Project Documentation

I. Data description

The data includes the Personal Identifying Information (PII) from the applications. There are in total 10 PII fields. The records are collected from January 1st 2017 to December 31st 2017, with a total of 1,000,000 rows. The objective is to look for identity fraud and identify fraud algorithms from the PII fields.

Summary Tables

a. Numerical Table

Field	% Populated	Min	Max	Mean	Standard Deviation	%Zero
Date	100%	2017-01-01	2017-12-31	n/a	n/a	0%
Dob	100%	1900-01-01	2016-10-31	n/a	n/a	0%

b. Categorical table

Field	% Populated	# Unique Values	Most Common Value
Record	100%	1,000,000	/
SSN	100%	835,819	999999999
Firstname	100%	78,136	EAMSTRMT
Lastname	100%	177,001	ERJSAXA
Address	100%	828,774	123 MAIN ST
Zip5	100%	26,370	68138
Homephone	100%	28,244	9999999999
Fraud_Label	100%	2	0

II. Data cleaning

In order to prepare the data for feature engineering, we first fix the frivolous fields. Since the fields are filled with default values from the business, we need to replace them with a unique value that will not cause it to link to a previous value, in this case the record number. Then, we encoded the categorical fields to be used for linking later to make new variables. We also conducted statistical smoothing for 'fraud_label' to avoid any imbalance of class distribution, which could underestimate the likelihood of fraud.

III. Variable creation

There are two most common modes of identity fraud:

1. A single fraudster using core information of multiple identities (SSN, Name, DOB) and his own contact number and address.
2. A person's core information is compromised in a data breach and used by multiple fraudsters.

As a result, we essentially create variables to count by different linkages and for each combination, we also created variables that account for number of days since that combination is seen, number of records seen over the past number of days (velocity), and ratio of the short-term velocity to a longer-term averaged velocity (relative velocity). The table below shows the total of 4050 variables created. After deduplication, there were 2240 variables left. However, after creating the variables. We realized that the max variables were improperly formed, so we decided to remove them.

Variable Summary Table

Description of Variables	# Variables Created
Age When Apply Age of the applicant upon application submission	1
Day of Week Target Encoding Average fraud percentage of that day	1
Days Since Number of days since an application with that entity was last seen	23
Velocity Number of records with the same entity over the last 0,1,3,7,14,30 days	138
Relative Velocity Ratio of the short-term velocity (0,1 days) to a longer-term (3,7,14,30 day) averaged velocity	184
Count by Entities Count of unique values for an entity comparing with another entity over the period of 0,1,3,7,14,30,60 days	3542
Maximum Indicator Maximum count of each entity over the period of 1,3,7,30 days	92
Age Indicator Include the maximum value, minimum value and the average value of applicants' age by each entity	69
Total Number of Variables	**4050**

