Credit Card Fraud Detection

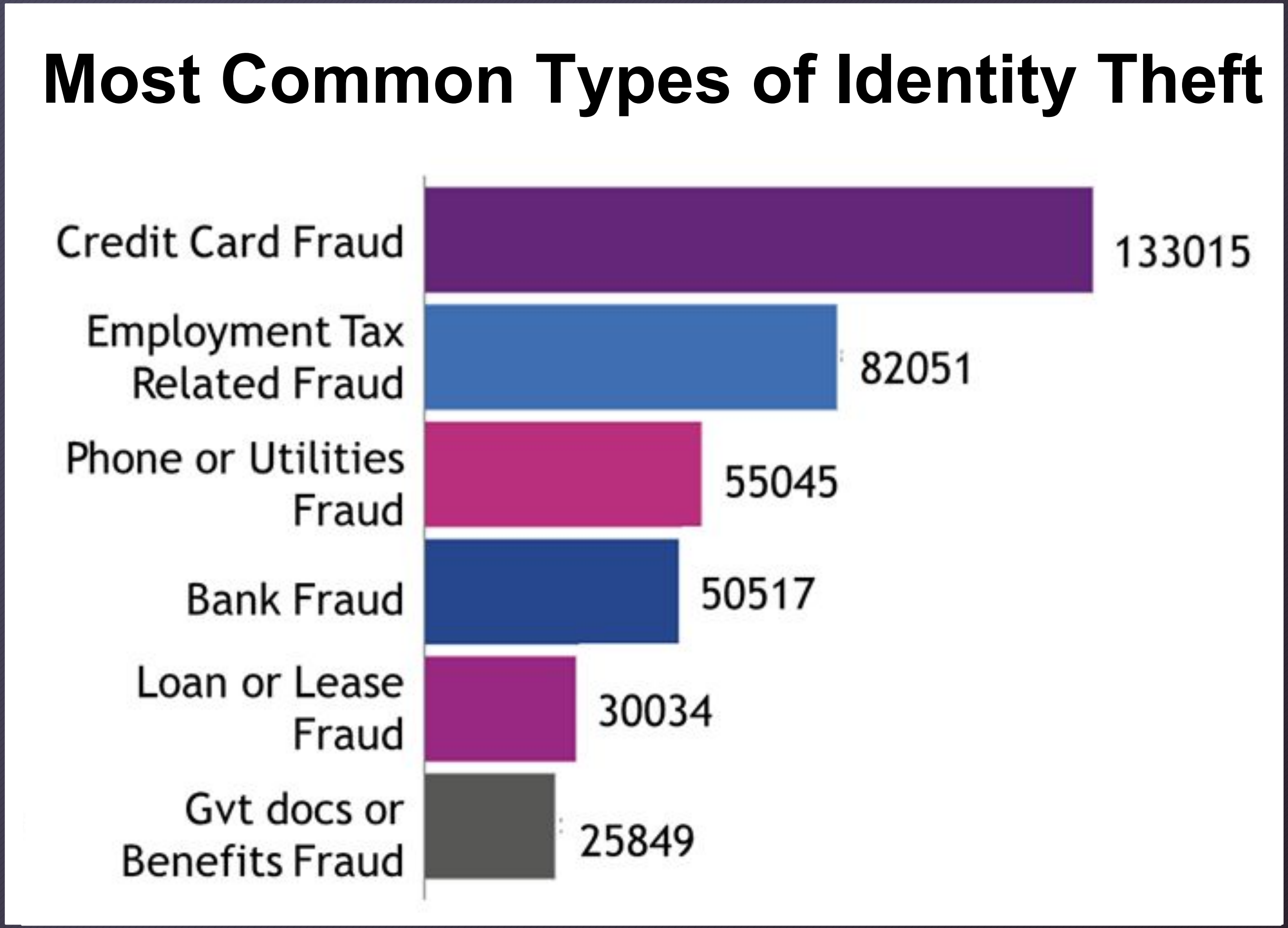
Team #3

Sean Ely, Xiwang Li, Pedram Tadayoni, Amir Rostami, Roxana Ruvalcaba, Arash Vahidnia

Credit Card Fraud Statistics

- Credit card fraud was most common form of identity theft in 2017 (133,015 reports) - *FTC 2017 Report*

- Credit card numbers exposed in 2017 = 14.2 million, up 88% over 2016 - *IRTC 2017 report*

