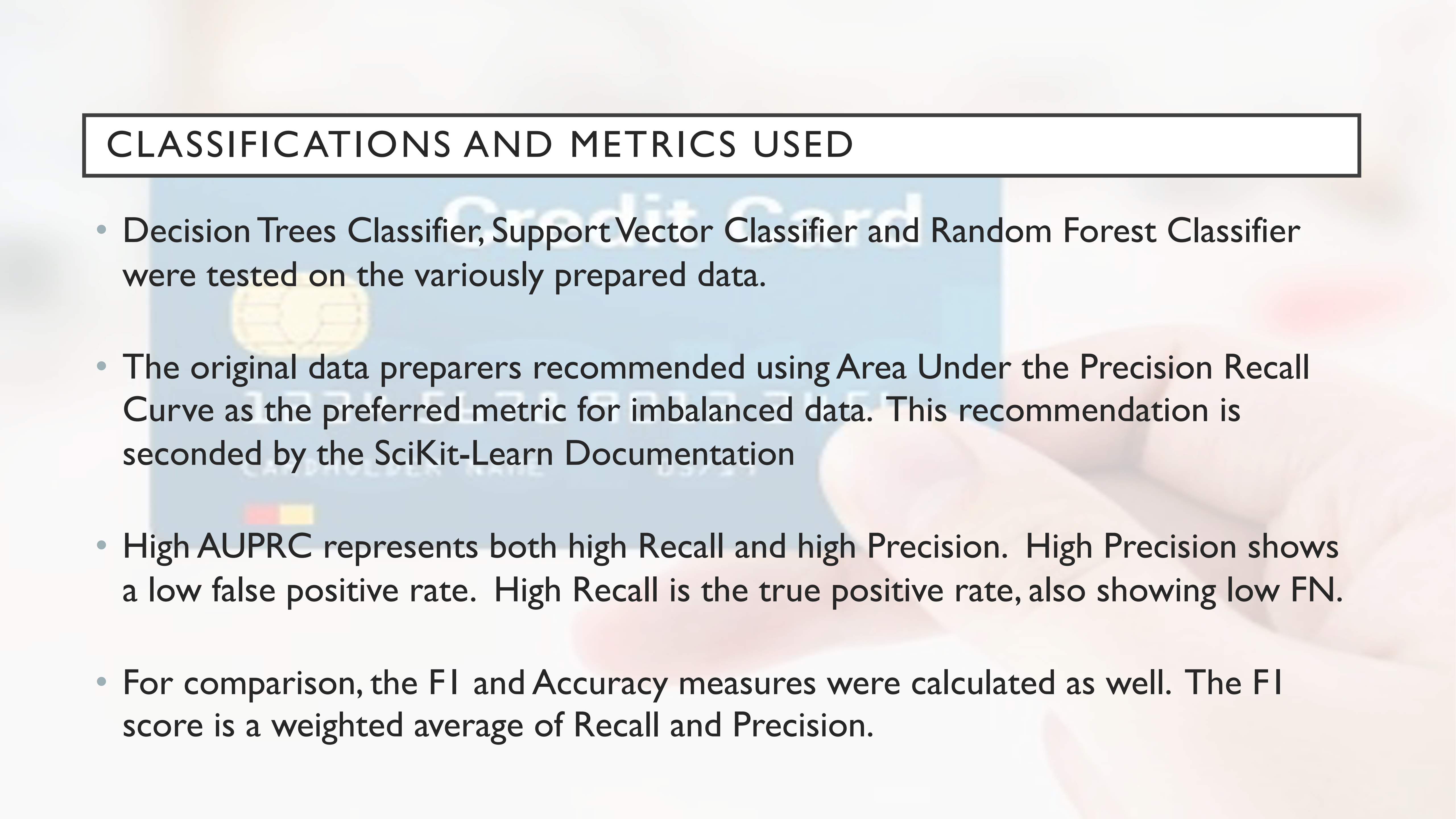
- Methods must be used prepare the data for better classification.

- Metrics other than Accuracy should be used to evaluate performance.

