Project 1 Summary

1. Description of Data

The data contains personal identity information (PII) about application and detected fraud transaction. The whole dataset comes from synthetic application data regarding credit cards and cell phones. The dataset is synthesized based on U.S. applications over about 10 years. In general, the dataset has 1,000,000 records and 10 fields, covers applications from 2017-01-01 to 2017-12-31.

a. Numeric Fields

Fields	% Populated	Min	Max	Mean	Standard Deviation	% Zeros
date	100.00	2017-01-01	2017-12-31	/	/	0.00
dob	100.00	1900-01-01	2016-10-31	/	/	0.00

b. Categorical Fields

Fields	% Populated	# Unique Values	Most Common Value
record	100.00	1,000,000	N/A
ssn	100.00	835,819	999999999
firstname	100.00	78,136	EAMSTRMT
lastname	100.00	177,001	ERJSAXA
address	100.00	828,774	123 MAIN ST
zip5	100.00	26,370	68138
homephone	100.00	28,244	9999999999
fraud_label	100.00	2	0

2. Data Cleaning

The dataset doesn't have missing fields and bad records. However, frivolous fields exist in the dataset. Frivolous fields are missing values that are filled with default values, such as 123 Main Street. These improperly common values (supposed to be other values instead) will cause the wrong counts when creating new variables and constructing linkage.

To erase negative effect of frivolous fields, we replace them with the record number those frivolous field correspond to. Those values will be unique and therefore won't link to a previous value.

3. Variable Creation

a. Mode of Identity Fraud (Identity Theft)

(1) Identity theft: The fraudsters acquired others' information through online purchasing or stealing. The fraudsters further apply for services such as credit cards through stolen core identity information such as name, SSN, date of birth and uses his own address and phone number to receive the products (such as credit cards).

We explore how many applications occur with same name, SSN, date of birth, address and phone number in the past certain days to capture the feature of fraud.

(2) Identity manipulation: The fraudsters change the information a little bit to avoid algorithm detection or compile different information of real identities together to apply for products or services. We explore the combination of different information (such as ssn and date of birth and similarities) and examine their occurrence in the past certain days to capture the feature of fraud.

(3) Synthetic identity fraud: The Fraudsters make up all the information by using algorithm. That information could be out of nowhere or pieces of real information. We explore irregular values (such as extreme high or low age) to capture the feature of fraud.

b. Variable Table

Description of Variables	# Variables Created
Date of Week Variable: Average fraud percentage of a certain day of week	1
Age Variable: age of the applicant when the application was made	1
Days since Variables: Number of days since the last time that an application with that entity was seen.	23
Velocity: Number of records with the same entity over the last {0,1,3,7,14,30} days	138
Relative Velocity: Number of applications with same entity over the last {0,1} day / Number of applications with same entity over the last {1, 3, 7, 14, 30} days	184
Counts by entity: Number of Unique entities of one field for a particular entity of another field over the last {0, 1, 3, 7, 14, 30, 60} days	3542

Age indicator: Maximum, Minimum and mean age of applicants for a certain entity that has ever seen	69
Total	3958

4. Feature selection

a. Motivation of Feature Selection

(1) Multicollinearity. We have created as many variables as possible and therefore it causes multicollinearity. Feature selection will take multicollinearity into consideration while selecting the variables.

(2) Reduce dimensionality. Since it's hard to fit in the nonlinear model with high dimensionality, we need to use feature selection to reduce the number of variables to reduce dimensionality.

(3) Exploration of potential variables. Existence of feature selection allows us to explore as many features as possible to create efficient variables.

(4) Convenient to explore models. Feature selection will return a sorted list of variables indicating significance. Therefore, it helps us to further add or drop variables into models.

b. Process of Feature Selection

(1) Filter. Filter runs test between dependent variable (Fraud_label) and other independent variables to select certain number of variables with highest performance. We applied two-sample Kolmogorov-Smirnov test to filter variables with top 200 highest KS score.

(2) Wrapper. Wrapper applies greedy algorithms to run models with multiple times. Each time the model adds or drops variables to seek for a better model based on current model. In this project, based on LGBM we built forward selection with 20 variables selected. Forward selection builds model between dependent variable and each independent variable and selected one with the best performance. Then we re-ran the model by adding one independent variable and selected the one with the best performance. We ran 20 times to eventually select 20 independent variables.

