Data Science Bootcamp

Project 2
Transaction Fraud Detection

Marwa Salah

Project Overview

Identify potential fraudulant activity (transactions) using ML based on existing card transaction data

Anonymized credit card transactions labeled as fraudulent or genuine
Transformed using PCA
2 Features were not transformed using PCA which are Time & Amount

k
source

Due to nature classes are highly unbalanced.

0.17% vs 99.8%
Class Balance

Transaction Faud Detection

WHY?

1B+

Transactions per Day

11K+

Transactions per Second

~19 Fradulant/sec

$2-3

Cost of each dollar lost to fraud

$ 29B

Fraud Losses 2019

$ 35B

Expected Fraud Losses 2022

Impacts

Individuals | Businesses | Financial Instituations

https://fortunly.com/statistics/cash-versus-credit-card-spending-statistics/#gref

https://www.ncr.com/blogs/payments/credit-card-fraud-detection

Objectives

Identify potentially fraudulent activity

Detecting fraudulent activity before those transactions are even completed

Primary

Minimize Accepted Fraudulant Transactions

FN

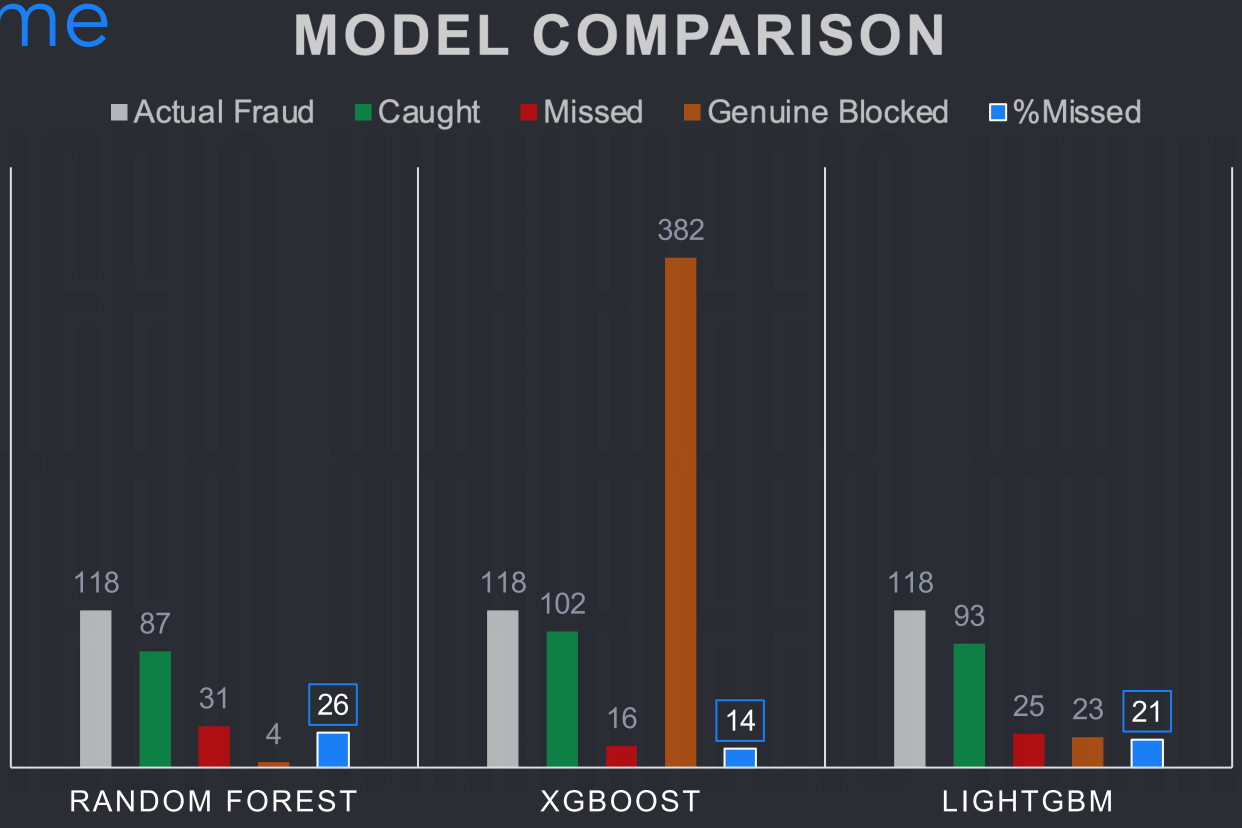
Secondary

Avoid Declining Genuine Transactions

FP

Outcome

~71k
#Transactions Tested



Conclusion & Next Step

Full Dataset

Fraud: $ 60K
Genuine: $ 25M

@86% Detection Rate →

Fraud: $ 9K
Genuine: $ 25M

Example based cost-sensitive learning & Cost–dependant Classification

	Actual Fraud $y_{true} = 1$	Actual Legitimate $y_{true} = 0$
Predicted Fraud $y_{pred} = 1$	*True Positive* $cost_{TP} = \text{Admin}$	*False Positive* $cost_{FP} = \text{Admin}$
Predicted Legitimate $y_{pred} = 0$	*False Negative* $cost_{FN} = \text{Transaction}$	*True Negative* $cost_{TN} = \$0$

https://towardsdatascience.com/fraud-detection-with-cost-sensitive-machine-learning-24b8760d35d9

Challenges

