Identity Fraud Detection

DSO 562: Fraud Analytics

Team 5

Team 5

Naiyuan Xiao

Yucen Zhao

Dongyue Wang

Jingke Gao

Jiaying Li

Anyu Li

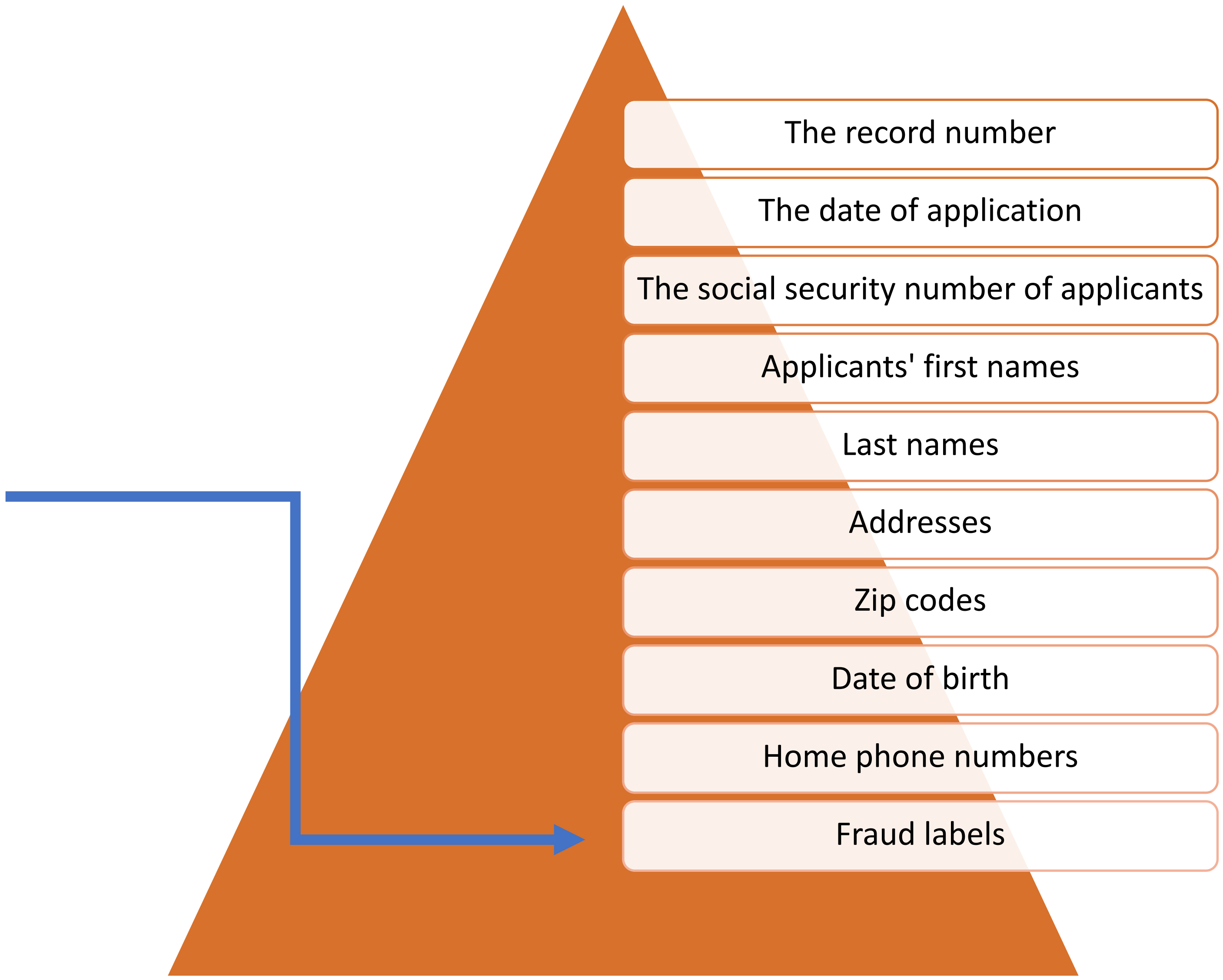
Agenda

- Dataset overview
- Data Preparation
- Model Building
- Result and Interpretation
- Future improvement