Selected variables are listed below:

Wrapper Order	Variable	Filter Score
1	fulladdress_day_since	0.3333
2	name_dob_count_30	0.2275
3	address_unique_count_for_name_homephone_60	0.2924
4	address_count_0_by_30	0.2919
5	address_unique_count_for_ssn_zip5_60	0.2897
6	address_day_since	0.3341
7	address_unique_count_for_homephone_name_dob_60	0.2914
8	fulladdress_count_0_by_30	0.2907
9	address_unique_count_for_ssn_name_60	0.2897
10	address_unique_count_for_ssn_homephone_60	0.2892
11	address_unique_count_for_ssn_firstname_60	0.2881
12	address_unique_count_for_ssn_name_dob_60	0.2876
13	address_unique_count_for_dob_homephone_60	0.2876
14	address_unique_count_for_ssn_lastname_60	0.2874
15	address_unique_count_for_ssn_60	0.2859
16	address_unique_count_for_ssn_name_dob_14	0.2769
17	address_unique_count_for_homephone_name_dob_30	0.2840
18	address_count_14	0.3224
19	fulladdress_unique_count_for_dob_homephone_3	0.2644
20	address_count_7	0.3017

5. Preliminary Models Exploration

Based on 20 variables selected, we run the model. The results are listed below.

		penalty	C	max_iter	solver	l_1 ratio	Train	Test	OOT	Performance
Logistic Regression	1	l2	1	3	lbfgs	N/A	0.5157	0.5100	0.4917	
	2	l2	1	20	lbfgs	N/A	0.4867	0.4848	0.4712	
	3	l2	1	50	lbfgs	N/A	0.4882	0.4881	0.4737	
	4	l2	1	100	lbfgs	N/A	0.4886	0.4932	0.4754	
	5	l2	1	500	lbfgs	N/A	0.4884	0.4884	0.4745	
	6	elasticnet	1	500	saga	0.2	0.4885	0.4875	0.4740	
		max_depth	min_samples_split	min_samples_leaf		max_features	Train	Test	OOT	Performance
Decision Tree	1	3	50	30		5	0.4587	0.4516	0.4371	Underfitting
	2	15	50	30		5	0.5333	0.5219	0.5018	
	3	15	40	20		5	0.5280	0.5217	0.5006	
	4	15	40	20		10	0.5326	0.5257	0.5014	
	5	15	30	15		10	0.5322	0.5235	0.5023	
	6	30	40	20		10	0.5386	0.5246	0.4999	
	7	60	40	20		10	0.5414	0.5186	0.5008	Overfitting
		n_estimators	max_depth	min_samples_split	min_samples_leaf	max_features	Train	Test	OOT	Performance
Random Forest	1	5	5	100	80	5	0.5158	0.5133	0.4920	Underfitting
	2	5	15	100	80	5	0.5312	0.5225	0.5027	
	3	25	5	100	80	5	0.5219	0.5181	0.4961	
	4	25	5	50	30	5	0.5211	0.5232	0.4916	
	5	25	5	50	30	10	0.5182	0.5208	0.4965	
	6	100	10	50	30	10	0.5287	0.5289	0.5045	
	7	25	25	50	30	10	0.5400	0.5179	0.5008	Overfitting
		hidden_layer_size	activation	solver	learning_rate	earning_rate_ini	max_iter	Train	Test	OOT	Performance
Neural Network	1	(1, 3)	relu	adam	constant	0.01	200	0.4206	0.4153	0.3914	Underfitting
	2	(1, 20)	relu	adam	constant	0.01	200	0.4759	0.4740	0.4578	Underfitting
	3	(10, 20)	relu	adam	constant	0.01	200	0.5244	0.5278	0.5059	
	4	(20, 20, 20)	relu	adam	constant	0.01	200	0.5256	0.5246	0.5024	
	5	(20, 20, 20)	relu	adam	adaptive	0.001	200	0.5281	0.5252	0.5062	
	6	(10, 10)	relu	lbfgs	adaptive	0.01	200	0.5252	0.5314	0.5053	
	7	(10, 10)	logistic	adam	adaptive	0.01	200	0.5232	0.5270	0.5013	
		max_depth	n_estimators	learning_rate		min_samples_leaf		Train	Test	OOT	Performance
Gradient Boosting	1	2	5	0.01		1		0.4932	0.4870	0.4738	Underfitting
	2	2	5	0.1		1		0.4935	0.4926	0.4765	Underfitting
	3	5	5	0.1		1		0.5234	0.5226	0.5006	
	4	10	5	0.1		1		0.5313	0.5252	0.5038	
	5	2	10	0.1		1		0.5101	0.5077	0.4883	
	6	5	10	0.1		1		0.5207	0.5292	0.5020	
	7	10	10	0.1		1		0.5324	0.5243	0.5012	Overfitting
	8	20	20	0.1		20		0.5411	0.5095	0.4904	Overfitting
		verbose	max_depth	iterations		learning_rate		Train	Test	OOT	Performance
Catboost	1	0	2	5		0.1		0.5003	0.5017	0.4793	Underfitting
	2	0	2	5		0.05		0.5002	0.5018	0.4793	Underfitting
	3	0	4	50		0.1		0.5225	0.5173	0.5001	
	4	0	4	50		0.05		0.5128	0.5100	0.4919	
	5	0	6	200		0.1		0.5311	0.5208	0.5052	
	6	0	6	200		0.05		0.5299	0.5207	0.5029	
	7	0	15	200		0.1		0.5327	0.5234	0.5053	