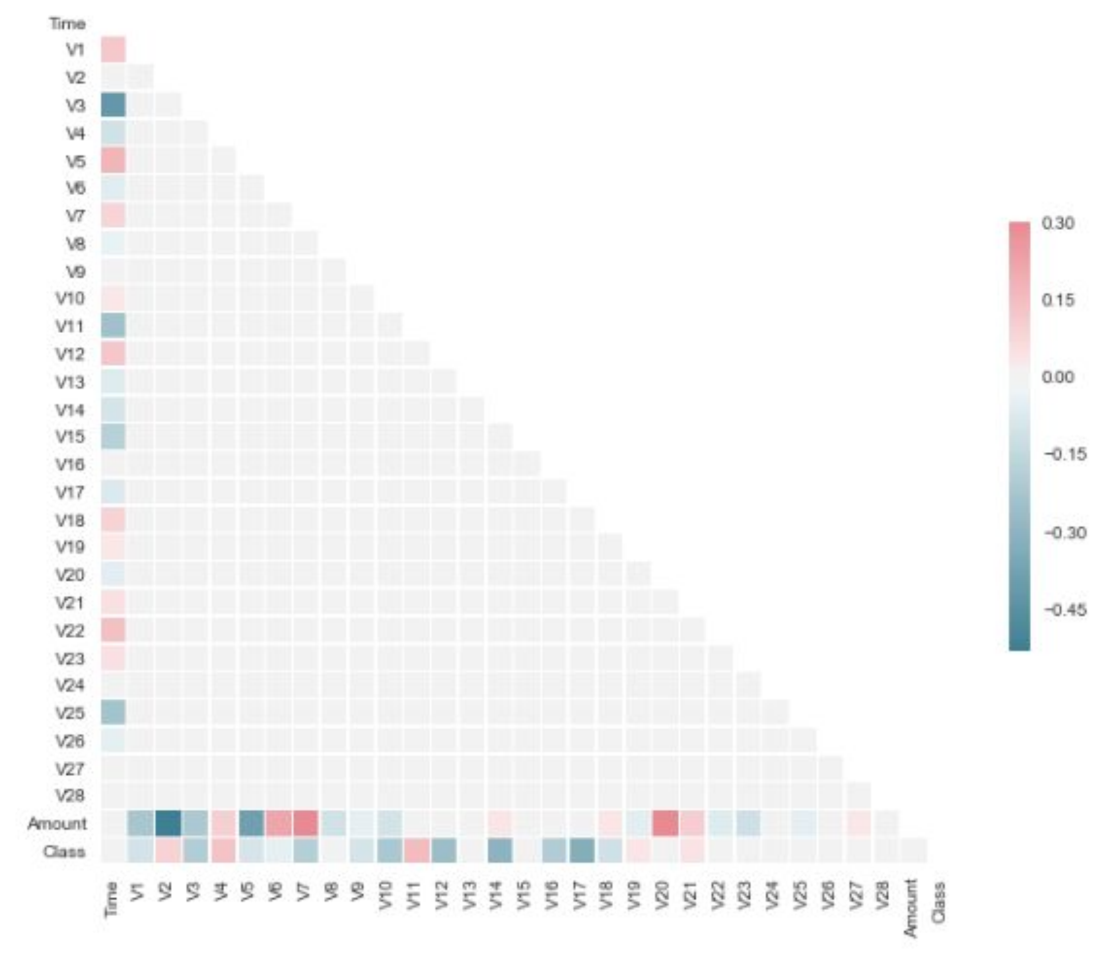


About the Data

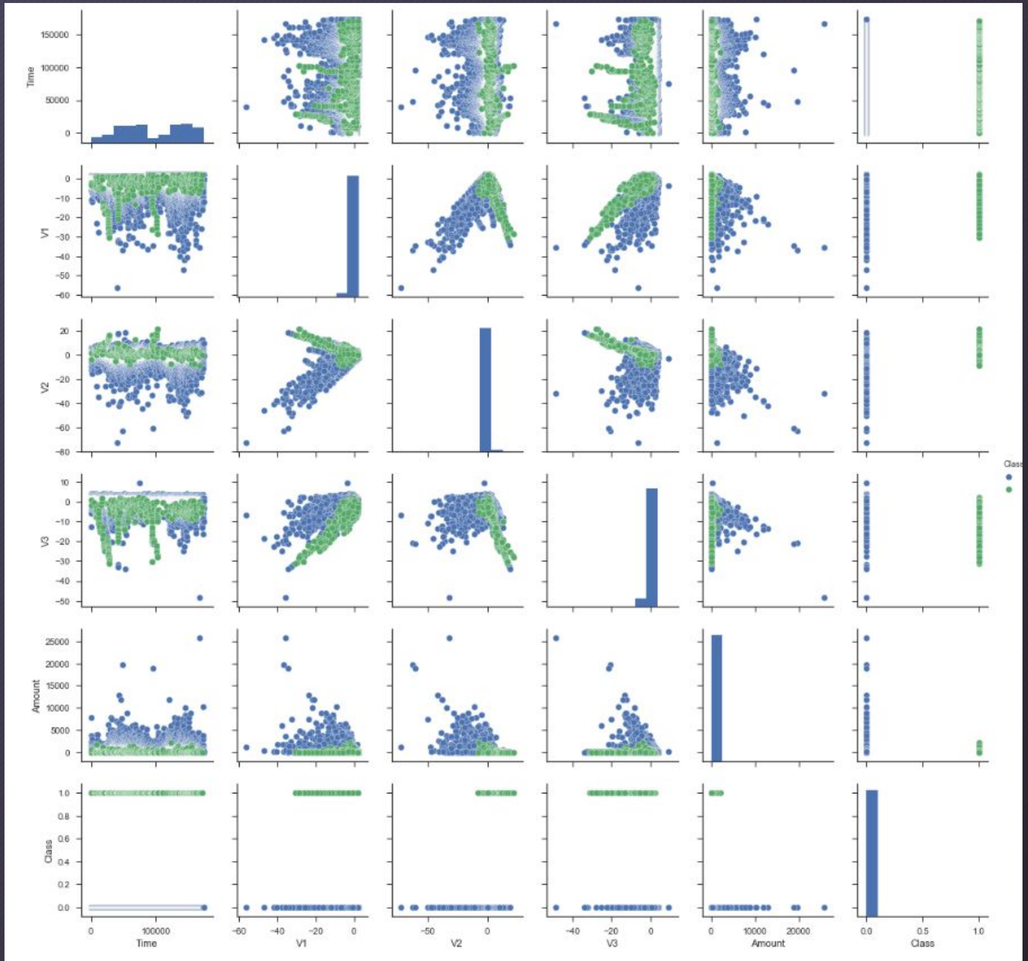
- Transactions made by credit cards in Sept 2013 by European cardholders over 2 days
- 284,807 transactions, 492 labeled as fraud (0.017% of all transactions)
- 31 variables including: ‘Time’, ‘Amount’ and 29 features already transformed using PCA to protect customer identities.
- No null values

Data Visualization

Covariance Plot



Pair Plot



Modeling Challenges and Solutions

Challenge

- **Unbalanced data**: 492 frauds out of 284,807 transactions (0.017% fraud)
- **Confidentiality**: no background available for 28 variables
- **Model**: Which predictive model to take

Approach

- Random under sample - Remove data from the majority class (no fraud)
- Random over sampling - augment the data set by replicating data from the minority class (fraud)
- Assessment metrics - confusion matrix, precision and recall will be used to evaluate the predictive modeling

Problems with Sampling Methods

Random Undersampling

- Removing examples from the majority class may cause the classifier to miss important concepts pertaining to the majority class.
  - Many false positives and thus poor precision

Random Oversampling

- Appending replicated data to the original data set creates multiple instances of certain examples and can lead to overfitting.
  - Training accuracy will be high, but the classification performance on unseen test data is generally far worse.