Dataset Overview

1,000,000 Records
10 fields – All categorical variables

Fraud	Count
0	985607
1	14393



Variable Creation

Velocity Variables	Total count by	firstname lastname fulladdress homephone nameDOB ssn	And some combinations	e.g ssn&fulladdress nameDOB&ssn ssn&homephone	over the past	0 days 1 day 3 days 7 days 14 days 30 days	(6+10) x 6=96
Days since Variables		Current date minus date of most recent application with same		e.g ssnnameDOB ssnfulladdress nameDOB lastnamessn			12
Risk Variables		Constellation ZIP3 Day of Week					3
Other Variables		Full Address Name+Dob					2

- Total: 113

Variable Creation

nameDOB	fulladdress	firstnamehomephone0	firstnamehomephone1	firstnamehomephone3	firstnamehomephone7	firstnamehomephone'
XRRAMMTR_SMJETJMJ_19070626	6861 EUTST PL_02765	1	1	1	1	
MAMSTUJR_RTTEMRRR_19340615	7280 URASA PL_57169	1	1	1	1	
SZMMUJEZS_EUSEZRAE_19070626	5581 RSREX LN_56721	1	1	1	1	
SJJZSXRSZ_ETJXTXXS_19440430	1387 UJZXJ RD_35286	1	1	1	1	
SSSXUEJMS_SSUUJXUZ_19980315	279 EAASA WY_03173	1	1	1	1	
XEEJJSTER_ERJSAXA_19480613	4322 USJXU LN_08391	1	1	1	1	
XZJRJUSRR_STSMJRUM_19640318	478 EEXUM LN_41640	1	1	1	1	
EJMRRSUXR_AMTZXRU_19190528	8906 UUAJ PL_60567	1	1	1	1	
RXTSZJATS_RSXMRJME_19900314	8266 SSEAR RD_37934	1	1	1	1	

Feature Selection

- Filter (60/113)

I. KS

II. FDR

III. Combination Rank

- Bidirectional Selection (27/60)

Days_since_fulladdress	fulladdress30	fulladdress14	fulladdress7	fulladdress3
ssn7	homephone3	fulladdress1	Days_since_ssn	ssn30
Days_since_firstname_ssn	Days_since_lastname_ssn	Days_since_fulladdresshomephone	Days_since_nameDOB	ssnnameDOB30
Days_since_ssnnameDOB	fulladdresshomephone30	nameDOB30	lastnamessn30	firstnamessn30
fulladdresshomephone14	nameDOB14	ssnnameDOB14	fulladdresshomephone7	nameDOB7
homephone7	zip3_risk			

Data Modeling

- Logistic Regression
- Neural Networks
- Random Forest
- Boosted Trees

Logistic Regression

- Baseline Model
  - Simple

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Assumptions:
  - No or little collinearity
  - Linear relationship between the log odds and the predictors.

- Validation Sets:
  - Randomly split train and test datasets
  - Repeated: 10 times
  - Result: Average

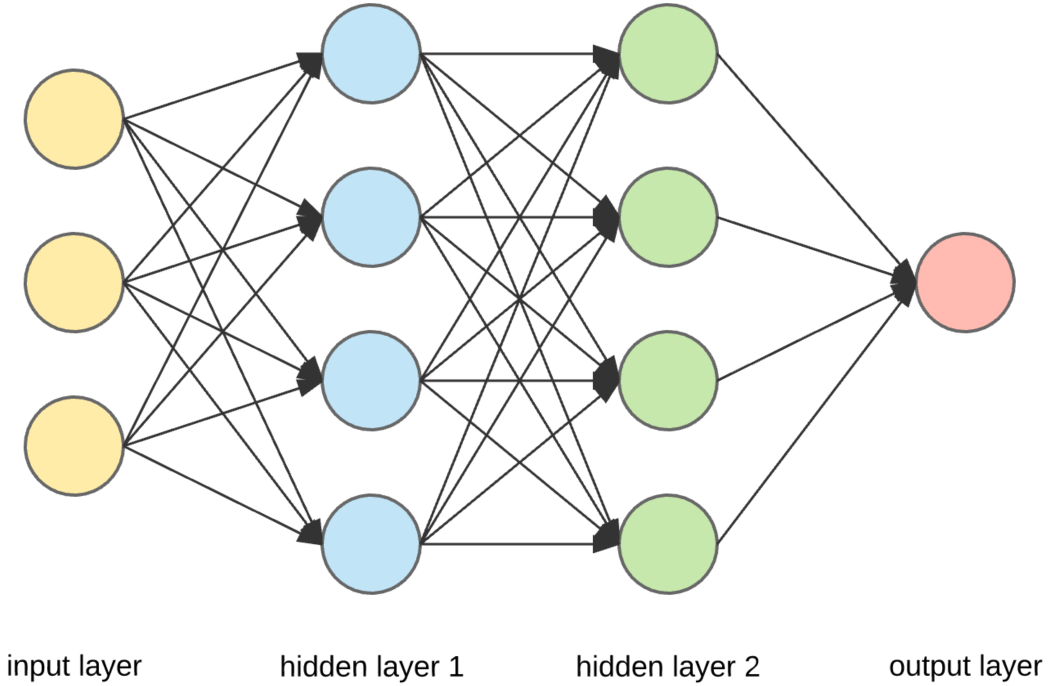
- Evaluation: FDR @3%
  - Rank-order on prediction
  - Slice Top 3%
  - Calculate: $\frac{\#\ frauds\ in\ Top\ 3\%}{Total\ \#\ frauds\ in\ each\ dataset}$
    - Separately for train/test/oot

