Fraudulent Credit Card Transaction Detection

Min Luo
Department of Statistics and Data Science
University of Central Florida

Introduction

According to statistics in 2018

- Over 40 billions credit card transactions made in the US
  - $3.8 trillion in dollar volume
- Credit card fraud increased by 18.4 % in 2018
- Payment card fraud resulted in lost of $24.26 Billion worldwide

- Credit card companies need to identify fraudulent transactions quickly to mitigate the loss of credit card holders

Data

- **Kaggle Credit Card Fraud Detection Dataset**
- 2 days of credit cards transactions in September 2013 in Europe
- 284,807 observations, 31 variables
    - Target variable: **Class** 1 = Fraud, 0 = Normal
    - 30 predictor variables
        - **V1 to V28**: 28 variables masked via PCA due to privacy protection
        - **Time**: Number of seconds elapsed between this transaction and the first transaction in the dataset
        - **Amount**: transaction amount

Challenges

- Dataset is highly unbalanced
  - Only 0.1727% (n = 492) of the transactions are fraudulent
- Lots of outliers
- V1 to V28 masked
  - Limited ability to do feature engineering using domain knowledge

Data Exploratory Analysis

Class (1=fraud, 0=normal)

Highly imbalance data

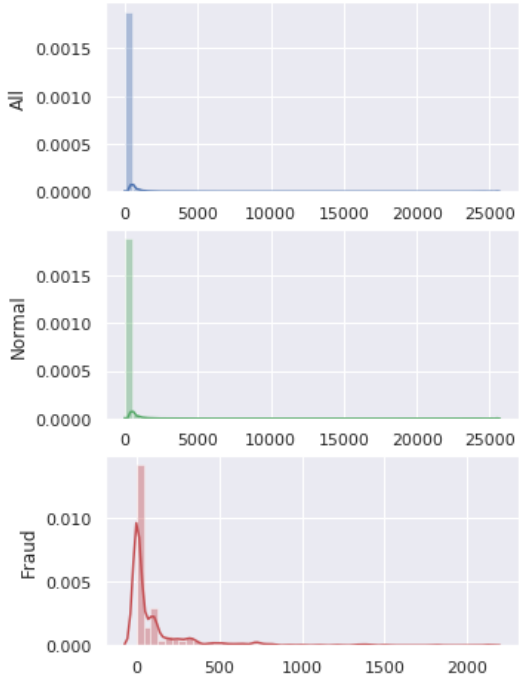
	Count
1 = Fraudulent	284,315 (99.8%)
0 = Normal	492 (0.17%)

Data Exploratory Analysis

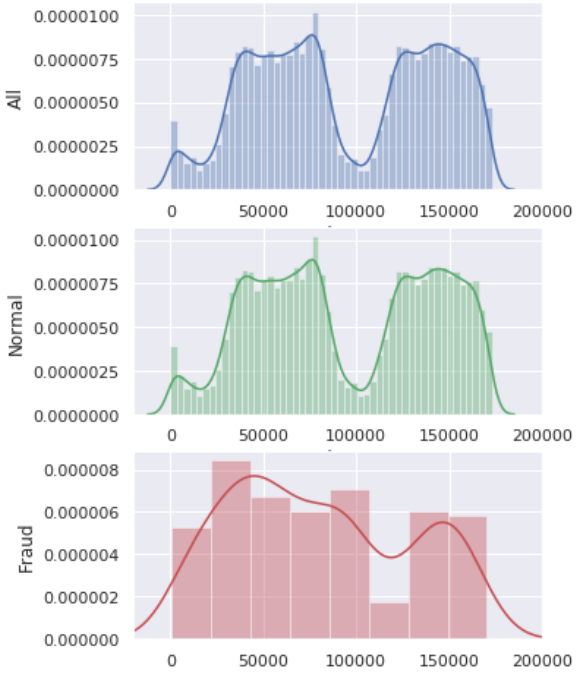
Amount, Time

Amount by Class

	Fraud	Normal
mean	122.211	82.291
std	256.683	250.105
min	0	0
max	2125.87	25691.16

