Background

Fraud losses incurred by banks and merchants on all credit, debit and prepaid general purpose payment cards issued globally hit **$21.8B** in 2015.

US account s for **38.7%** of the total at **$8.5B**.

By 2020, the US fraud losses due to card fraud could surpass **$12B**, indicating a **45%** growth rate.

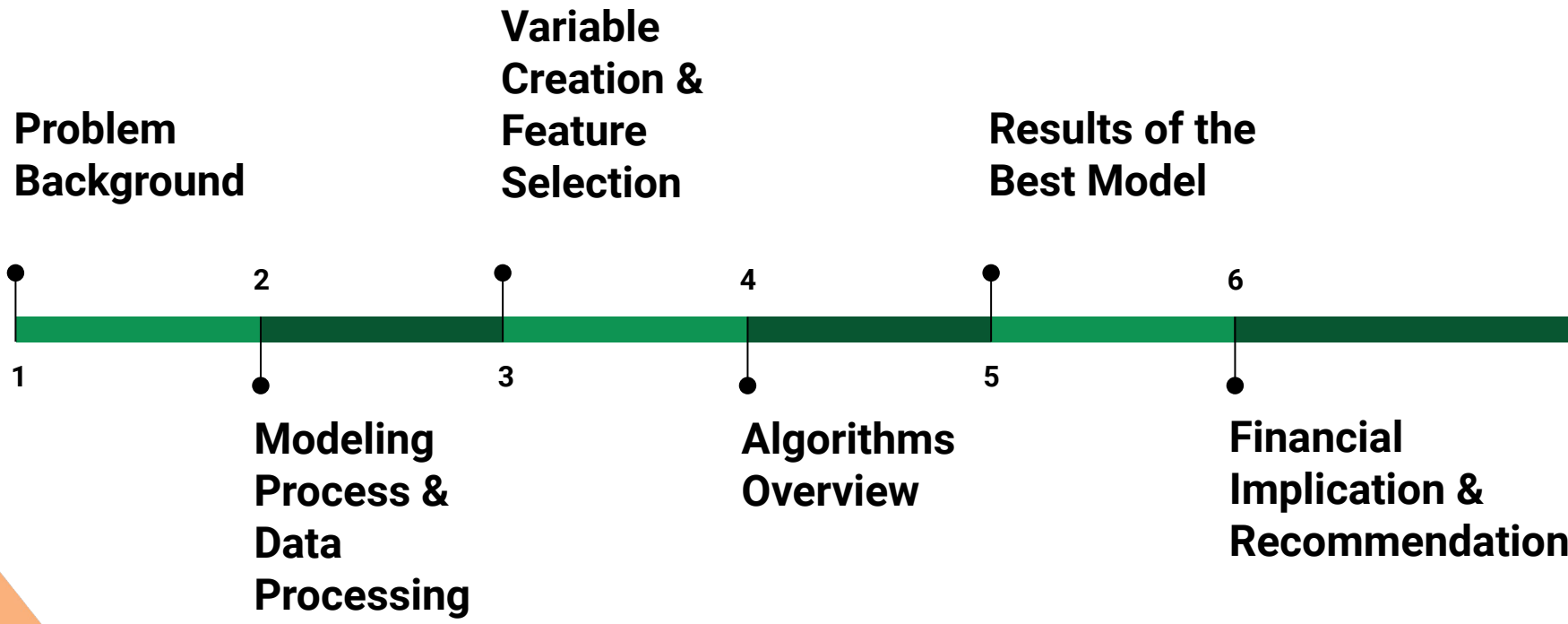
Resource: Forbes

Credit Card Transaction Fraud

Team 5:

Rui Xin Wu, Zhuolin Ouyang, Xiaoxuan Feng, Zimeng Cao, Feifei Shao, Zihao Geng, Zijian Wang

Agenda


Problem Background
Variable Creation & Feature Selection
Results of the Best Model
1
2
3
4
5
6
Modeling Process & Data Processing
Algorithms Overview
Financial Implication & Recommendation

Five people, 62 credit cards, and a nationwide fraud scheme stopped by Milwaukee investigators