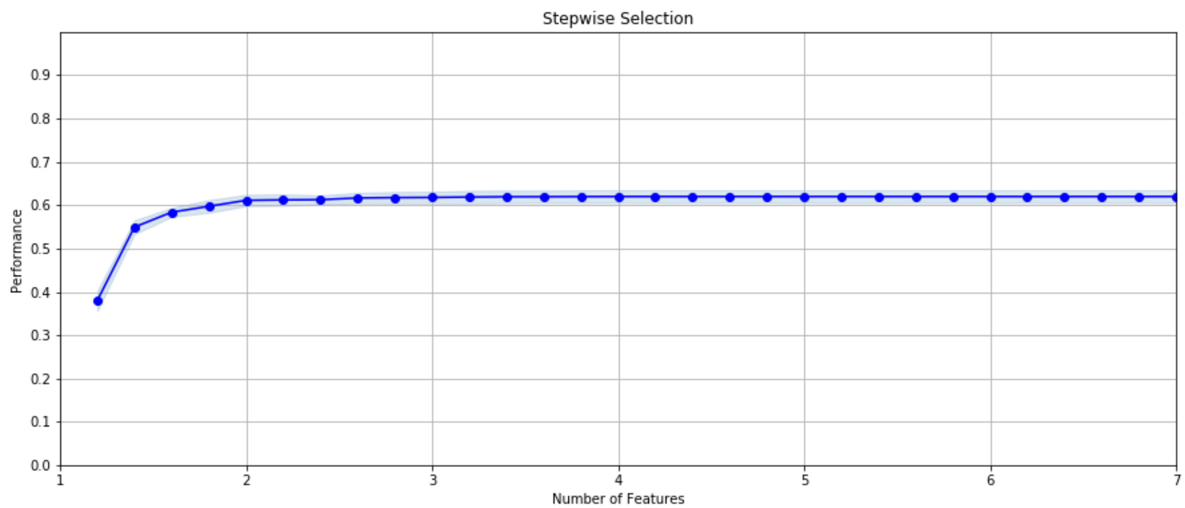
IV. Feature selection

In order for the model to run faster and allow more hyperparameter tuning, feature selection is needed. After making more than 4000 variables, we ran a filter to get down to 200 candidate variables. Then, we chose a simple, fast nonlinear model for the wrapper and ran it 20 times. In the end, we have a final list of 20 variables that are sorted in terms of their univariate KS's (the order of importance). Below is the final list of variables and the plot showing model performance versus number of variables.

List of Final variables

Wrapper order	Variable	Filter Score
1	fulladdress_day_since	0.3332690
2	ssn_firstname_day_since	0.2264275
3	fulladdress_unique_count_for_ssn_name_30	0.2819330
4	address_count_30	0.3326480
5	address_count_7	0.3017353
6	fulladdress_unique_count_for_ssn_dob_14	0.2762090
7	address_count_14	0.3224360
8	fulladdress_unique_count_for_ssn_homephone_60	0.2899910
9	ssn_count_30	0.2268940
10	address_unique_count_for_name_homephone_30	0.2845160
11	address_unique_count_for_ssn_zip5_7	0.2732480
12	fulladdress_unique_count_for_ssn_lastname_30	0.2818810
13	address_day_since	0.3341400
14	address_count_0_by_30	0.2919220
15	fulladdress_count_0_by_30	0.2907220
16	address_unique_count_for_ssn_zip5_60	0.2897236
17	address_unique_count_for_homephone_name_dob_60	0.2914098
18	address_unique_count_for_dob_homephone_60	0.2875560
19	fulladdress_unique_count_for_ssn_homephone_60	0.2899906
20	address_unique_count_for_ssn_name_60	0.2896792

Model Performance



V. Preliminary models exploration

After having our final list of variables, we choose the final 20 variables for model exploration. We first ran a linear logistic regression model as a baseline and explored with a few nonlinear models, which includes decision tree, random forest, LGBM, neural network, GBC, Catboost, XGBoost. The table below shows the results of multiple runs with different parameters and number of variables. We also attempted some overfitting and underfitting results.

