Credit Card Fraud Detection

Jalal Kiani and Sharon Kwak

Situation

IEEE-CIS wants to improve the effectiveness of fraudulent transaction alerts

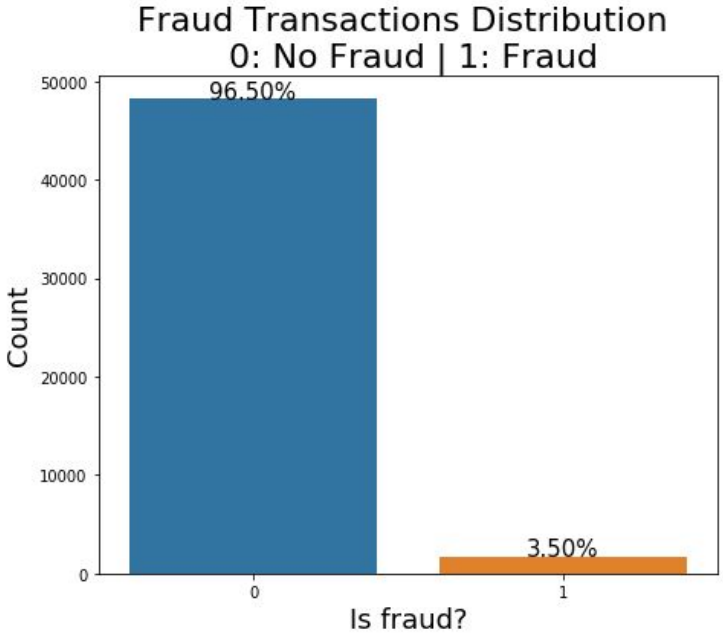
- save millions of dollars for consumers around the world
- help businesses reduce their fraud loss and increase their revenue

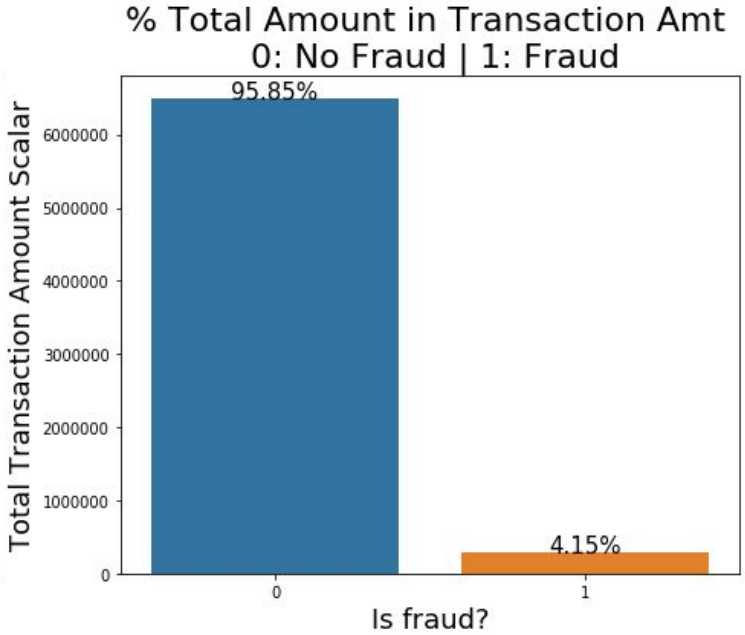
*IEEE-CIS: Institute of Electrical and Electronics Engineers Computational Intelligence Society

Data

- IEEE Fraud Detection Data (Kaggle) - used 10% of the data
- From Vesta's real-world e-commerce transactions
- Contains a wide range of features (434 total), some listed below:
  - Purchase card information (company, type, bank, country, etc.)
  - Device information (type, network connection, etc.)
  - Product bought in transaction
  - Transaction payment amount (USD)