Ashley Luthern, Milwaukee Journal Sentinel Published 6:00 a.m. CT April 3, 2018 | Updated 12:01 p.m. CT April 3, 2018

(Photo: Milwaukee County Office)

Buffalo woman convicted of credit card fraud sentenced

By Ha

Recommend 210 3

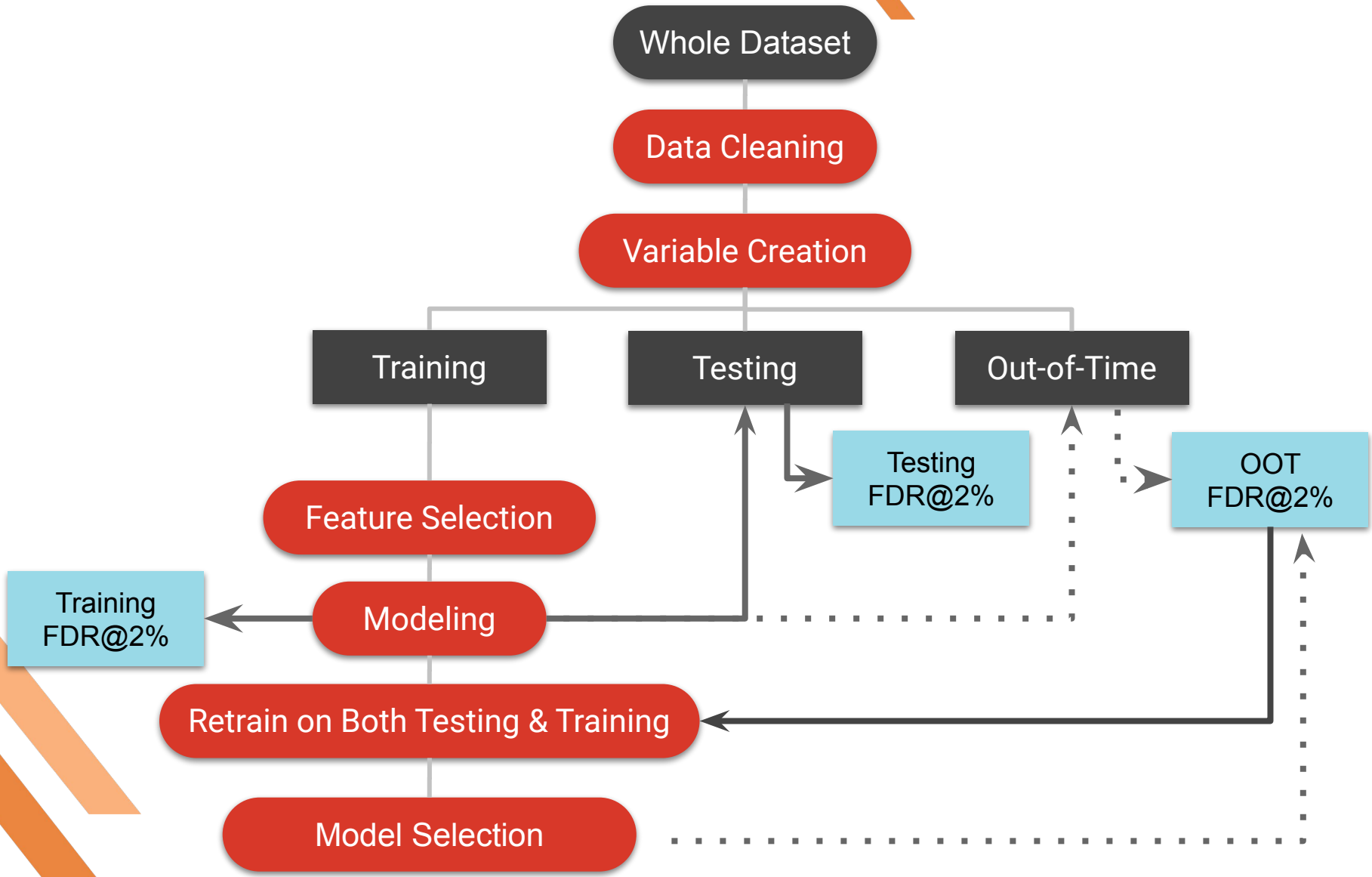
Local Businessman Sentenced for Credit Card Fraud and ID Theft