Model	Parameter							Average FDR at 3%			
Logistic regression	**# of variables**	**penalty**	**C**	**Solver**		**l1_ratio**		**Train**	**Test**	**OOT**	
1	10	none	none	none		none		0.485405	0.487786	0.471081	
2	10	l1	none	saga		none		0.487824	0.488927	0.473931	
3	10	l2	1	saga		none		0.488076	0.488086	0.473764	
4	10	elasticnet	none	saga		0.4		0.486106	0.480378	0.469153	
5	10	none	none	liblinear		0.4		0.489868	0.484578	0.474099	
6	5	none	none	liblinear		0.4		0.477955	0.475607	0.463286	
Decision Tree	**# of variables**	**Max_depth**		**min_sample_split**		**min_sample_leafs**		**Train**	**Test**	**OOT**	
1	10	5		50		30		0.51256	0.511225	0.489187	
2	10	10		40		20		0.531522	0.518441	0.506119	
3	10	15		30		10		0.532135	0.523659	0.503521	
4	10	20		20		5		0.539516	0.516225	0.501341	
5	10	30		10		3		0.539554	0.519933	0.49715	
6	5	30		10		3		0.532687	0.520083	0.499749	
7	20	30		10		3		0.545973	0.514263	0.499581	
Random Forest	**# of variables**	**n_estimators**	**max_depth**	**min_sample_split**		**min_samples_leaf**	**max_features**	**Train**	**Test**	**OOT**	
1	10	10	5	50		30	3	0.51525	0.519802	0.493965	
2	10	30	15	40		20	5	0.531948	0.527312	0.504107	
3	10	50	25	30		10	8	0.539951	0.521735	0.501509	
4	10	100	30	20		5	10	0.539889	0.522605	0.500671	
5	15	100	30	20		5	10	**0.543988**	**0.516966**	0.500671	OVERFITTING
LightGBM	**# of variables**	**n_estimators**	**max_depth**	**num_leaves**	**col_samplebytree**	**learning_rate**	**eval_metric**	**Train**	**Test**	**OOT**	
1	10	20	2	2	1	0.1	none	0.511695	0.512854	0.488852	
2	10	100	3	4	0.8	0.03	auc	0.509624	0.50829	0.485163	
3	10	500	5	8	0.8	0.01	auc	0.521047	0.528574	0.502012	
4	10	1000	6	10	0.8	0.01	logloss	0.529548	0.523249	0.506203	
5	15	1000	6	10	0.8	0.01	logloss	0.531117	0.519609	0.504694	
Neural Network	**# of variables**	**hidden_layer_size**	**activation**	**alpha**	**learning_rate**	**solver**	**learning_rate_init**	**Train**	**Test**	**OOT**	
1	10	5	logistic	0.1	constant	adam	0.01	0.498557	0.497713	0.479212	
2	10	10	relu	0.001	constant	adam	0.001	0.526276	0.523445	0.50285	
3	10	20	relu	0.0001	adaptive	lbfgs	0.0001	0.529691	0.520945	0.506119	
4	20	20	relu	0.0001	adaptive	lbfgs	0.0001	0.525618	0.529903	0.507544	
5	15	20	relu	0.0001	adaptive	lbfgs	0.0001	0.528445	0.524422	0.505616	
6	15	(20,20,20)	logistic	0.0001	constant	lbfgs	0.0001	**0.467195**	**0.468398**	**0.45482**	UNDERFITTING
GBC	**# of variables**	**n_estimators**	**max_depth**	**min_samples_leaf**		**subsample**		**Train**	**Test**	**OOT**	
1	10	10	5	1		1		0.526357	0.517669	0.501509	
2	10	50	10	3		0.8		0.53505	0.524572	0.501844	
3	10	100	15	5		0.5		0.541146	0.519249	0.499162	
4	5	100	15	5		0.5		0.534136	0.517501	0.500671	
5	15	100	15	5		0.5		**0.543792**	**0.516226**	0.499329	OVERFITTING
Catboost	**# of variables**	**bootstrap_type**	**verbose**	**max_depth**		**iteration**	**random_state**	**Train**	**Test**	**OOT**	
1	10	none	0	2		5	None	0.540728	0.543309	0.513412	
2	10	Bayesian	0	5		10	10	0.522148	0.52321	0.500503	
3	10	Bayesian	0	16		15	10	0.526529	0.523115	0.501928	
4	10	MVS	0	16		20	5	0.528429	0.523049	0.503521	
5	20	MVS	0	16		20	5	0.531203	0.524023	0.504946	
XGBoost	**# of variables**	**max_depth**	**n_estimators**	**tree_method**		**subsample**		**Train**	**Test**	**OOT**	
1	10	2	5	auto		1		0.542421	0.543366	0.515507	
2	10	10	50	approx		0.8		0.530224	0.524862	0.507376	
3	10	20	100	auto		0.5		0.5378	0.524138	0.501425	
4	5	30	100	auto		0.5		0.529266	0.525789	0.5	
5	15	30	1000	hist		1		**0.545249**	**0.510646**	0.498324	OVERFITTING

VI. Summary of results

The final model we choose is LightGBM with the following parameters:

Number of variables = 10

n_estimators= 1000

Max_depth = 6

Num_leaves = 10

Col_samplebytree = 0.8

Learning_rate = 0.01

Eval_metric = logloss

RESULTS

	trn	tst	oot
0.0000	0.5315	0.5195	0.5080
1.0000	0.5291	0.5241	0.5071
2.0000	0.5268	0.5294	0.5050
3.0000	0.5257	0.5323	0.5042
4.0000	0.5265	0.5301	0.5063
Average	**0.5279**	**0.5271**	**0.5061**

The reason we chose this model is because the average train and test of this model is close to 0.525, while training is a bit better than testing, and OOT is near to 0.5.