6. Summary of Results

We choose Random Forest as the final model. In the model, 100 trees are set, maximum depth is 10, minimum samples required to split an internal node is 50, minimum of samples required at a leaf node is 30, and 10 variables are included. In general, we achieve fraud detection rate at 3% (FDR 3%). That means our model can capture 53.09% fraud of training data, 52.42% fraud of testing data and 50.38% fraud of out of time data at top 3% of the data (Our model only needs to reject 3% of the data). Summary tables are listed below:

a. Training Data

Training	# Records	# Goods	# Bads	Fraud Rate
	583454	574963	8491	0.01455

	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
1	5835	1561	4274	26.75%	73.25%	5835	1561	4274	0.27%	50.34%	50.06	0.37
2	5834	5672	162	97.22%	2.78%	11669	7233	4436	1.26%	52.24%	50.99	1.63
3	5835	5763	72	98.77%	1.23%	17504	12996	4508	2.26%	53.09%	50.83	2.88
4	5834	5766	68	98.83%	1.17%	23338	18762	4576	3.26%	53.89%	50.63	4.10
5	5835	5789	46	99.21%	0.79%	29173	24551	4622	4.27%	54.43%	50.16	5.31
6	5834	5783	51	99.13%	0.87%	35007	30334	4673	5.28%	55.03%	49.76	6.49
7	5835	5799	36	99.38%	0.62%	40842	36133	4709	6.28%	55.46%	49.17	7.67
8	5834	5789	45	99.23%	0.77%	46676	41922	4754	7.29%	55.99%	48.70	8.82
9	5835	5801	34	99.42%	0.58%	52511	47723	4788	8.30%	56.39%	48.09	9.97
10	5834	5788	46	99.21%	0.79%	58345	53511	4834	9.31%	56.93%	47.62	11.07
11	5835	5788	47	99.19%	0.81%	64180	59299	4881	10.31%	57.48%	47.17	12.15
12	5834	5792	42	99.28%	0.72%	70014	65091	4923	11.32%	57.98%	46.66	13.22
13	5835	5787	48	99.18%	0.82%	75849	70878	4971	12.33%	58.54%	46.22	14.26
14	5835	5781	54	99.07%	0.93%	81684	76659	5025	13.33%	59.18%	45.85	15.26
15	5834	5802	32	99.45%	0.55%	87518	82461	5057	14.34%	59.56%	45.22	16.31
16	5835	5796	39	99.33%	0.67%	93353	88257	5096	15.35%	60.02%	44.67	17.32
17	5834	5797	37	99.37%	0.63%	99187	94054	5133	16.36%	60.45%	44.09	18.32
18	5835	5790	45	99.23%	0.77%	105022	99844	5178	17.37%	60.98%	43.62	19.28
19	5834	5792	42	99.28%	0.72%	110856	105636	5220	18.37%	61.48%	43.10	20.24
20	5835	5791	44	99.25%	0.75%	116691	111427	5264	19.38%	62.00%	42.62	21.17