METHODS TO ADDRESS IMBALANCED DATASETS



- Undersampling – randomly select data from the majority class to match the number of minority class records. Due to the large number of records removed from the majority class, classification information may be lost.
- Oversampling – repeatedly sample from the minority class to match the number of majority class records. Due to minority records simply being repeated, no new information is gained about the minority class.
- SMOTE – Synthetic Minority Oversampling Technique. SMOTE synthesizes new examples of the minority class from the existing records by using a form of 'nearest neighbor'.
- SMOTE with Tomek Links – Combines SMOTE with Tomek Links, which removes members of the majority class that are close to members of the minority class.



CLASSIFICATIONS AND METRICS USED

- Decision Trees Classifier, Support Vector Classifier and Random Forest Classifier were tested on the variously prepared data.

- The original data preparers recommended using Area Under the Precision Recall Curve as the preferred metric for imbalanced data. This recommendation is seconded by the SciKit-Learn Documentation

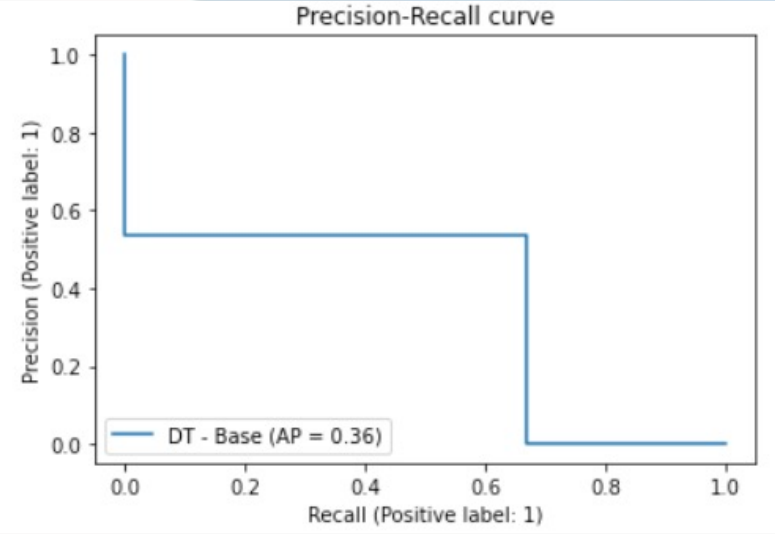
- High AUPRC represents both high Recall and high Precision. High Precision shows a low false positive rate. High Recall is the true positive rate, also showing low FN.

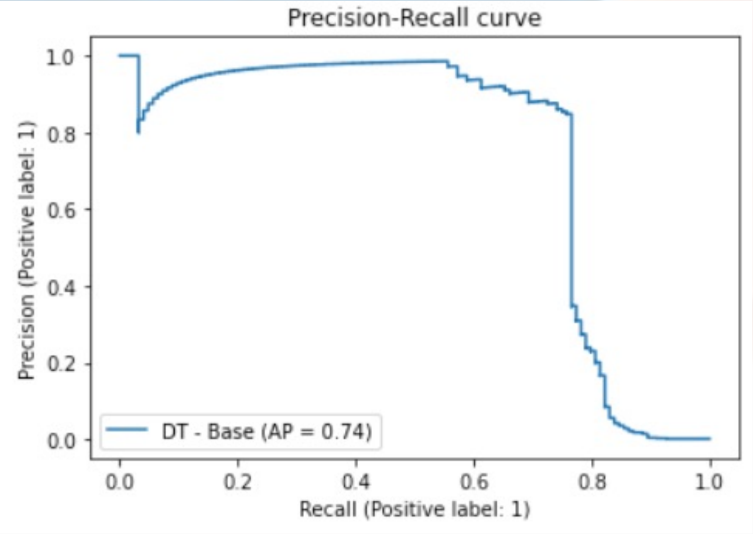
- For comparison, the F1 and Accuracy measures were calculated as well. The F1 score is a weighted average of Recall and Precision.