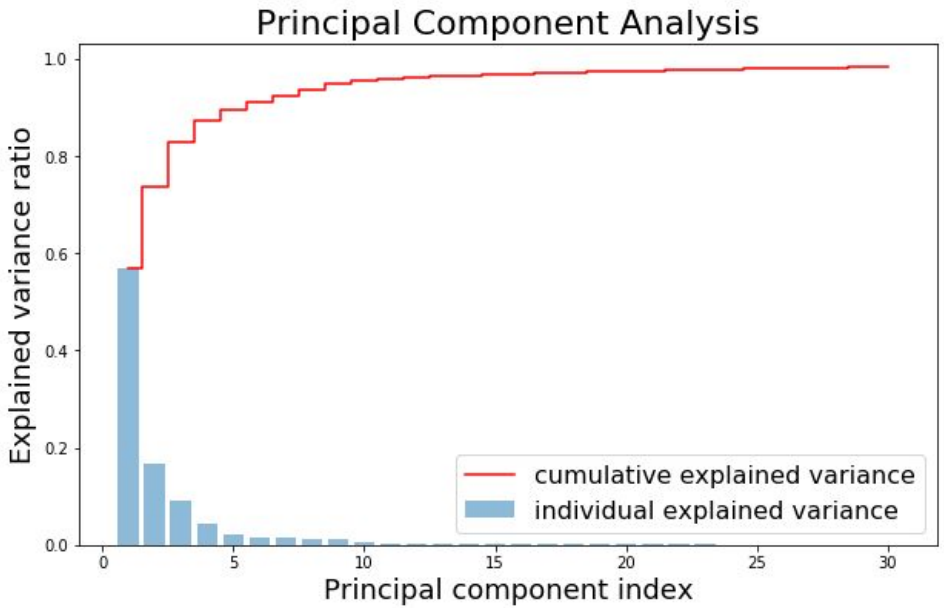
Fraud vs. No-Fraud Transactions





Feature Selection

- Extracted “Vesta-engineered features” to reduce the dimension
- Combines 339 features into 10 new features, dropping the least important variables but retaining the most valuable parts
- The 10 PCA features explain variance of 96% of the data



Methodology

Classification

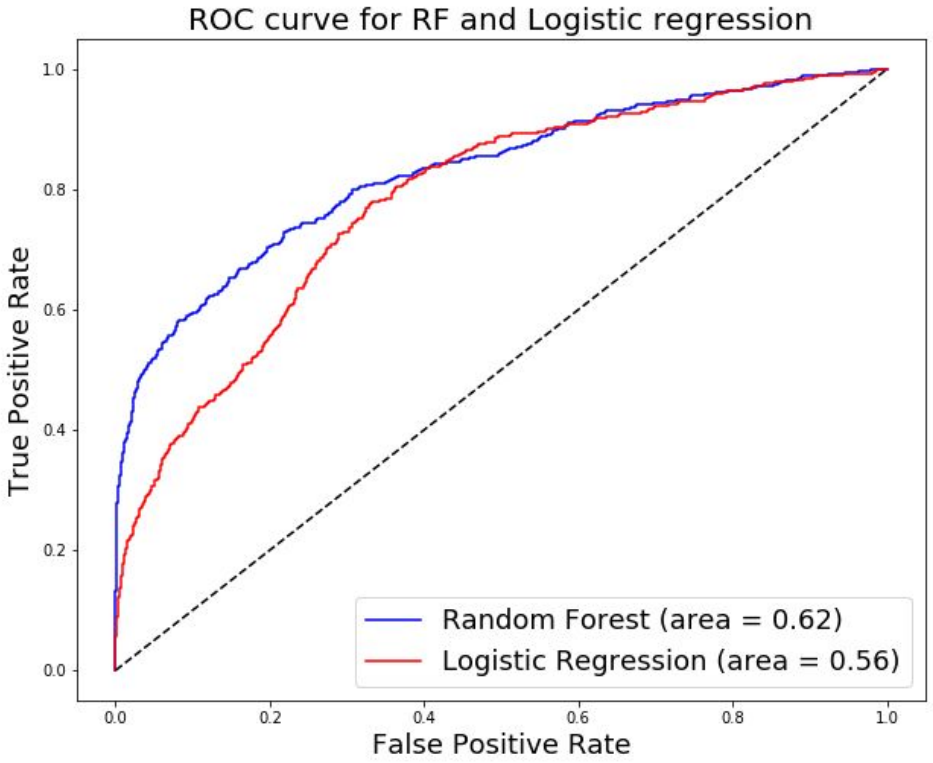
- Logistic Regression
- Random Forest

Metrics

- ROC (Receiver Operating Characteristic) Curve
- AUC score (area under the curve)

Logistic or Random Forest?

Metric	Logistic Regression	Random Forest
Accuracy	0.97	0.97
Precision	0.51	0.87
Recall	0.13	0.25
F1 score	0.2	0.39