Posted: Apr 18, 2018 12:11 AM PDT
Updated: Apr 18, 2018 12:11 AM PDT

CORPUS CHRISTI, Texas - The owner of American Auto Pros in Corpus Christi has been ordered to federal prison after admitting he used numerous credit card accounts belonging to others in order to obtain things of value without authorization, announced U.S. Attorney Ryan K. Patrick. Antonio Arteaga, 38, of Corpus Christi, pleaded guilty Nov. 2, 2017, to one count of credit card fraud and one count of aggravated identity theft.

Antonio Arteaga was arrested by Corpus Christi police officers back on August 18th.

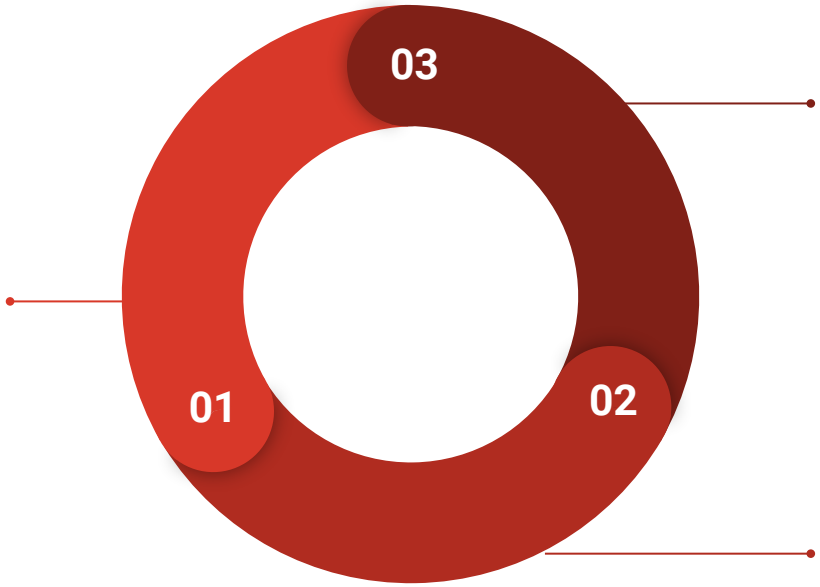
Modeling Process



Data Processing

Filtering out the irrelevant data

Excluding transaction type A,D and Y, **using only transaction type P**



Removing the outlier

Removing Record 52594, whose amount was erroneously recorded in Mexican Peso

Dealing with missing values

- Filling “blank” into **MerchantState** and **MerchantZip**
- Filling serial number from 1 to 3198 into the 3198 missing values in **Merchantnum**

Expert Variables

Transaction Frequency

eg. the transaction frequency of a specific card within the last n days

Purchasing Pattern

eg. how many different merchants a card was used for transaction within the last n days

Location

eg. the difference in merchant's zips between the current transaction of a card and its last transaction

Seasonality

eg. how common it is for a transaction to happen on a certain day of week

Profiling

eg. the ratio of the current amount of a specific card to its past n-day average amount, excluding the past fraudulent transactions

102 Variables

Feature Selection

Methodology

Kolmogorov-Smirnov Test

measures how separate is one distribution from the other distribution.

- Independent of any modeling methods
- can be easily implemented to remove the obvious insignificant variables
- Faster computation time

- A total of 63 variables are selected.