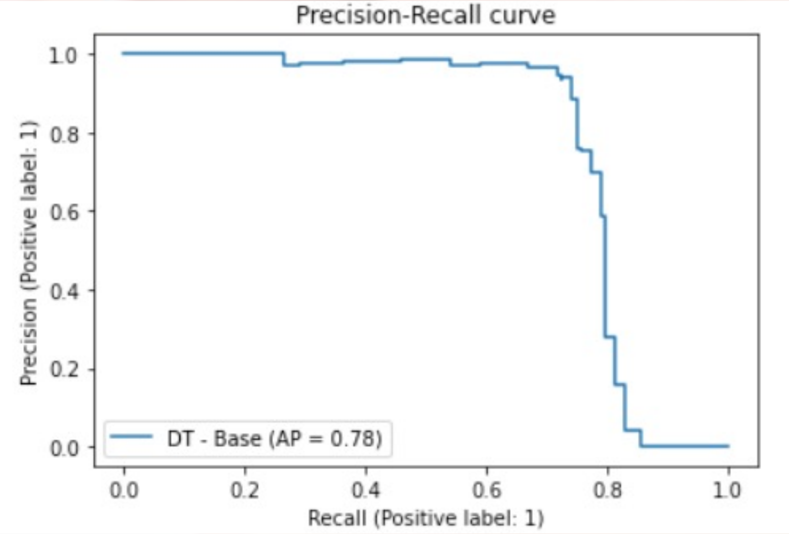
RESULTS

- Raw Data, no preparation

	Decision Tree	SVC	Random Forest
Accuracy	0.999	0.999	0.9995
F1	0.595	0.678	0.790
AUPRC	0.359	0.516	0.646