Logistic Regression - Result

Logistic Regression				
Logistic Regression	**Number of Variables**	**TRAIN**	**TEST**	**OOT**
1	20	49.87%	49.44%	47.42%
2	26	52.67%	52.35%	50.15%
3	**27 w/ zip3_risk variable**	**53.42%**	**53.01%**	**50.27%**

Neural Networks

- Algorithm:
  - Linear Combination of the previous layers
  - Activation function for outputs
- Parameter Tuning:
  - # of Hidden Layer: 2
  - # of nodes in each hidden layer
- Activation function:
  - *Sigmoid logistic function*



Neural Networks - Result

Neural Networks (2 hidden layers)				
# nodes 1st hidden layer	**# nodes 2nd hidden layer**	**Train**	**Test**	**OOT**
2	**3**	**51.58%**	**51.73%**	**49.20%**
2	5	49.94%	50.06%	47.03%
9	2	49.80%	49.71%	46.39%
10	4	49.71%	49.67%	45.76%

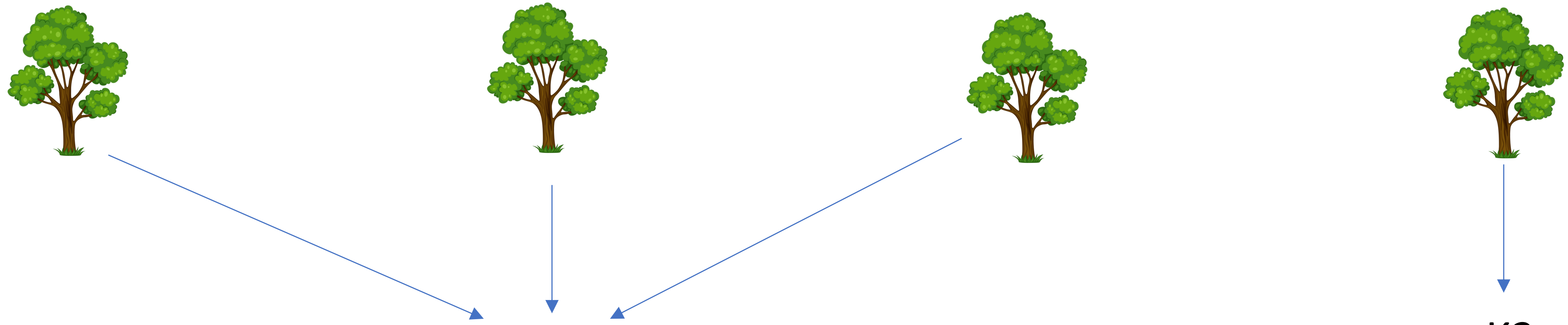
Random Forest - Model Description

$$X_1, X_2, X_3 \ldots\ldots X_{27}$$

$$m = \sqrt{p}$$

$$m \approx 5$$

$X_1\ X_3\ X_{11}\ X_{20}\ X_{24}$ | $X_4\ X_8\ X_{10}\ X_2\ X_{25}$ | $X_7\ X_{\mathbf{9}}\ X_{12}\ X_{14}\ X_{18}$ | $X_5\ X_6\ X_{13}\ X_{16}\ X_{22}$