- Annomyzation of Data
- Diversity of Data

- Time
- Compute Resources

Technical Outcome Summary

Outcome Summary

	Model	precision	recall	f1-score
0 Genuine	Random Forest	1	1	1
	XGBoost	1	0.99	1
	LightGBM	1	1	1
1 Fraud	Random Forest	0.96	0.74	0.83
	XGBoost	0.21	0.86	0.33
	LightGBM	0.83	0.80	0.81

Confusion Matrix

74%
Recall

96%
Precision

True Label	0 Genuine	70809	5	70814
	1 Fraud	31	87	118
		0 Genuine	1 Fraud	
		Predicted Label		

Random Forest

Confusion Matrix

86%
Recall

21%
Precision

XGBoost

True Label		0 Genuine	1 Fraud	
	0 Genuine	70424	390	70814
	1 Fraud	16	102	118

Predicted Label

Confusion Matrix

79%
Recall

80%
Precision

LightGBM

True Label		Predicted 0 Genuine	Predicted 1 Fraud	
	0 Genuine	70795	19	70814
	1 Fraud	24	94	118

Predicted Label

Project Description

- In this project we were initially tasked with selecting a couple of datasets for analysis that meet the following rules:
    - Available to use
    - Suitable for use in a professional environment
    - Does not contain any personal information
    - Not used for any assignment, lecture or task from this environment

- My Proposed Sets: Credit Card Transaction Data & Financial Sales Data. Former was approved.

- So far, the project covers the following progress:
    - Data Cleaning & EDA
    - Evaluating & Tuning 3 Different Models (RandomForest, XGBoost, LightGBM)
    - Testing Impact of example-based cost sensitive learning
    - Model performance comparison

Data Description

Anonymized credit card transactions labeled as fraudulent or genuine
Transformed using PCA
Contains transactions made by European credit card holders in Sep 2013
2 Features were not transformed using PCA which are Time & Amount

Data source is kaggle.

Due to nature classes are highly unbalanced.