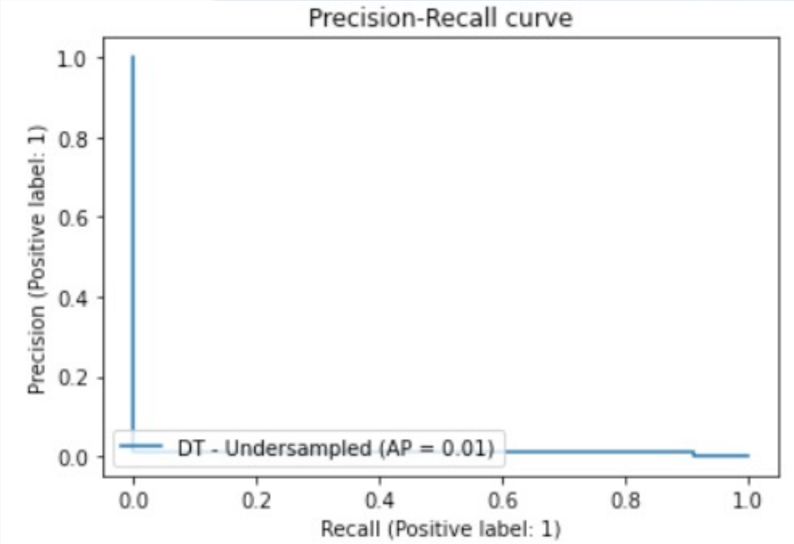


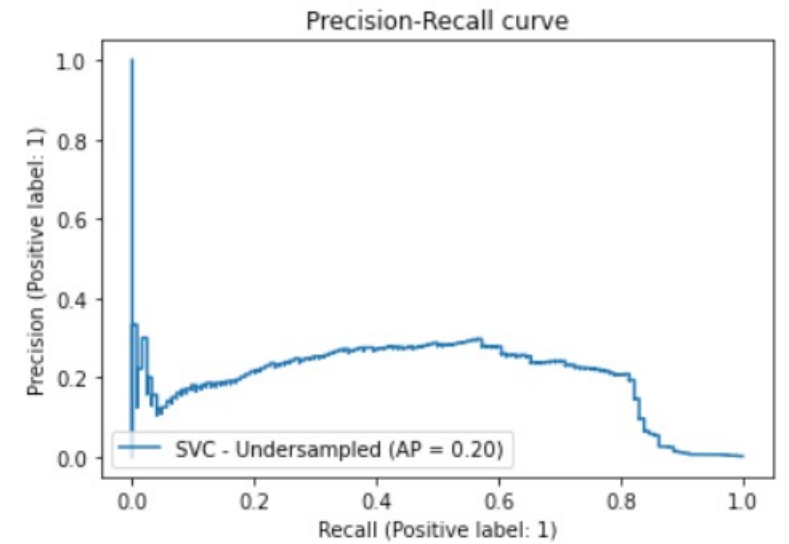


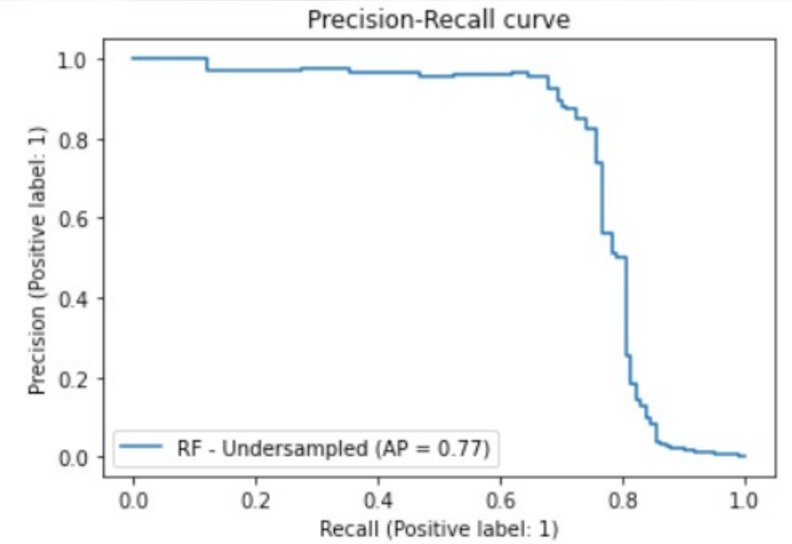
RESULTS (CONT.)

- Under-sampled Data

	Decision Tree	SVC	Random Forest
Accuracy	0.870	0.949	0.983
F1	0.018	0.043	0.114
AUPRC	0.008	0.020	0.053