Feature Selection Results

Top 10 Variables

No.	Variable	KS Score
1	Amount	50.62
2	amount_sum_28	47.17
3	amount_avg_28	44.03
4	amount_max_28	43.37
5	merch_amount_avg_28	43.31
6	merch_amount_median_28	42.98
7	amount_sum_14	42.74
8	merch_amount_median_14	41.13
9	merch_amount_avg_14	41.06
10	amount_max_14	40.43

Algorithms Overview

	Train	Test	OOT
Logistic Regression	50.6%	54.1%	53%
Neural Network	87.62%	81.82%	68.93%
Random Forest	91.0%	92.8%	74.6%
XGBoost	99.82%	91.97%	73.67%
SVM	89.6%	85.8%	71.8%

Random Forest

Key Characteristics:

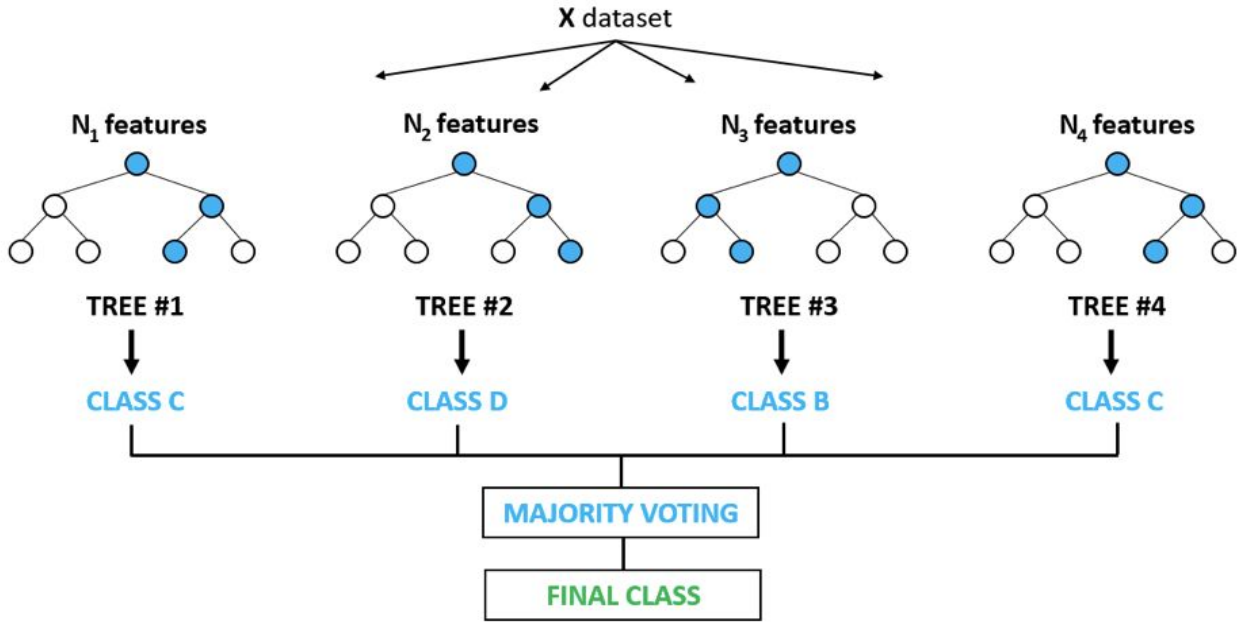
1. Based on Classification Decision Trees
2. The method combines Breiman's "bagging" idea and the random selection of features.

FDR@2% Population:

- Training: 91.0%
- Testing: 92.8%

-----Retrained the model on both Training and Testing Dataset and Predicted on OOT set-----

- OOT: 74.56% (Caught all 252 Frauds for top 2% of the OOT dataset)