K1 | K2

Majority vote

Random Forest – Result

- Package: ranger
- Parameters tuning:
  Number of trees: 50, 100, 400, 1000
  Number of variables in each tree: 5, 7

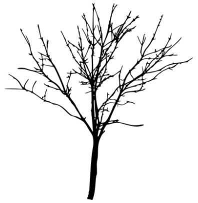
Best model: 100 trees, 5 variables

Random Forest				
# Trees	**# Vars**	**Train**	**Test**	**OOT**
50	5	57.62%	57.35%	55.48%
100	**5**	**57.59%**	**57.35%**	**55.53%**
400	5	57.63%	57.42%	55.51%
400	7	57.55%	57.27%	55.32%
1000	7	57.55%	57.29%	55.33%

Boosted Trees - Model Description

- Sum of weak learners
- Trees are grown sequentially
- Fit the tree using current residuals

$$Set\ f(x) = 0, r_i = y_i \longrightarrow \begin{array}{l} f(x) \leftarrow f(x) + \lambda f^b(x) \\ r_i \leftarrow r_i - \lambda f^b(x_i) \end{array} \longrightarrow f(x) = \sum_{b=1}^{\infty} \left(\lambda f^b(x)\right)$$



$\longrightarrow$

Boosted Tree – Result

- Package : gbm
- Parameters tuning:
  Number of trees: 500, 1000
  Interaction depth: 2, 3, 4
  Shrinkage (learning rate): 0.1
  Distribution: Bernoulli

Best model: 1000 trees, Interaction depth 3

Boosting Trees				
# Trees	**Depth**	**Train**	**Test**	**OOT**
500	2	57.39%	56.92%	54.93%
500	3	58.09%	56.92%	55.63%
500	4	58.16%	57.59%	55.56%
1000	2	58.04%	57.47%	55.39%
1000	**3**	**58.25%**	**57.62%**	**55.81%**