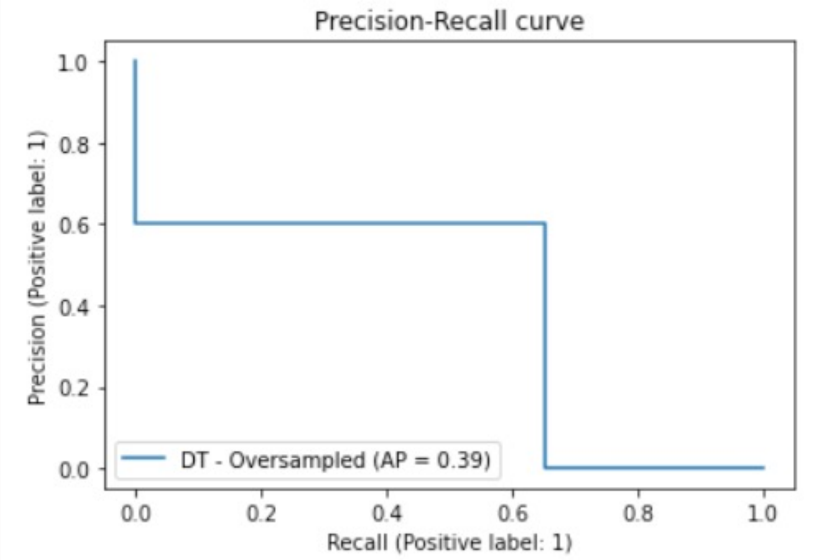


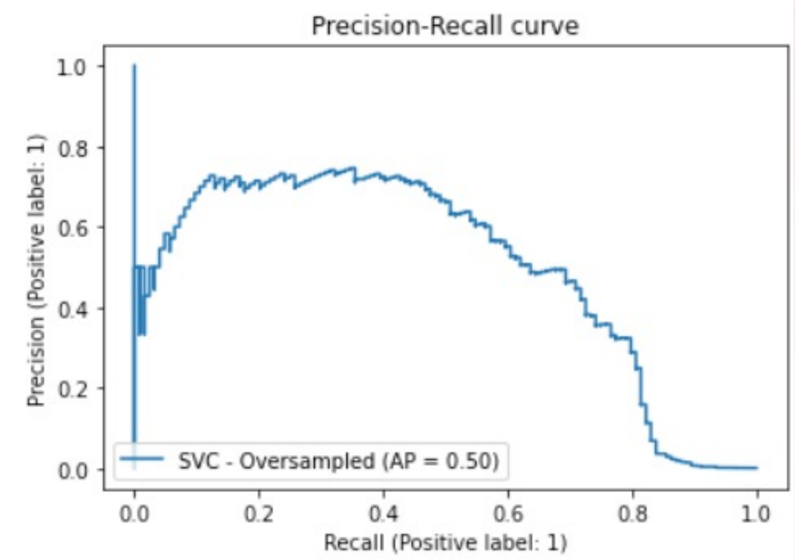


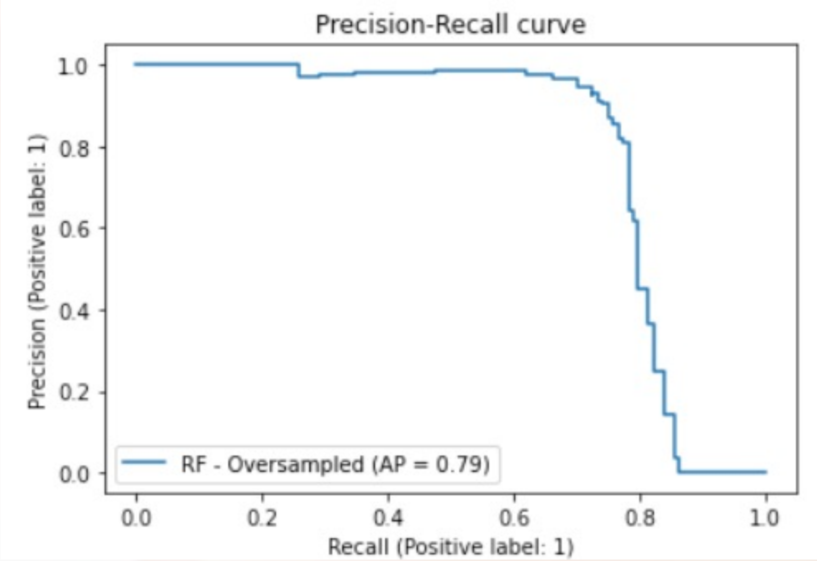
RESULTS (CONT.)

- Over-sampled Data

	Decision Tree	SVC	Random Forest
Accuracy	0.999	0.999	0.9996
F1	0.625	0.566	0.796
AUPRC	0.392	0.324	0.654