Random Forest

Why we chose Random Forest

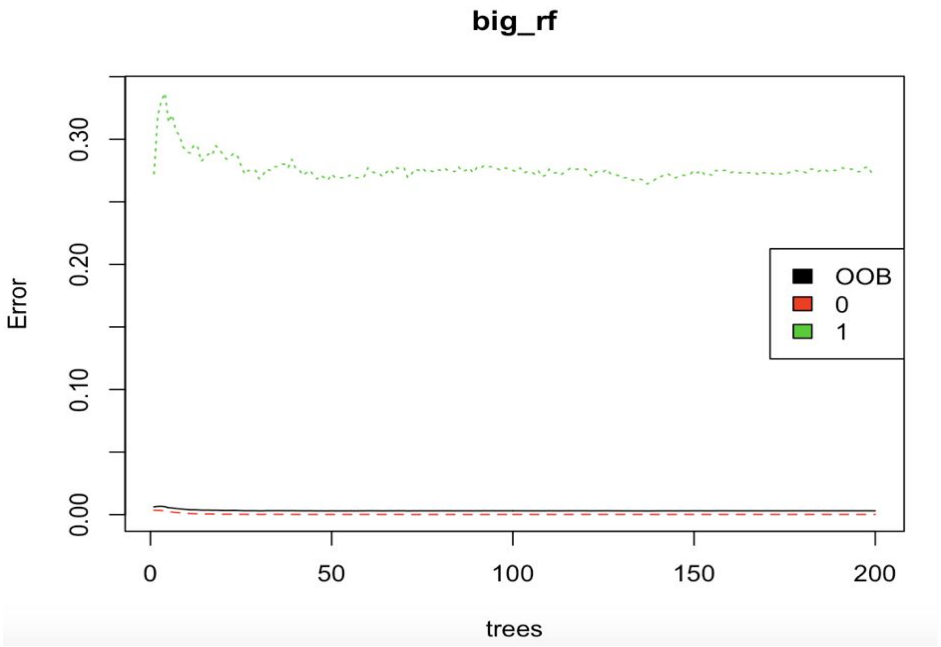
1. Randomness will prevent the model from overfitting.
2. Randomness will make the model robust, not suffering too much noise.
3. Random Forests can take in high dimension dataset and it can output the importance index for each variable.
4. Random Forests do not need cross validation for the performance measure since it will generate the OOB (Out-of-Bag) Estimation.