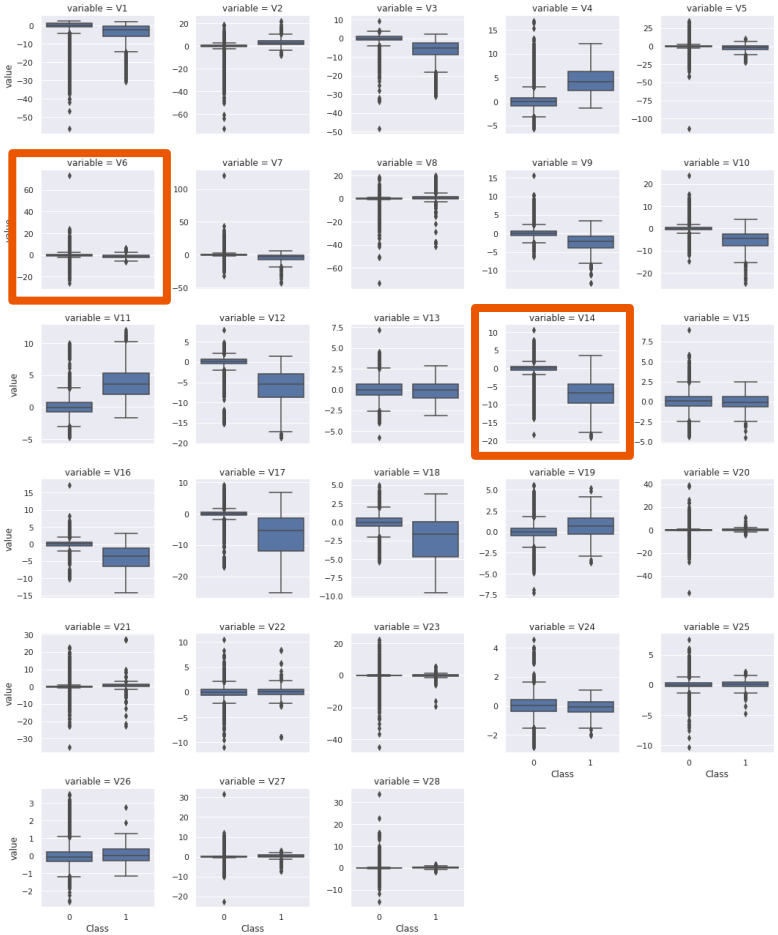


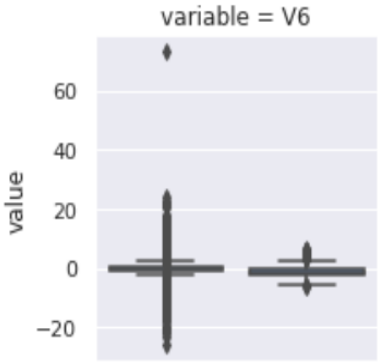
Amount

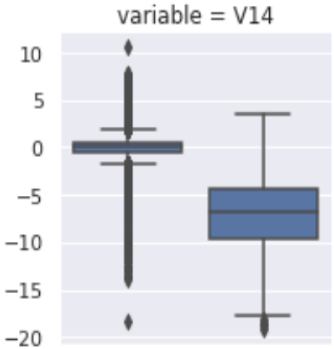


Time

Boxplot of masked variables V1 - V28 by class







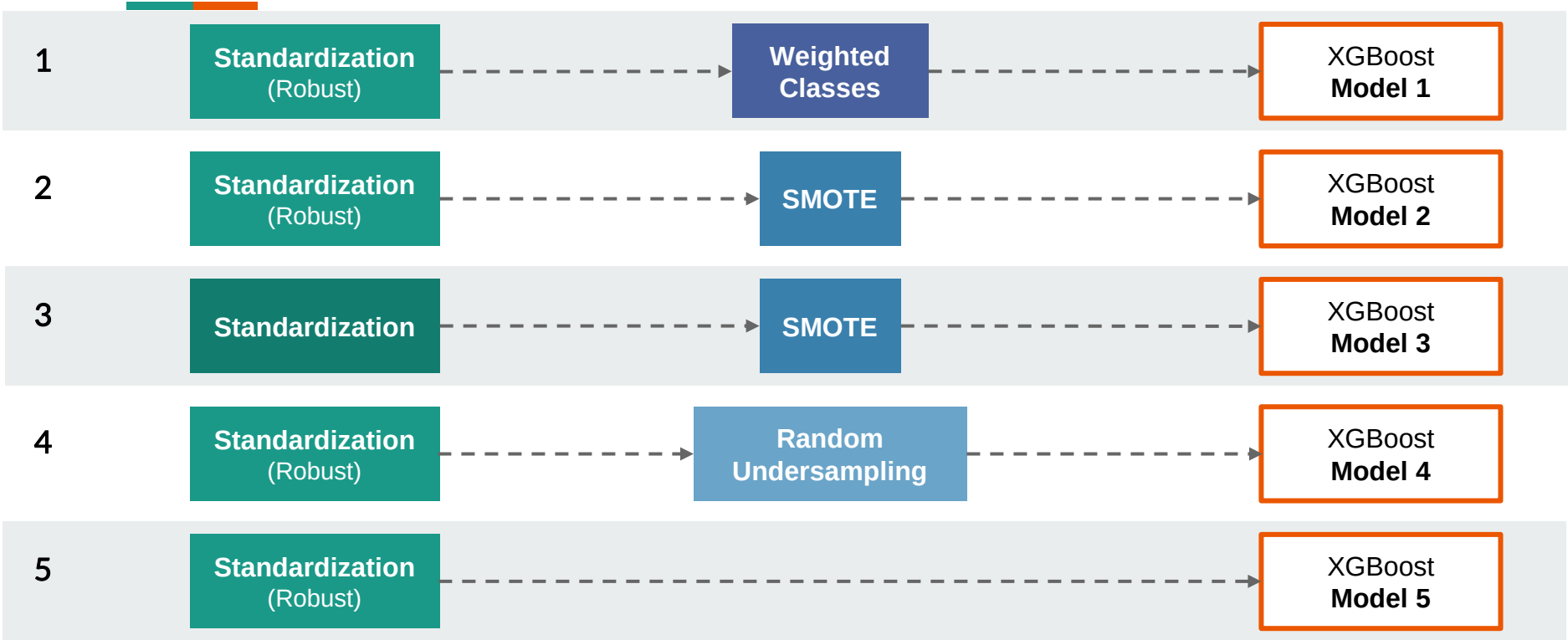
Models Selection

- XGBoost (e**X**treme **G**radient **Boost**ing)
  - Advanced implementation of gradient boosting machine (GBM)
  - Flexible model not bounded by structure of data
- Imbalance data
  - Rebalancing data via under- or over-sampling
  - Cost sensitive learning by assigning different cost/weight to each class