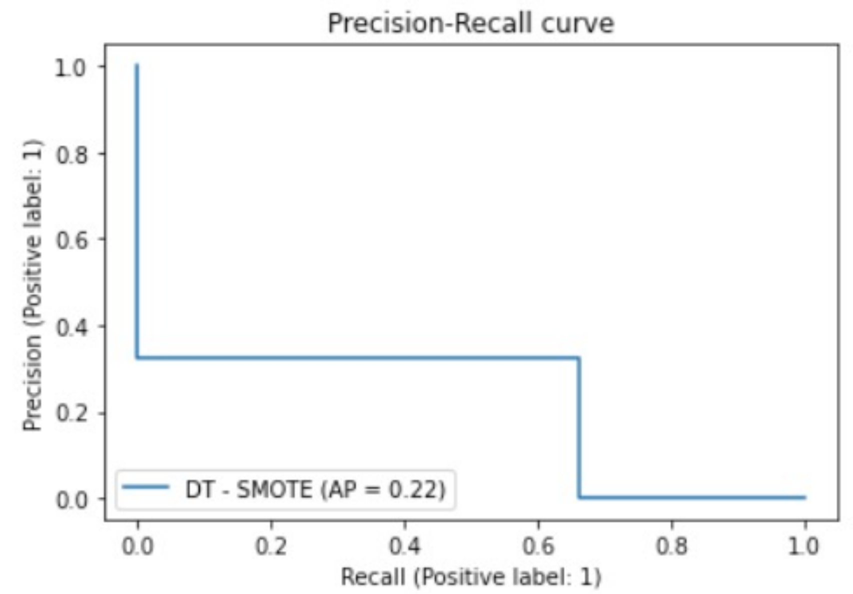


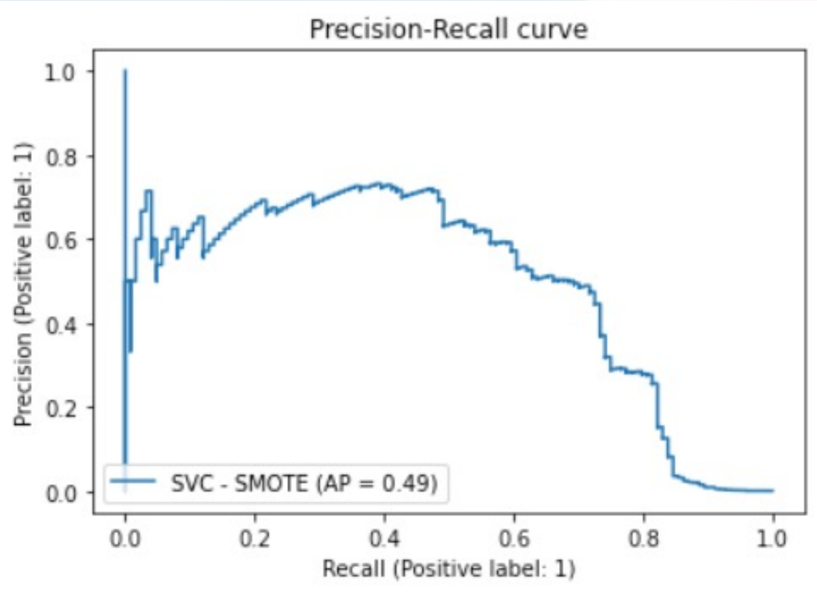


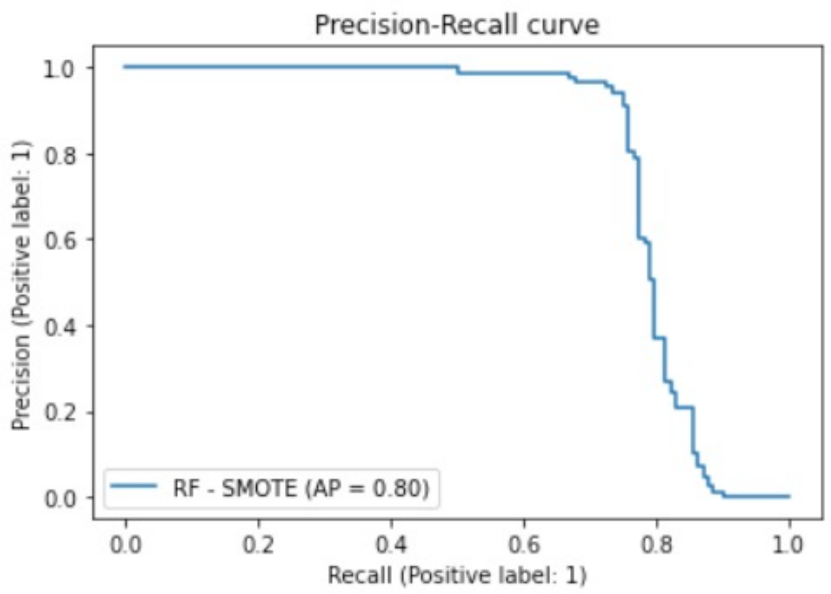
RESULTS (CONT.)

- SMOTE Over-sampled Data

	Decision Tree	SVC	Random Forest
Accuracy	0.998	0.999	0.9996
F1	0.437	0.578	0.834
AUPRC	0.216	0.338	0.705







RESULTS (CONT.)