b. Testing Data

Training	# Records	# Goods	# Bads	Fraud Rate
	250053	246537	3516	0.01406

	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
1	2501	747	1754	29.87%	70.13%	2501	747	1754	0.30%	49.89%	49.58	0.43
2	2500	2452	48	98.08%	1.92%	5001	3199	1802	1.30%	51.25%	49.95	1.78
3	2501	2460	41	98.36%	1.64%	7502	5659	1843	2.30%	52.42%	50.12	3.07
4	2500	2486	14	99.44%	0.56%	10002	8145	1857	3.30%	52.82%	49.51	4.39
5	2501	2486	15	99.40%	0.60%	12503	10631	1872	4.31%	53.24%	48.93	5.68
6	2500	2479	21	99.16%	0.84%	15003	13110	1893	5.32%	53.84%	48.52	6.93
7	2501	2477	24	99.04%	0.96%	17504	15587	1917	6.32%	54.52%	48.20	8.13
8	2500	2483	17	99.32%	0.68%	20004	18070	1934	7.33%	55.01%	47.68	9.34
9	2501	2479	22	99.12%	0.88%	22505	20549	1956	8.34%	55.63%	47.30	10.51
10	2500	2481	19	99.24%	0.76%	25005	23030	1975	9.34%	56.17%	46.83	11.66
11	2501	2474	27	98.92%	1.08%	27506	25504	2002	10.34%	56.94%	46.59	12.74
12	2500	2481	19	99.24%	0.76%	30006	27985	2021	11.35%	57.48%	46.13	13.85
13	2501	2485	16	99.36%	0.64%	32507	30470	2037	12.36%	57.94%	45.58	14.96
14	2500	2483	17	99.32%	0.68%	35007	32953	2054	13.37%	58.42%	45.05	16.04
15	2501	2492	9	99.64%	0.36%	37508	35445	2063	14.38%	58.67%	44.30	17.18
16	2500	2481	19	99.24%	0.76%	40008	37926	2082	15.38%	59.22%	43.83	18.22
17	2501	2473	28	98.88%	1.12%	42509	40399	2110	16.39%	60.01%	43.62	19.15
18	2501	2488	13	99.48%	0.52%	45010	42887	2123	17.40%	60.38%	42.99	20.20
19	2500	2485	15	99.40%	0.60%	47510	45372	2138	18.40%	60.81%	42.40	21.22
20	2501	2485	16	99.36%	0.64%	50011	47857	2154	19.41%	61.26%	41.85	22.22

c. Out of Time Data

Training	# Records	# Goods	# Bads	Fraud Rate
	166493	164107	2386	0.01433

	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
1	1665	513	1152	30.81%	69.19%	1665	513	1152	0.31%	48.28%	47.97	0.45
2	1665	1638	27	98.38%	1.62%	3330	2151	1179	1.31%	49.41%	48.10	1.82
3	1665	1642	23	98.62%	1.38%	4995	3793	1202	2.31%	50.38%	48.07	3.16
4	1665	1642	23	98.62%	1.38%	6660	5435	1225	3.31%	51.34%	48.03	4.44
5	1665	1654	11	99.34%	0.66%	8325	7089	1236	4.32%	51.80%	47.48	5.74
6	1665	1653	12	99.28%	0.72%	9990	8742	1248	5.33%	52.31%	46.98	7.00
7	1665	1654	11	99.34%	0.66%	11655	10396	1259	6.33%	52.77%	46.43	8.26
8	1664	1653	11	99.34%	0.66%	13319	12049	1270	7.34%	53.23%	45.88	9.49
9	1665	1654	11	99.34%	0.66%	14984	13703	1281	8.35%	53.69%	45.34	10.70
10	1665	1655	10	99.40%	0.60%	16649	15358	1291	9.36%	54.11%	44.75	11.90
11	1665	1656	9	99.46%	0.54%	18314	17014	1300	10.37%	54.48%	44.12	13.09
12	1665	1655	10	99.40%	0.60%	19979	18669	1310	11.38%	54.90%	43.53	14.25
13	1665	1653	12	99.28%	0.72%	21644	20322	1322	12.38%	55.41%	43.02	15.37
14	1665	1654	11	99.34%	0.66%	23309	21976	1333	13.39%	55.87%	42.48	16.49
15	1665	1656	9	99.46%	0.54%	24974	23632	1342	14.40%	56.24%	41.84	17.61
16	1665	1653	12	99.28%	0.72%	26639	25285	1354	15.41%	56.75%	41.34	18.67
17	1665	1654	11	99.34%	0.66%	28304	26939	1365	16.42%	57.21%	40.79	19.74
18	1665	1652	13	99.22%	0.78%	29969	28591	1378	17.42%	57.75%	40.33	20.75
19	1665	1656	9	99.46%	0.54%	31634	30247	1387	18.43%	58.13%	39.70	21.81
20	1665	1644	21	98.74%	1.26%	33299	31891	1408	19.43%	59.01%	39.58	22.65