How we perform Random Forest

Input Variables = 63

Chose Mtree = 50

Chose Mtry = 14

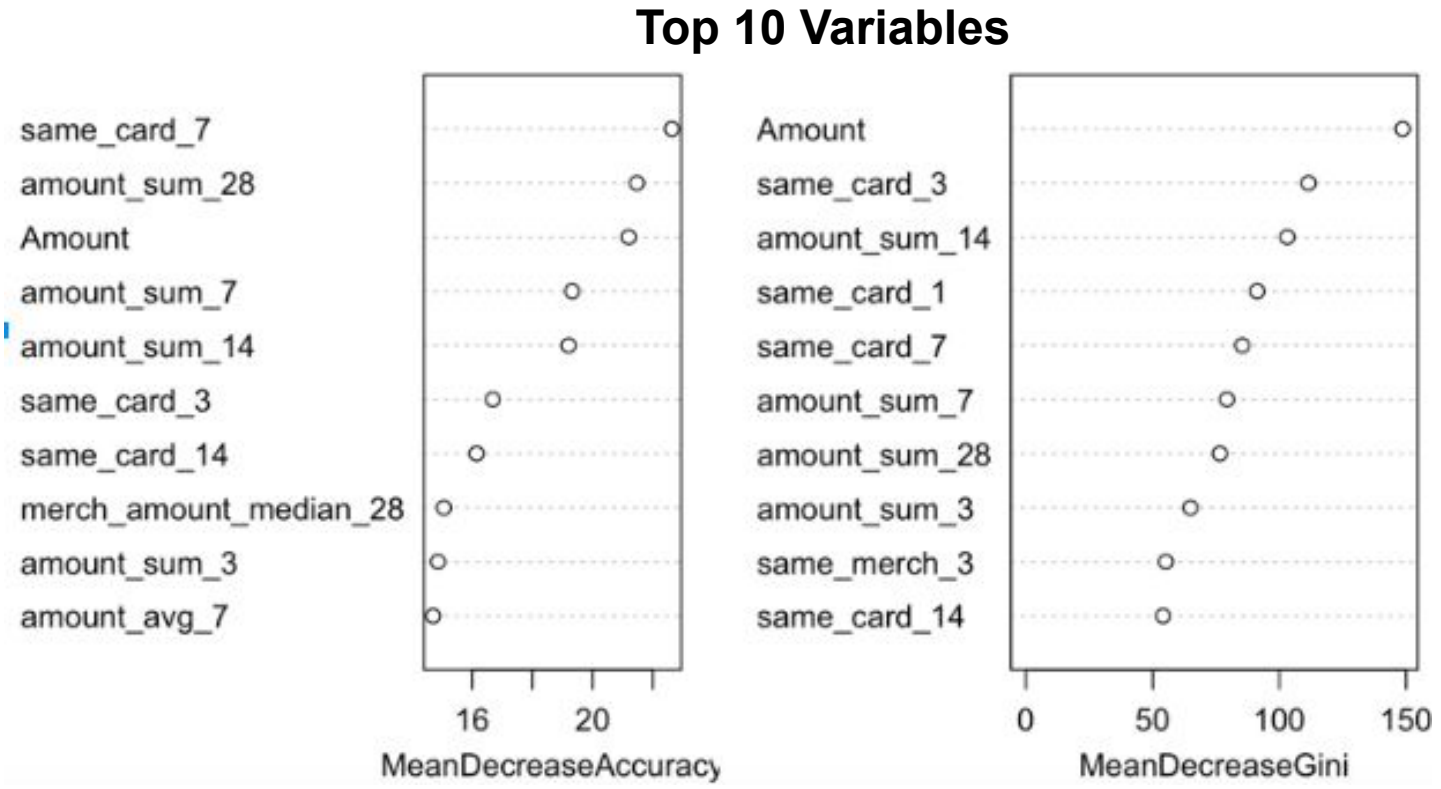


Random Forest FDR Table

The model captures 74.56% fraudulent records at 2% population, which is also the maximum it could achieve at that cutoff. With 0.5% precision, it captures all the fraud at 3% population.