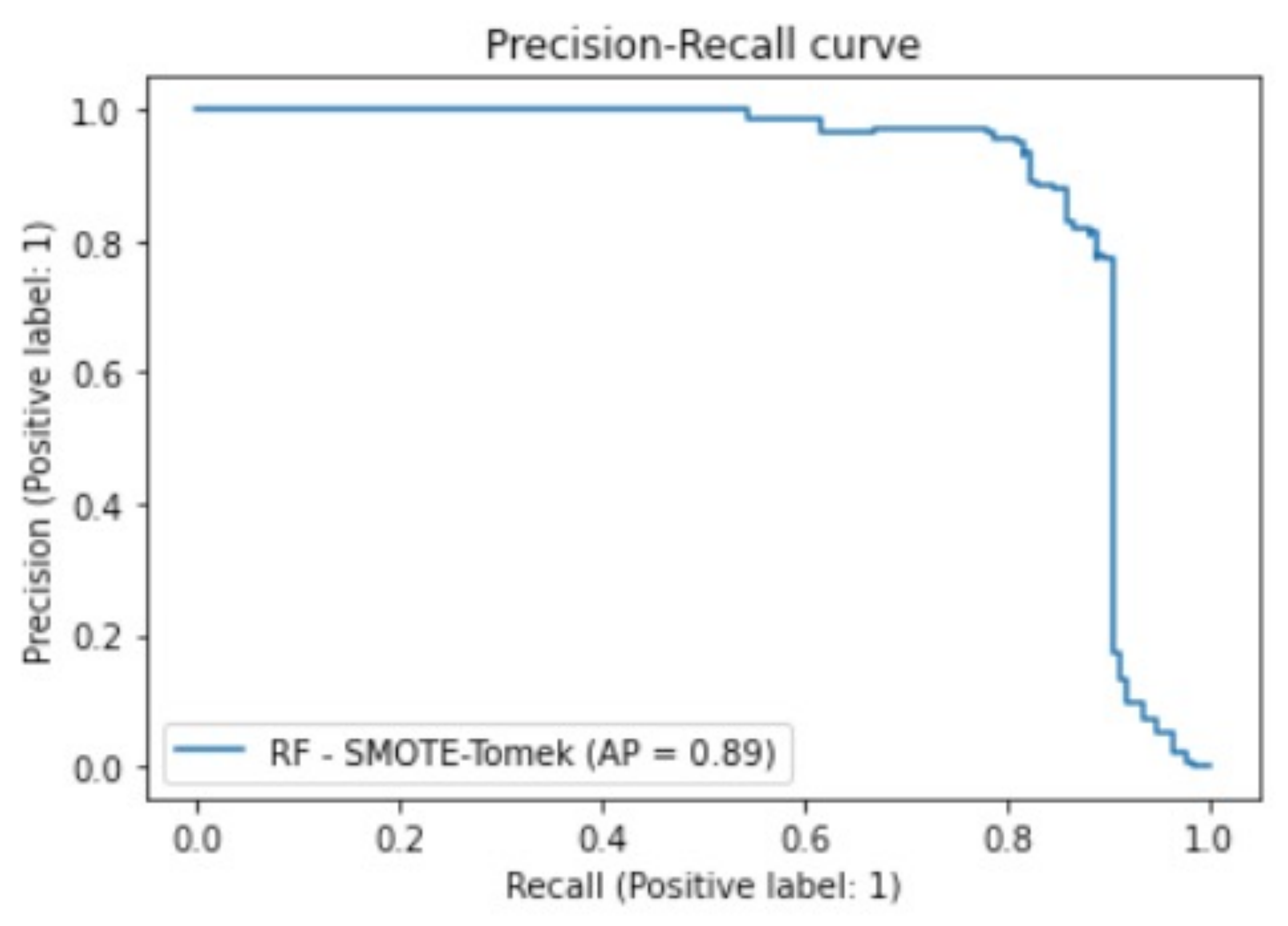
- SMOTE-Tomek Links Prepared Data, Random Forest Tuned Model, 70%/30% split