Boosted Tree: 1000 trees & depth 3

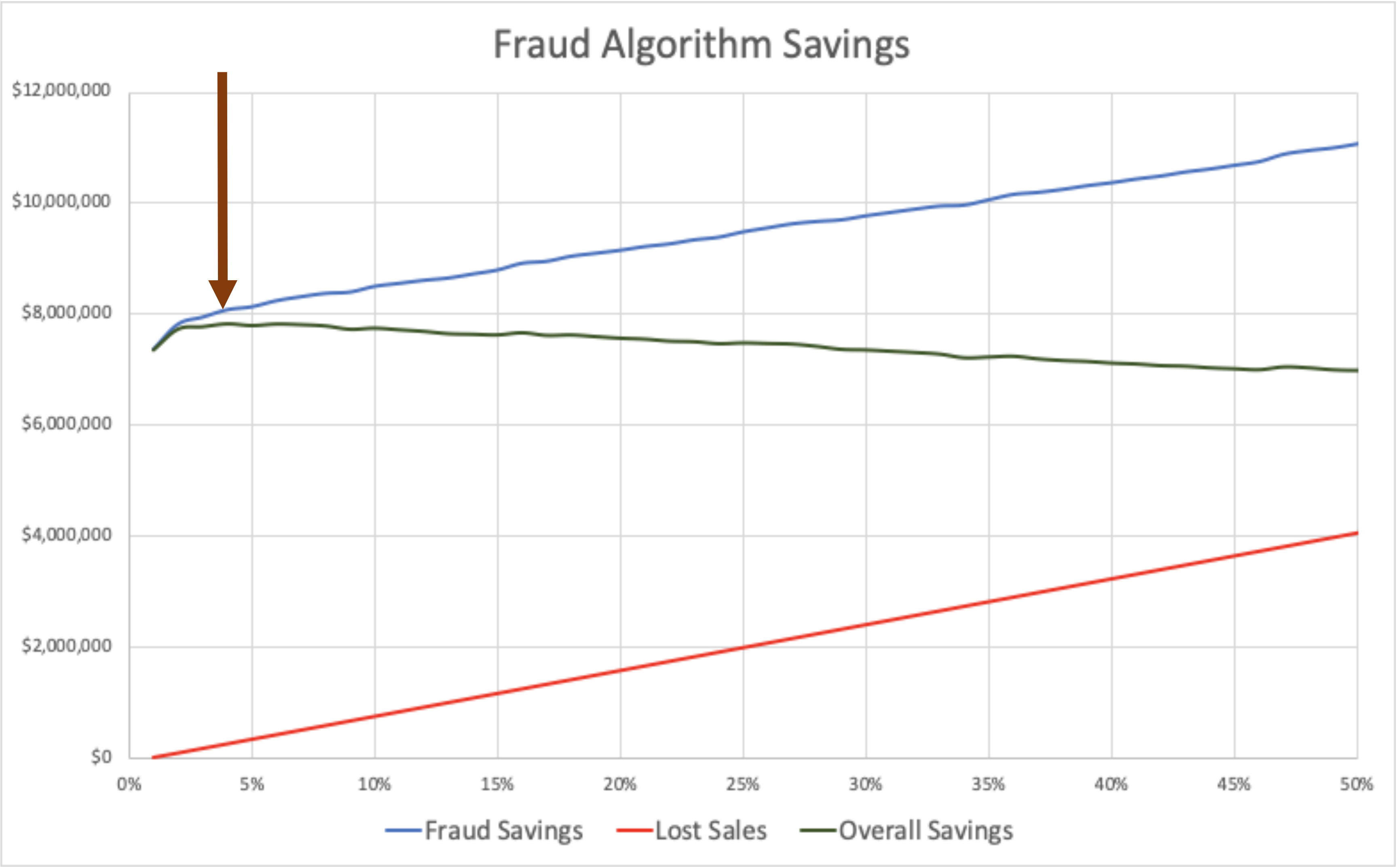
Training	# Records			# Goods			# Bads			Fraud Rate		
	583454			575137			8317			0.0143		
	Bin Statistic					Cumulative						
Population Bin	# Records	# Goods	# Bads	% Goods	% Bads	Total # records	Cumulative Goods	Cumulative bads	% All goods	% Bads (FDR)	KS	FPR
1	5835	1303	4532	22.3	77.7	5835	1303	4532	0.23	54.49	54.26	0.3
2	5834	5593	241	95.9	4.1	11669	6896	4773	1.20	57.39	56.19	1.4
3	5835	5751	84	98.6	1.4	17504	12647	4857	2.20	58.40	56.20	2.6
4	5834	5748	86	98.5	1.5	23338	18395	4943	3.20	59.43	56.23	3.7
5	5835	5768	67	98.9	1.1	29173	24163	5010	4.20	60.24	56.04	4.8
6	5834	5776	58	99.0	1.0	35007	29939	5068	5.21	60.94	55.73	5.9
7	5835	5766	69	98.8	1.2	40842	35705	5137	6.21	61.77	55.56	7.0
8	5834	5778	56	99.0	1.0	46676	41483	5193	7.21	62.44	55.23	8.0
9	5835	5787	48	99.2	0.8	52511	47270	5241	8.22	63.02	54.80	9.0
10	5834	5770	64	98.9	1.1	58345	53040	5305	9.22	63.79	54.56	10.0
11	5835	5778	57	99.0	1.0	64180	58818	5362	10.23	64.47	54.24	11.0
12	5834	5785	49	99.2	0.8	70014	64603	5411	11.23	65.06	53.83	11.9
13	5835	5788	47	99.2	0.8	75849	70391	5458	12.24	65.62	53.39	12.9
14	5835	5798	37	99.4	0.6	81684	76189	5495	13.25	66.07	52.82	13.9
15	5834	5789	45	99.2	0.8	87518	81978	5540	14.25	66.61	52.36	14.8
16	5835	5790	45	99.2	0.8	93353	87768	5585	15.26	67.15	51.89	15.7
17	5834	5780	54	99.1	0.9	99187	93548	5639	16.27	67.80	51.54	16.6
18	5835	5780	55	99.1	0.9	105022	99328	5694	17.27	68.46	51.19	17.4
19	5834	5782	52	99.1	0.9	110856	105110	5746	18.28	69.09	50.81	18.3
20	5835	5787	48	99.2	0.8	116691	110897	5794	19.28	69.66	50.38	19.1