Modeling Without Sampling

Random Forest

- Test Accuracy = 99.94%
- Recall = 74.29%
- Precision = 88.89%

	Predicted No Fraud	Predicted Fraud
Actual No Fraud	85,322	13
Actual Fraud	36	104

Logistic Regression

- Test Accuracy = 99.91%
- Recall = 50%
- Precision = 87.5%

	Predicted No Fraud	Predicted Fraud
Actual No Fraud	85,325	10
Actual Fraud	70	70

Modeling Without Sampling

KNN

- Test Accuracy = 99.95%
- Recall = 79.59%
- Precision = 87.96%

	Predicted No Fraud	Predicted Fraud
Actual No Fraud	85280	16
Actual Fraud	30	117

Decision Tree

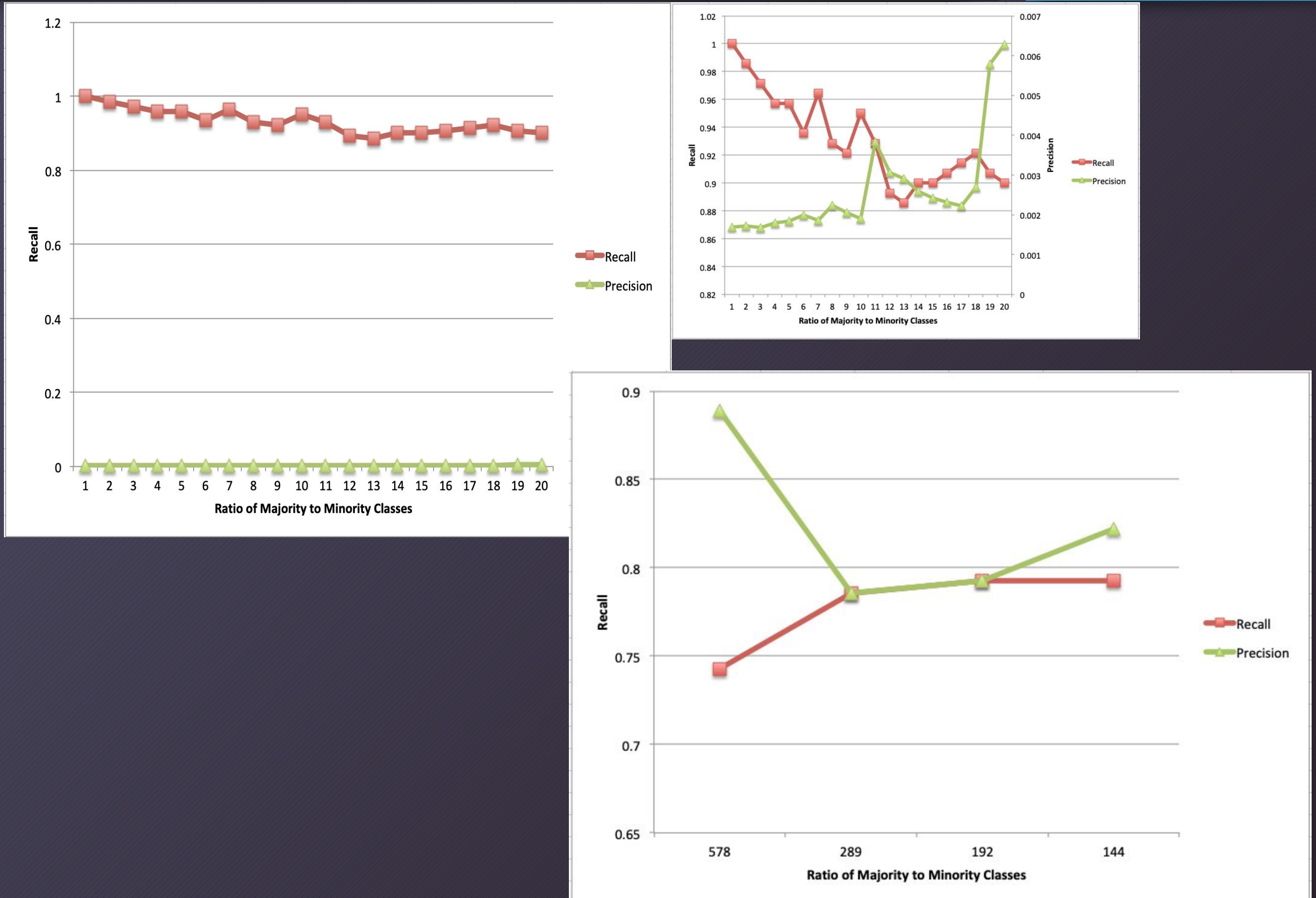
- Test Accuracy = 99.93%
- Recall = 75.51%
- Precision = 80.43%

	Predicted No Fraud	Predicted Fraud
Actual No Fraud	85269	27
Actual Fraud	36	111

Undersampling - Recall vs Precision

Using a undersampling and a Random Forest model, it was observed that by increasing the size of the majority class, Precision can be improved at the cost of decreasing recall. Note this model is not particularly useful for the following reasons:

1. Precision is very low .62% at best, meaning many customers will be notified of incorrect fraudulent activity.
2. Recall drops with rising precision, meaning that fraud is not detected.



Oversampling - SMOTE (Synthetic Minority Over-sampling Technique)

Random Forest

- Test Accuracy = 99.86%
- Recall = 90.44%
- Precision = 53.25%

	Predicted No Fraud	Predicted Fraud
Actual No Fraud	85,199	108
Actual Fraud	13	123

RF with Scaling

Scaled amount and dropped Time

- Test Accuracy = 99.75%
- Recall = 89.71%
- Precision = 37.65%

	Predicted No Fraud	Predicted Fraud
Actual No Fraud	85,105	202
Actual Fraud	14	122

Summary

- The KNN and Random Forest models showed good recall and precision compared to other algorithms (model without sampling).
- Undersampling showed great recall but horrible precision.
- Oversampling (SMOTE) showed better recall than unsampled models, with moderate precision.