	Random Forest
Accuracy	0.9995
F1	0.854
AUPRC	0.730

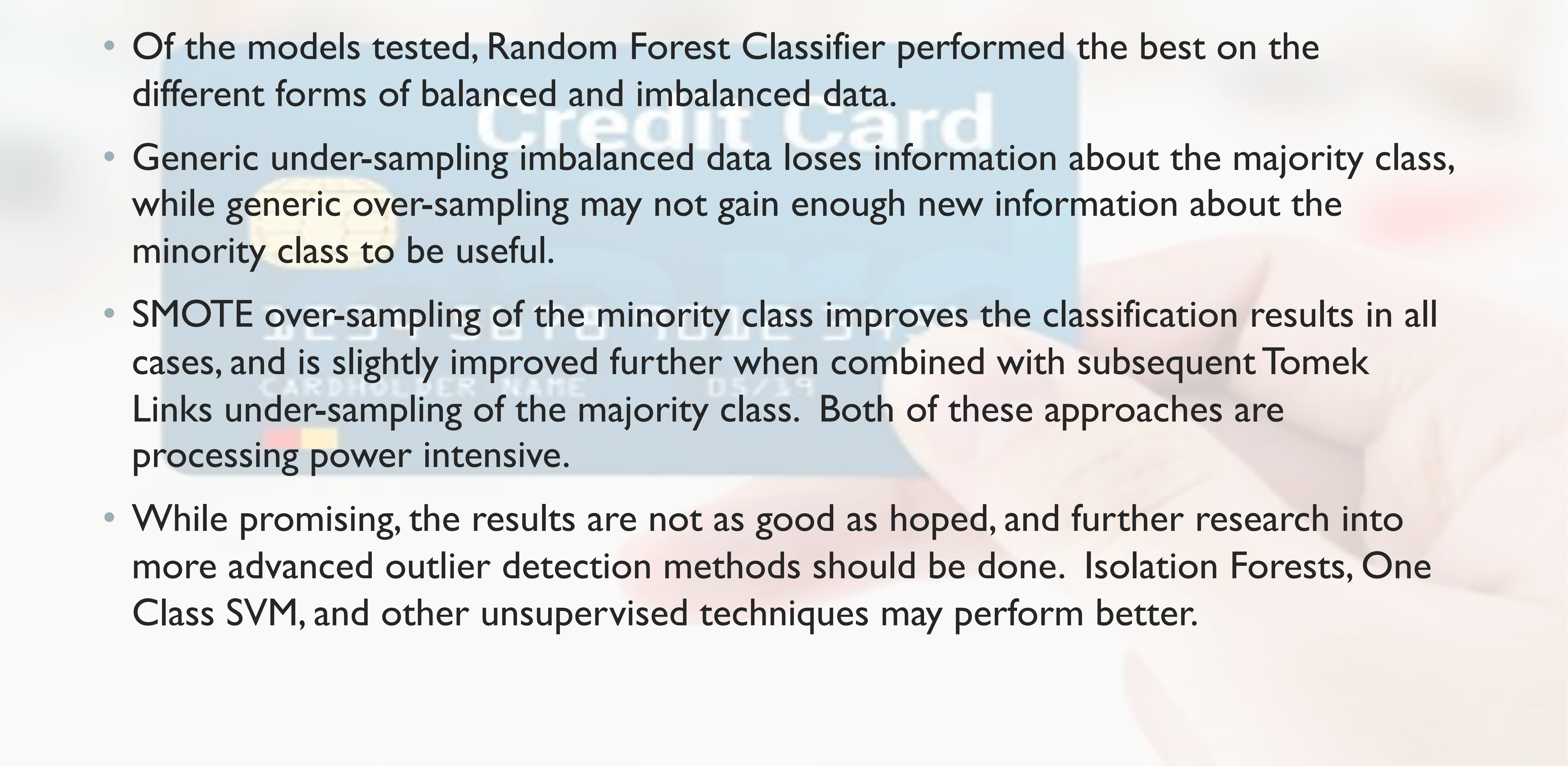
Confusion Matrix:

```
[[85286  21]
 [ 19    117]]
```

136 actual Fraud cases



CONCLUSION



- Of the models tested, Random Forest Classifier performed the best on the different forms of balanced and imbalanced data.
- Generic under-sampling imbalanced data loses information about the majority class, while generic over-sampling may not gain enough new information about the minority class to be useful.
- SMOTE over-sampling of the minority class improves the classification results in all cases, and is slightly improved further when combined with subsequent Tomek Links under-sampling of the majority class. Both of these approaches are processing power intensive.
- While promising, the results are not as good as hoped, and further research into more advanced outlier detection methods should be done. Isolation Forests, One Class SVM, and other unsupervised techniques may perform better.