Boosted Tree: 1000 trees & depth 3

Testing	# Record				# Good		# Bad			#Fraud Rate		
	250053				246363		3690			0.0148		
	Bin Statistics					Cumulative Statistics						
Population Bin %	# record	# good	# bad	% good	% bad	Total # record	Cumulative goods	Cumulative bads	% goods	% bad (FDR)	KS	FPR
1	2501	564	1937	22.6	77.4	2501	564	1937	0.23	52.49	52.3	0.29
2	2500	2361	139	94.4	5.6	5001	2925	2076	1.19	56.26	55.1	1.41
3	2501	2464	37	98.5	1.5	7502	5389	2113	2.19	57.26	55.1	2.55
4	2500	2473	27	98.9	1.1	10002	7862	2140	3.19	57.99	54.8	3.67
5	2501	2468	33	98.7	1.3	12503	10330	2173	4.19	58.89	54.7	4.75
6	2500	2477	23	99.1	0.9	15003	12807	2196	5.20	59.51	54.3	5.83
7	2501	2480	21	99.2	0.8	17504	15287	2217	6.21	60.08	53.9	6.90
8	2500	2471	29	98.8	1.2	20004	17758	2246	7.21	60.87	53.7	7.91
9	2501	2482	19	99.2	0.8	22505	20240	2265	8.22	61.38	53.2	8.94
10	2500	2478	22	99.1	0.9	25005	22718	2287	9.22	61.98	52.8	9.93
11	2501	2466	35	98.6	1.4	27506	25184	2322	10.22	62.93	52.7	10.85
12	2500	2479	21	99.2	0.8	30006	27663	2343	11.23	63.50	52.3	11.81
13	2501	2479	22	99.1	0.9	32507	30142	2365	12.23	64.09	51.9	12.75
14	2500	2482	18	99.3	0.7	35007	32624	2383	13.24	64.58	51.3	13.69
15	2501	2480	21	99.2	0.8	37508	35104	2404	14.25	65.15	50.9	14.60
16	2500	2471	29	98.8	1.2	40008	37575	2433	15.25	65.93	50.7	15.44
17	2501	2466	35	98.6	1.4	42509	40041	2468	16.25	66.88	50.6	16.22
18	2501	2471	30	98.8	1.2	45010	42512	2498	17.26	67.70	50.4	17.02
19	2500	2483	17	99.3	0.7	47510	44995	2515	18.26	68.16	49.9	17.89
20	2501	2469	32	98.7	1.3	50011	47464	2547	19.27	69.02	49.8	18.64

Boosted Tree: 1000 trees & depth 3

OOT	# Records			# Goods			# Bads			Fraud Rate	
	166493			164107			2386			0.0143	
	Bin Statistic					Cumulative					
Population Bin	# Records	# Goods	# Bads	% Goods	% Bads	Total # records	Cumulative Goods	Cumulative bads	% All goods	% Bads (FDR)	KS	FPR
1	1665	434	1231	26.1	73.9	1665	434	1231	0.3	51.6	51.3	0.4
2	1665	1590	75	95.5	4.5	3330	2024	1306	1.2	54.7	53.5	1.5
3	1665	1644	21	98.7	1.3	4995	3668	1327	2.2	55.6	53.4	2.8
4	1665	1643	22	98.7	1.3	6660	5311	1349	3.2	56.5	53.3	3.9
5	1665	1656	9	99.5	0.5	8325	6967	1358	4.2	56.9	52.7	5.1
6	1665	1647	18	98.9	1.1	9990	8614	1376	5.2	57.7	52.4	6.3
7	1665	1653	12	99.3	0.7	11655	10267	1388	6.3	58.2	51.9	7.4
8	1664	1654	10	99.4	0.6	13319	11921	1398	7.3	58.6	51.3	8.5
9	1665	1661	4	99.8	0.2	14984	13582	1402	8.3	58.8	50.5	9.7
10	1665	1648	17	99.0	1.0	16649	15230	1419	9.3	59.5	50.2	10.7
11	1665	1656	9	99.5	0.5	18314	16886	1428	10.3	59.8	49.6	11.8
12	1665	1656	9	99.5	0.5	19979	18542	1437	11.3	60.2	48.9	12.9
13	1665	1658	7	99.6	0.4	21644	20200	1444	12.3	60.5	48.2	14.0
14	1665	1653	12	99.3	0.7	23309	21853	1456	13.3	61.0	47.7	15.0
15	1665	1653	12	99.3	0.7	24974	23506	1468	14.3	61.5	47.2	16.0
16	1665	1645	20	98.8	1.2	26639	25151	1488	15.3	62.4	47.0	16.9
17	1665	1659	6	99.6	0.4	28304	26810	1494	16.3	62.6	46.3	17.9
18	1665	1650	15	99.1	0.9	29969	28460	1509	17.3	63.2	45.9	18.9
19	1665	1656	9	99.5	0.5	31634	30116	1518	18.4	63.6	45.3	19.8
20	1665	1656	9	99.5	0.5	33299	31772	1527	19.4	64.0	44.6	20.8



Overall Savings = Fraud Savings – Lost Sales

Assume:

$6,000 gain for every fraud caught

$50 loss for every false positive

Cutoff: 4%

Almost $8M savings !!!