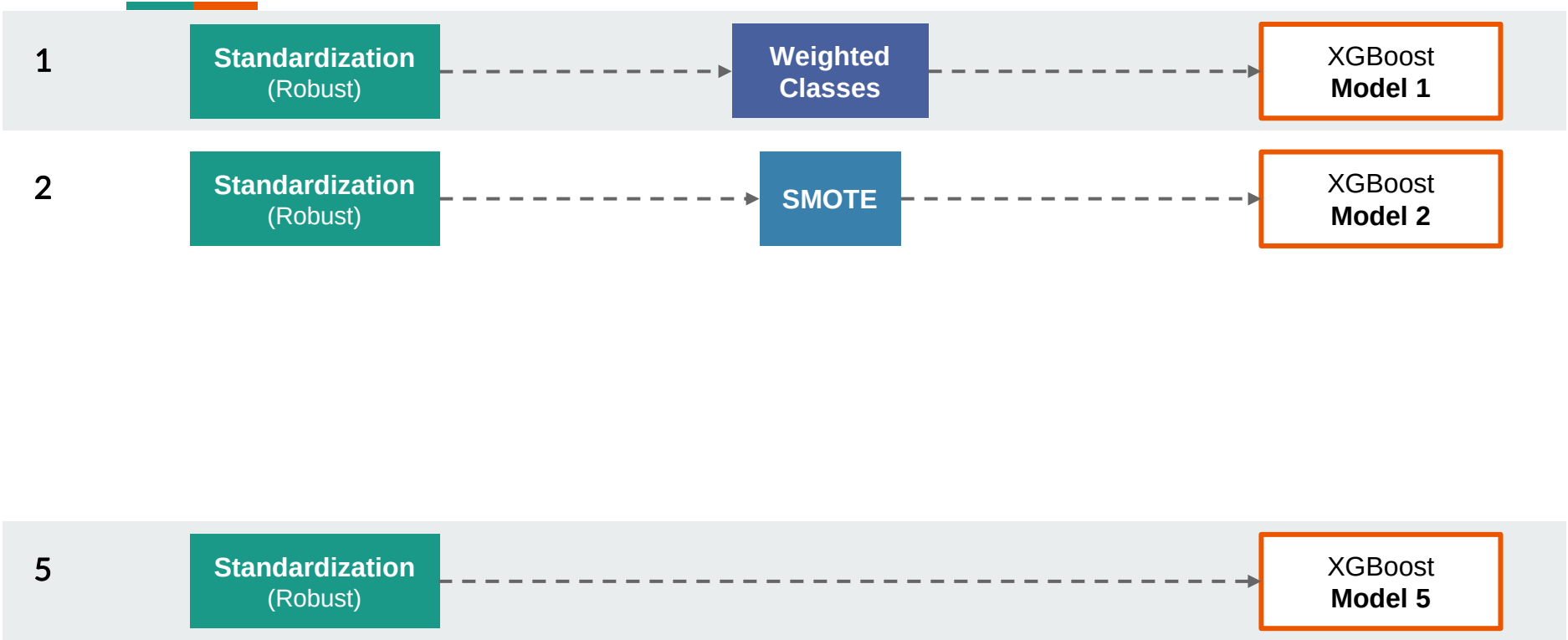
Models Selection

- Model “pipeline” using scikit-learn in python
  - **Round 1**: proof of concept/prototyping via 5-fold cross validation
  - **Round 2**: hyperparameter tuning via grid search 5-fold cross validation
- Metrics
  - Precision, recall, and , AUC

Model Selection Pipelines Round 1: Prototyping (5-fold CV)



Model Selection Pipelines Round 2: Hyperparameter Tuning (Grid Search 5-fold CV)



Result

		Model 1 Weighted Class	**Model 2** SMOTE	**Model 5** No Re-balance
AUC	Train	**0.9989**	0.9988	0.9863
	Test	**0.9711**	0.9726	0.9644
Precision	Train	0.9470	**0.9785**	0.8343
	Test	0.8317	**0.8730**	0.7758
Recall	Train	0.8599	0.3058	**0.9788**
	Test	0.6445	0.1967	**0.8340**

Summary

- An xgboost model with carefully tuned hyperparameters out performed other xgboost models with re-balancing
- scikit-learn does not support GPU
    - To fully take advantage of xgboost, there are other libraries that support that
- In terms of application, a fraud detection model should focus on minimizing the false negative rate rather than increasing the accuracy

Questions?

Sources Cited

1. Altini, Marco. Dealing with Imbalanced Data: Undersampling, Oversampling and Proper Cross-Validation. 17 Aug. 2015, https://www.marcoaltini.com/blog/dealing-with-imbalanced-data-undersampling-oversampling-and-proper-cross-validation.
2. Analytics Vidhya. (2019). How to handle Imbalanced Classification Problems in machine learning?. [online] Available at: https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/ [Accessed 24 Nov. 2019].
3. Brownlee, Jason. “A Gentle Introduction to XGBoost for Applied Machine Learning.” Machine Learning Mastery, 21 Aug. 2019, https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/.
4. Credit Card Fraud Statistics. (2019, October). Retrieved November 2019, from https://shiftprocessing.com/credit-card-fraud-statistics/.
5. Holmes, Tamara E. “Credit Card Market Share Statistics.” CreditCards.com, 12 Sept. 2019, www.creditcards.com/credit-card-news/market-share-statistics.php.
6. “Introduction to Boosted Trees.” Introduction to Boosted Trees - Xgboost 1.0.0-SNAPSHOT Documentation, https://xgboost.readthedocs.io/en/latest/tutorials/model.html.
7. Machine Learning Group. “Credit Card Fraud Detection.” Kaggle, 23 Mar. 2018, www.kaggle.com/mlg-ulb/creditcardfraud
8. Mishra, Satwik. Handling Imbalanced Data: SMOTE vs. Random Undersampling. International Research Journal of Engineering and Technology, Aug. 2017, https://www.irjet.net/archives/V4/i8/IRJET-V4I857.pdf.

Image
“thief” by By Adrien Coquet/ CC BY