Data Shape is:

- 28 Components, 2 Features & 1 Label
- 284807 Rows (Observations)

Data Types:
All features are Integers of type int64

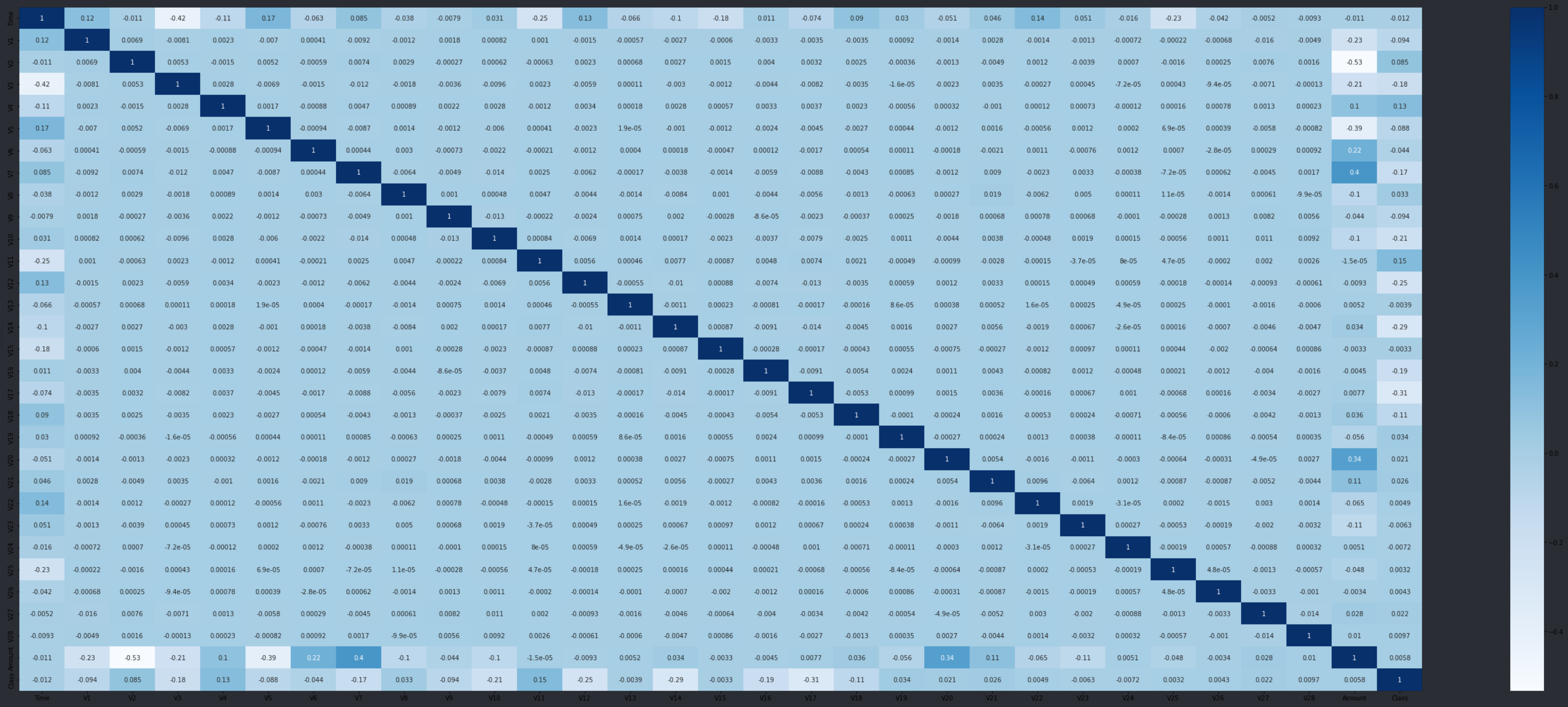
ML Problem Type is:
Classification

Classification Target:
Feature 'Class' is the response variable, and it takes value 1 in case of fraud and 0 otherwise.

Data Cleaning

- Due to PCA preprocessing Data Cleaning needed was very limited.
- No Nulls found
- No Data Type Inconcistencies
- No Categorical Inconsistencies would exist
- No Illogical Values
- A Number of Duplicates were there (~1000 rows) which were dropped

Visuals



Time
V1
V2
V3
V4
V5
V6
V7
V8
V9
V10
V11
V12
V13
V14
V15
V16
V17
V18
V19
V20
V21
V22
V23
V24
V25
V26
V27
V28
Amount
Class
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Outcome

~71k
#Transactions Tested