Below are the three tables of the final model results (the first and last 10 bins of train, test, and OOT).

TRAIN

Total of Goods = 574997
Total of Bads = 8457
Fraud Rate (%Cumulative Bad) = 0.014707

Statistics by bin							Cumulative statistics					
bin	#records	#good	#bad	%good	%bad	total	cumulative good	cumulative bad	%cumulative good	FDR	KS	FPR
0	0	0	0	0.00	0.00	0	0	0	0.00	0.000000	0.000000	0.000000
1	5835	1591	4244	27.27	72.73	5835	1591	4244	0.28	50.183280	49.906583	0.374882
2	5834	5686	148	97.46	2.54	11669	7277	4392	1.27	51.933310	50.667738	1.656876
3	5835	5774	61	98.95	1.05	17504	13051	4453	2.27	52.654606	50.384855	2.930833
4	5834	5779	55	99.06	0.94	23338	18830	4508	3.27	53.304954	50.030155	4.177019
5	5835	5781	54	99.07	0.93	29173	24611	4562	4.28	53.943479	49.663283	5.394783
6	5834	5802	32	99.45	0.55	35007	30413	4594	5.29	54.321864	49.032619	6.620157
7	5835	5797	38	99.35	0.65	40842	36210	4632	6.30	54.771195	48.473771	7.817358
8	5834	5790	44	99.25	0.75	46676	42000	4676	7.30	55.291475	47.987089	8.982036
9	5835	5794	41	99.30	0.70	52511	47794	4717	8.31	55.776280	47.464237	10.132287
10	5834	5792	42	99.28	0.72	58345	53586	4759	9.32	56.272910	46.953557	11.259929
...												
91	5834	5796	38	99.35	0.65	530943	522769	8174	90.92	96.653660	5.736838	63.955102
92	5835	5800	35	99.40	0.60	536778	528569	8209	91.93	97.067518	5.141995	64.388963
93	5834	5808	26	99.55	0.45	542612	534377	8235	92.94	97.374956	4.439340	64.890953
94	5835	5799	36	99.38	0.62	548447	540176	8271	93.94	97.800639	3.856496	65.309636
95	5834	5787	47	99.19	0.81	554281	545963	8318	94.95	98.356391	3.405809	65.636331
96	5835	5809	26	99.55	0.45	560116	551772	8344	95.96	98.663829	2.702980	66.127996
97	5834	5799	35	99.40	0.60	565950	557571	8379	96.97	99.077687	2.108312	66.543860
98	5835	5804	31	99.47	0.53	571785	563375	8410	97.98	99.444247	1.465475	66.988704
99	5834	5804	30	99.49	0.51	577619	569179	8440	98.99	99.798983	0.810814	67.438270
100	5835	5818	17	99.71	0.29	583454	574997	8457	100.00	100.000000	0.000000	67.990659

TEST

Total of Goods = 246504
Total of Bads = 3550
Fraud Rate (%Cumulative Bad) = 0.014401