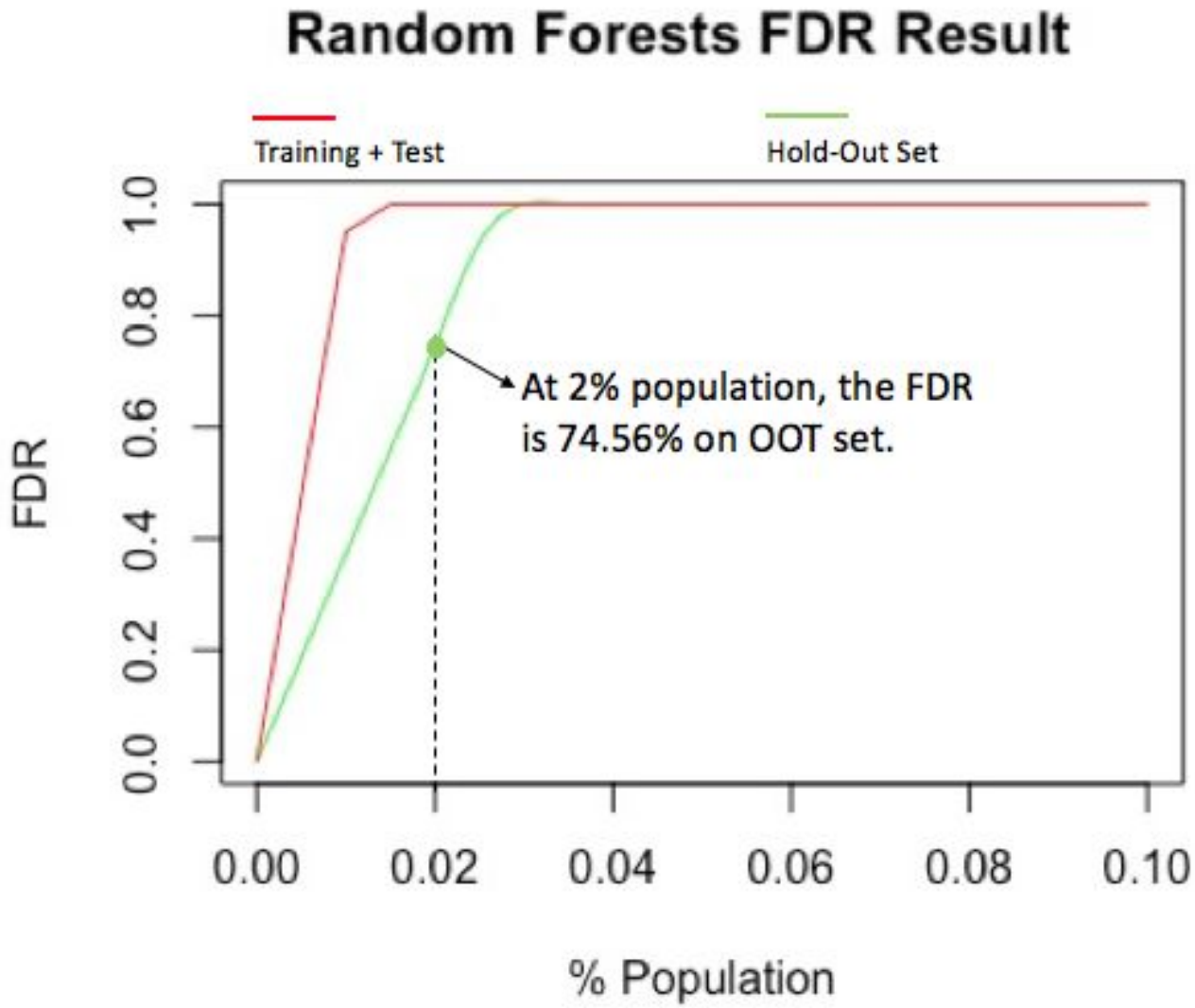
Overall Bad rate is 2.69%	Bin Statistics					Cumulative Statistics					
Population Bin%	Total # Record	#Good	#Bad	%Good	%Bad	Cml # Good	Cml # Bad	Cml % Good	Cml % Bad(FDR)	KS	False Positive Ratio
0.5	63	0	63	0	100	0	63	0	18.64	18.64	0
1	63	0	63	0	100	0	126	0	37.28	37.28	0
1.5	63	0	63	0	100	0	189	0	55.92	55.92	0
2	63	0	63	0	100	0	252	0	74.56	74.56	0
2.5	63	0	63	0	100	0	315	0	93.2	93.2	0
3	63	40	23	63.49	36.51	40	338	0.33	100	99.67	0.12
3.5	63	63	0	100	0	103	338	0.84	100	99.16	0.3
4	62	62	0	100	0	165	338	1.35	100	98.65	0.49
4.5	63	63	0	100	0	228	338	1.86	100	98.14	0.67
5	63	63	0	100	0	291	338	2.38	100	97.62	0.86
5.5	63	63	0	100	0	354	338	2.89	100	97.11	1.05
6	63	63	0	100	0	417	338	3.4	100	96.6	1.23
6.5	63	63	0	100	0	480	338	3.92	100	96.08	1.42
7	63	63	0	100	0	543	338	4.43	100	95.57	1.61
7.5	63	63	0	100	0	606	338	4.95	100	95.05	1.79
8	63	63	0	100	0	669	338	5.46	100	94.54	1.98
8.5	63	63	0	100	0	732	338	5.98	100	94.02	2.17
9	63	63	0	100	0	795	338	6.49	100	93.51	2.35
9.5	63	63	0	100	0	858	338	7.01	100	92.99	2.54
10	63	63	0	100	0	921	338	7.52	100	92.48	2.72