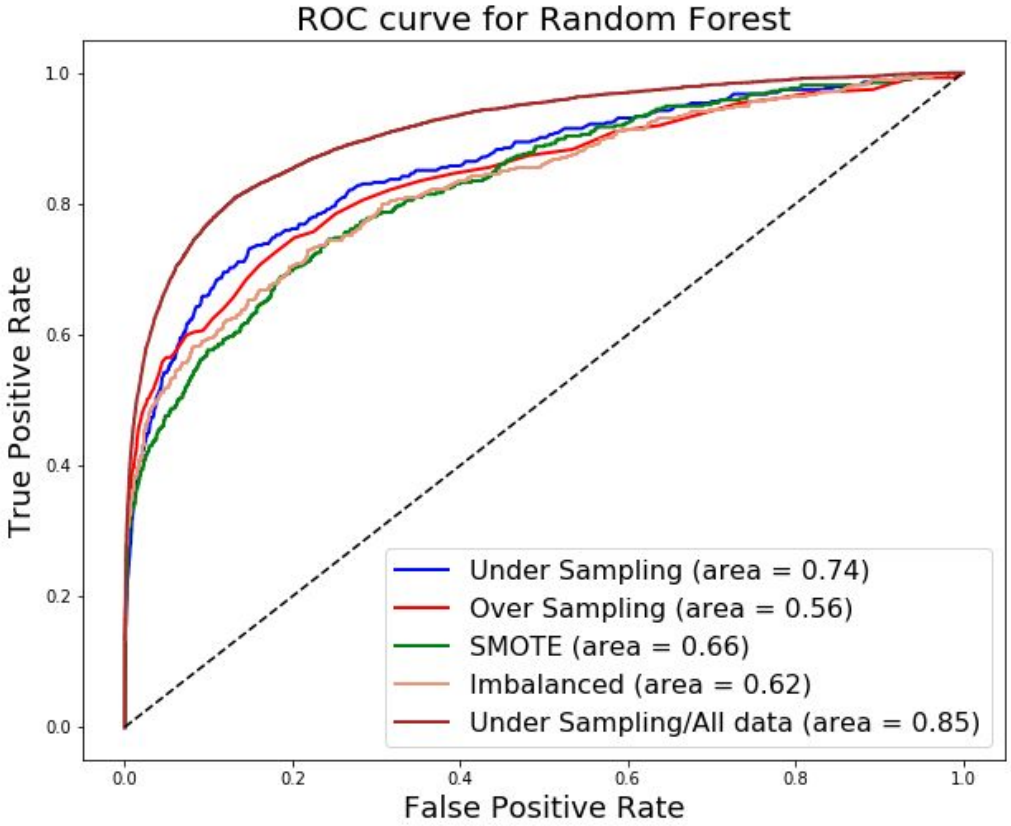


Dealing with Imbalanced Dataset

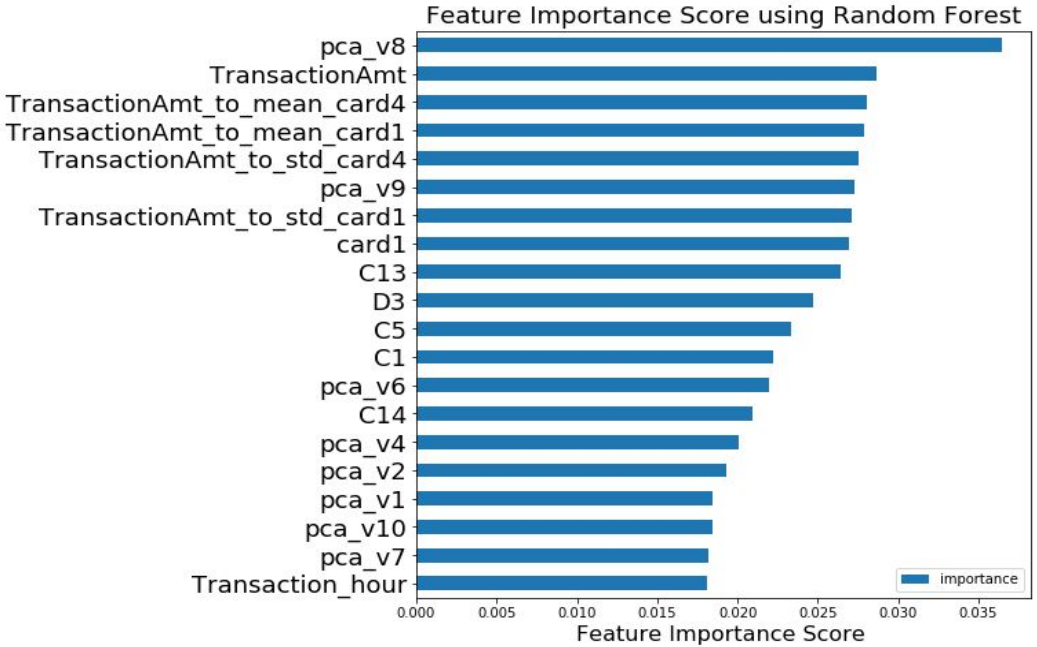
Oversampling
Randomly replicating “Fraud” data

Undersampling
Randomly removing “No-Fraud” data

SMOTE
Creating new, synthetic “Fraud” data using the K-Nearest Neighbors (KNN) algorithm



Top 20 Features in Predicting Credit Card Fraud



Conclusions

- Random Forest performed better than Logistic Regression
- Undersampling performed best

Next Steps

- Run the model on the whole dataset
- Using XGboost