Statistics by bin							Cumulative statistics					
bin	#records	#good	#bad	%good	%bad	total	cumulative good	cumulative bad	%cumulative good	FDR	KS	FPR
0	0	0	0	0.00	0.00	0	0	0	0.00	0.000000	0.000000	0.000000
1	2501	713	1788	28.51	71.49	2501	713	1788	0.29	50.366197	50.076951	0.398770
2	2500	2446	54	97.84	2.16	5001	3159	1842	1.28	51.887324	50.605798	1.714984
3	2501	2461	40	98.40	1.60	7502	5620	1882	2.28	53.014085	50.734193	2.986185
4	2500	2478	22	99.12	0.88	10002	8098	1904	3.29	53.633803	50.348650	4.253151
5	2501	2481	20	99.20	0.80	12503	10579	1924	4.29	54.197183	49.905552	5.498441
6	2500	2484	16	99.36	0.64	15003	13063	1940	5.30	54.647887	49.348560	6.733505
7	2501	2479	22	99.12	0.88	17504	15542	1962	6.30	55.267606	48.962611	7.921509
8	2500	2484	16	99.36	0.64	20004	18026	1978	7.31	55.718310	48.405620	9.113246
9	2501	2489	12	99.52	0.48	22505	20515	1990	8.32	56.056338	47.733924	10.309045
10	2500	2486	14	99.44	0.56	25005	23001	2004	9.33	56.450704	47.119783	11.477545
...												
91	2500	2494	6	99.76	0.24	227548	224115	3433	90.92	96.704225	5.786468	65.282552
92	2501	2487	14	99.44	0.56	230049	226602	3447	91.93	97.098592	5.171921	65.738903
93	2500	2481	19	99.24	0.76	232549	229083	3466	92.93	97.633803	4.700654	66.094345
94	2501	2480	21	99.16	0.84	235050	231563	3487	93.94	98.225352	4.286130	66.407514
95	2500	2493	7	99.72	0.28	237550	234056	3494	94.95	98.422535	3.471967	66.987979
96	2501	2491	10	99.60	0.40	240051	236547	3504	95.96	98.704225	2.743121	67.507705
97	2500	2489	11	99.56	0.44	242551	239036	3515	96.97	99.014085	2.043257	68.004552
98	2501	2493	8	99.68	0.32	245052	241529	3523	97.98	99.239437	1.257262	68.557763
99	2500	2487	13	99.48	0.52	247552	244016	3536	98.99	99.605634	0.614546	69.009050
100	2501	2487	14	99.44	0.56	250053	246503	3550	100.00	100.000000	0.000000	69.437465

OOT

Total of Goods = 164107
Total of Bads = 2386
Fraud Rate (%Cumulative Bad) = 0.014539

Statistics by bin							Cumulative statistics					
bin	#records	#good	#bad	%good	%bad	total	cumulative good	cumulative bad	%cumulative good	FDR	KS	FPR
0	0	0	0	0.00	0.00	0	0	0	0.00	0.000000	0.000000	0.000000
1	1665	511	1154	30.69	69.31	1665	511	1154	0.31	48.365465	48.054083	0.442808
2	1665	1638	27	98.38	1.62	3330	2149	1181	1.31	49.497066	48.187555	1.819644
3	1665	1638	27	98.38	1.62	4995	3787	1208	2.31	50.628667	48.321026	3.134934
4	1665	1645	20	98.80	1.20	6660	5432	1228	3.31	51.466890	48.156855	4.423453
5	1665	1658	7	99.58	0.42	8325	7090	1235	4.32	51.760268	47.439916	5.740891
6	1665	1653	12	99.28	0.72	9990	8743	1247	5.33	52.263202	46.935580	7.011227
7	1665	1655	10	99.40	0.60	11655	10398	1257	6.34	52.682313	46.346204	8.272076
8	1664	1649	15	99.10	0.90	13319	12047	1272	7.34	53.310981	45.970039	9.470912
9	1665	1653	12	99.28	0.72	14984	13700	1284	8.35	53.813915	45.465703	10.669782
10	1665	1655	10	99.40	0.60	16649	15355	1294	9.36	54.233026	44.876326	11.866306
...												
91	1665	1658	7	99.58	0.42	151509	149210	2299	90.92	96.353730	5.431344	64.902131
92	1665	1651	14	99.16	0.84	153174	150861	2313	91.93	96.940486	5.012049	65.223087
93	1664	1651	13	99.22	0.78	154838	152512	2326	92.93	97.485331	4.550843	65.568358
94	1665	1657	8	99.52	0.48	156503	154169	2334	93.94	97.820620	3.876425	66.053556
95	1665	1662	3	99.82	0.18	158168	155831	2337	94.96	97.946354	2.989405	66.679932
96	1665	1656	9	99.46	0.54	159833	157487	2346	95.97	98.323554	2.357508	67.130009
97	1665	1654	11	99.34	0.66	161498	159141	2357	96.97	98.784577	1.810651	67.518456
98	1665	1655	10	99.40	0.60	163163	160796	2367	97.98	99.203688	1.221274	67.932404
99	1665	1655	10	99.40	0.60	164828	162451	2377	98.99	99.622800	0.631897	68.342869
100	1665	1656	9	99.46	0.54	166493	164107	2386	100.00	100.000000	0.000000	68.779128