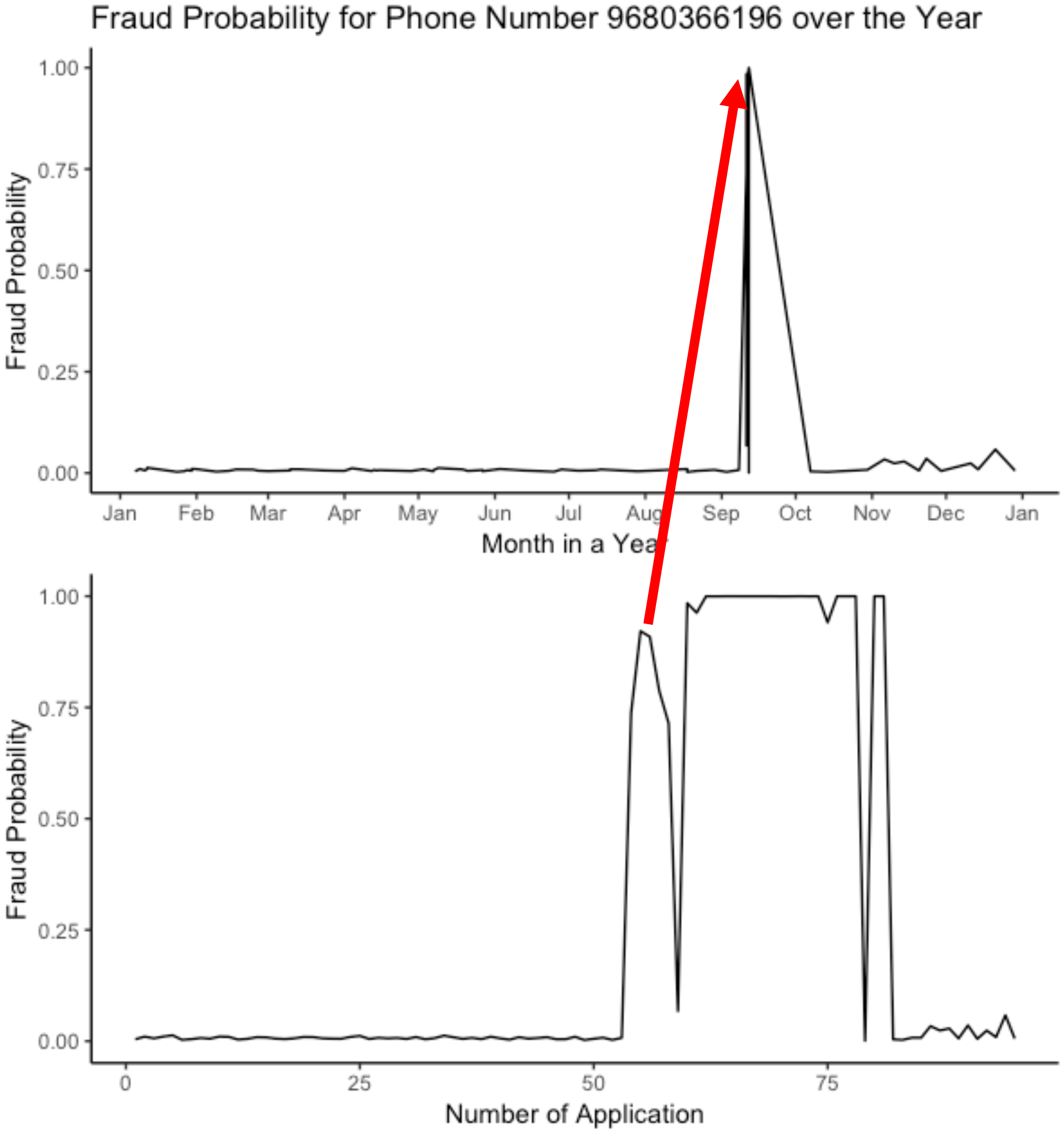
Business Insights

How does our model function in a business content?