Variable Importance



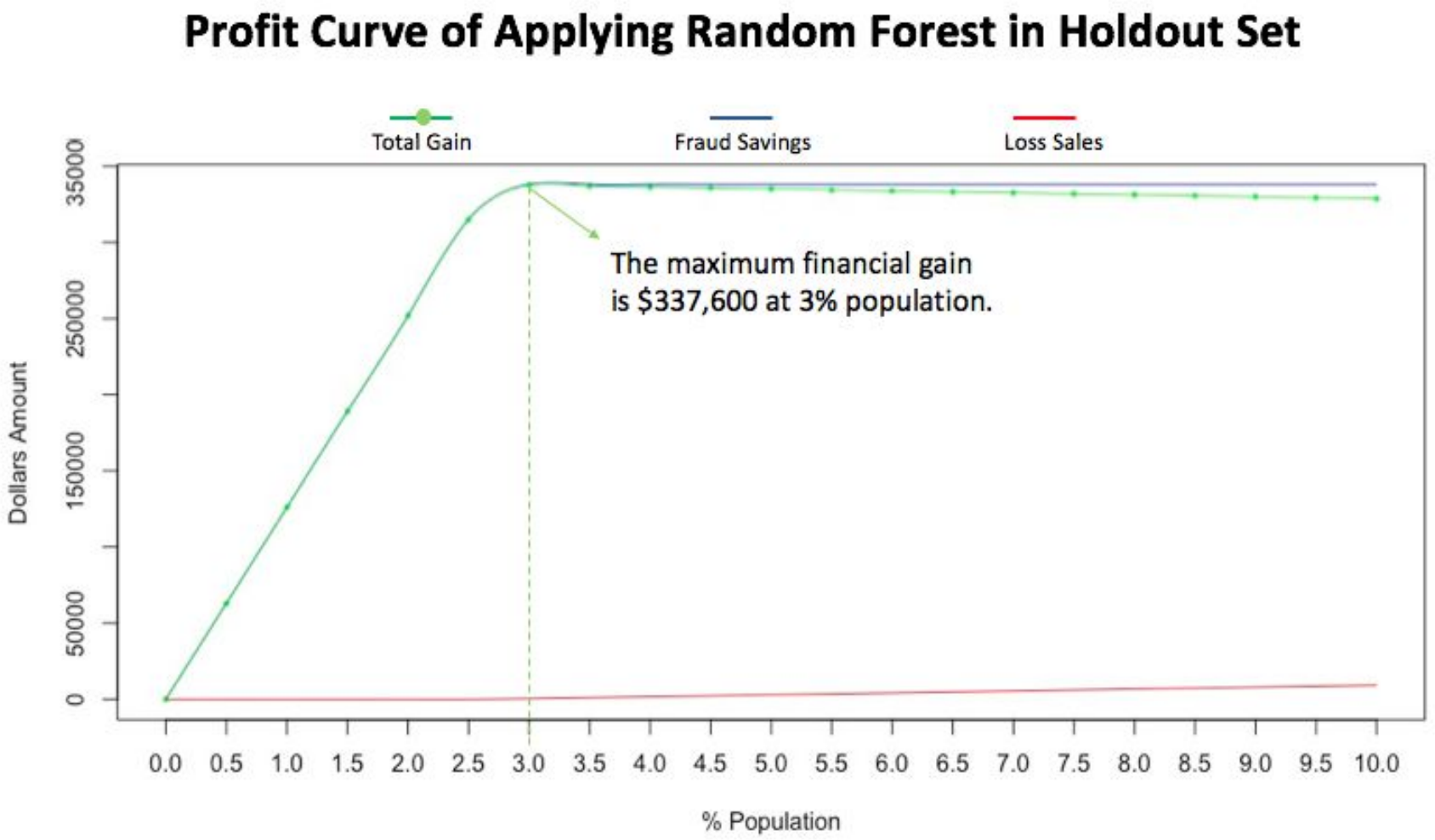
Transaction Frequency Variables (3/10): same_card_3, same_card_7, same_card_14
Profiling Variables (5/10): amount_sum_3, amount_sum_7, amount_avg_7, amount_sum_14, amount_sum_28

FDR Result


Random Forests FDR Result
Training + Test
Hold-Out Set
FDR
0.0
0.2
0.4
0.6
0.8
1.0
At 2% population, the FDR
is 74.56% on OOT set.
0.00
0.02
0.04
0.06
0.08
0.10
% Population

Financial Implication

Fraud Saving = $1000 * # True Positive
Loss Sales = $10 * # False Positive
Total Gain = Fraud Saving - Loss Sales



Recommendations

1. Apply the Random Forest model to detect transaction fraud
2. Set the cut-off at 3% to gain the maximum profit
3. Give sufficient communication before determining that the account is fraudulent so as to decrease false positive as far as possible

Thank You!

Team 5:

Rui Xin Wu, Zhuolin Ouyang, Xiaoxuan Feng,
Zimeng Cao, Feifei Shao, Zihao Geng, Zijian Wang