1. Catching new fraud is important, but we don't want to put barriers for customers to create a new account.

2. It's not efficient to check whether the address, phone number, etc., are valid or not.

How does our model take these into account?



Fraud Probability for Phone Number 9680366196 over the Year
Fraud Probability
1.00
0.75
0.50
0.25
0.00
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Month in a Year
Fraud Probability
1.00
0.75
0.50
0.25
0.00
0
25
50
75
Number of Application

2016/9/3 - 2016/9/11

record	fraud_label	Days_since_fulladdress	fulladdress30	fulladdress14	fulladdress7
673422	0	365	1	1	1
685702	0	365	1	1	1
694483	1	0	1	1	1
696365	1	0	7	7	7
694580	1	0	2	2	2

nameDOB14	ssnnameDOB14	fulladdresshomephone7	nameDOB7	ssn7	homephone3	pred
1	1	1	1	1	1	0.0027918604
1	1	1	1	1	1	0.0075436074
1	1	1	1	1	2	0.7382900996
1	1	7	1	1	9	0.9216079640
1	1	2	1	1	3	0.9087503985

Importance Table

Number of times that the full address used in the past 30 days
Number of times that same SSN, name and DOB used in the past 14 days
Risk Variable - Zip3
Number of times that the phone number used in the past 3 days
Number of days since the last time the same full address was used
Number of days since the last time the same name and dob were used
Number of times that SSN used in the past 30 days
Number of times that phone number used in the past 7 days
Number of times that full address used in the past 1 day
Number of days since last time the same full address and home phone were used

Address information is easy to be stolen or manipulated for fraud.

Future Improvements

External Datasets + New Variables + Balanced Dataset + Support Vector Machine = Perfection!

External Datasets & New Variables

- External Datasets

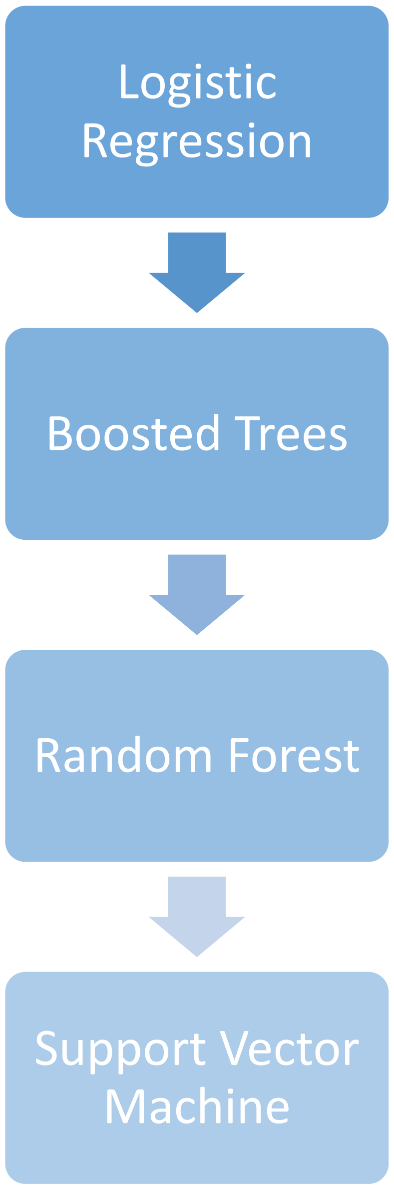
a. Identity Fraud Hot Spot Dataset: Zip, address
b. U.S Common Scams and Fraud: phone number
c. Credit Card information for the existing applicants

- More variables using expert knowledge

a. Special Identity: phone number, SSN, address, name_DOB
b. More velocity variables
c. Interaction Effect

Balanced Dataset & SVM

Neural Network		
Parameters	**OOT(Unbalanced)**	**OOT(Balanced)**
2 layers (2,3)	49.20%	51.15%
2 layers (2,5)	47.03%	43.25%
2 layers (9,2)	46.39%	50.28%
2 layers (10,4)	45.76%	48.84%
Average	**47.10%**	**48.38%**



Thank You!

Any Question?