Project 1 Documentation

I. Description of Data

The data contains individuals personal identifying information (PII) that is used to identify potential credit card application fraud. It contains records from Jaunuary 1st 2017 to December 31st 2017. There are 10 fields and 1,000,000 records available.

a. Numerical Table

Field Name	% Populated	Min	Max	Std Dev	% Zero
Date	100 %	2017-01-01	2017-12-31	/	0 %
D.o.b	100 %	1900-01-01	2016-10-31	/	0 %

b. Categorical Table

Field Name	% 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

Frivolous values in datasets are identical values that may cause false linkages in variables. Such identical values might be imputed as a method to fill in missing data, however, they might lead to inaccurate analysis. We addressed this by converting dates to datetime format, right-aligning zip codes, and modifying frivolous values in the address, SSN, DOB, and homephone fields. We used the record number as a unique identifier to prevent frivolous values from being linked to previous values.

III. Variable Creation

Identity fraud typically involves two core entities - the fraudster and the victim. There are three primary types of identity fraud:

a. Identity theft: The fraudster impersonates as the victim through victim's personal information obtained
b. Identity manipulation: The fraudster alters existing personal information, which may be their own
c. Synthetic identity: The fraudster creates a completely false identity

In response, various unique identifiers variables are created based on linking personal identifying information (PII) such as last name, address, and others. Additional variables are made to track the velocity of similar records and monitor the length of time since the groups appeared. By doing so, we can identify patterns and monitor variations in the data.

* Variables from maximum indicator are later removed from the feature selection process as it poses a concern of target leak.

Description of Variables	# Variables Created
Age when Apply Contains age of the applicants when application was made.	1
Day of Week Target Encoding Contains target encoded day of week (DOW), where dow_risk demonstrates risk of fraud in the particular day (exp: Monday, Tuesday...)	1
Days Since Variable Amount of days since an application with the applicants occurred.	23
Velocity Amount of records with the same attributes over the past [0,1,3,7,14,30] days.	138
Relative Velocity Demonstrates frequency where an attribute is appearing in the recent days [0,1] versus other days [3,7,14,30].	184
Count by entities Count the amount of unique values for one attribute in comparison to other attribute over the time period of [0,1,3,7,14,30,60] days.	3,542
Maximum Indicator* The maximum count of the occurrences of each attribute within a certain period of time.	92
Age Indicator The age of the applicants when they applied, and it contains the min, max and mean age of the applicants by each attributes.	69
Total Variables Before removing duplicates. Including 2 orignial fields of 'Record' and 'fraud_label', however, it is not considered as a independent variable.	4052

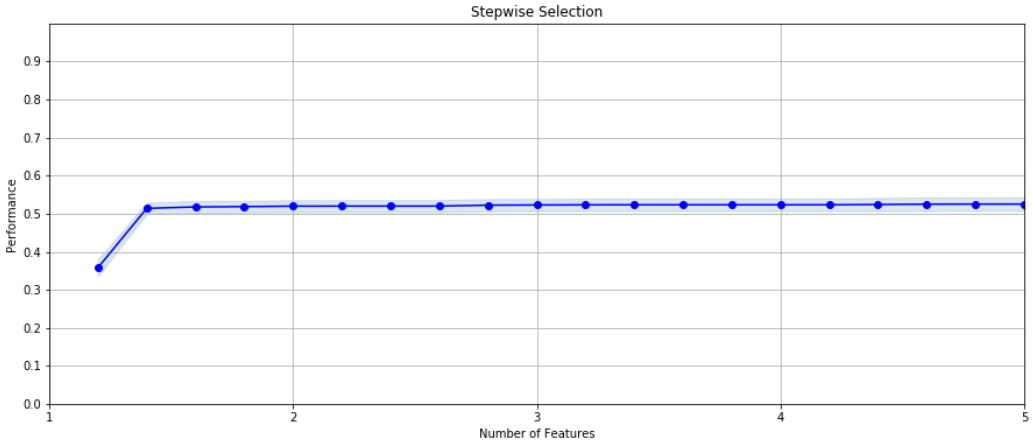
IV. Feature Selection

To improve model performance, we select some most relevant features to reduce dimensionality in the model, it also allows for flexible model exploration. The results yielded 20 variables, sorted by its filter score. In the process, we first

a. 'Filter' the variables: By sorting and evaluating variables by its relevance to predicting the dependent variable, with univariate KS method to normalize and separate into two classes.
b. 'Wrapped' model around features: By creating a non-linear model, with stepwise selection, to measure model performance by fraud detection rate (FDR).
c. List of final variables along with respective univariate KS's (filter score)

Wrapper Order	Variable	Filter Score
1	fulladdress_day_since	0.333269
2	ssn_dob_day_since	0.228626
3	fulladdress_unique_count_for_ssn_name_30	0.281933
4	address_count_30	0.332648
5	fulladdress_count_7	0.301666
6	fulladdress_unique_count_for_ssn_dob_14	0.276209
7	address_count_14	0.322436
8	fulladdress_unique_count_for_ssn_homephone_60	0.289991
9	ssn_count_30	0.226894
10	address_unique_count_for_name_homephone_30	0.284516
11	address_unique_count_for_ssn_zip5_7	0.273248
12	fulladdress_unique_count_for_ssn_lastname_30	0.281881
13	address_day_since	0.33414
14	address_count_0_by_30	0.291922
15	fulladdress_count_0_by_30	0.290722
16	address_unique_count_for_ssn_homephone_60	0.289166
17	address_unique_count_for_name_homephone_60	0.292438
18	address_unique_count_for_dob_homephone_60	0.287556
19	fulladdress_unique_count_for_ssn_60	0.286764
20	address_unique_count_for_homephone_name_dob_60	0.29141

d. Model Performance Graph



V. Preliminary Model Exploration

There are 7 models in total that are explored in this summary report. Different number of variables and parameter values are utilized to explore model performance, especially to observe its effect on over-fitting or under-fitting. The baseline logistic regression yielded a result of 0.48, in regard to keeping FDR at 3%. The at-best result for train and test set are around 0.52-0.53, while certain overfitting examples pushed train sets to 0.54, but nothing beyond 0.55. The OOT sets performance averages around 0.49-0.50.

Model	Iteration	Dataset NVARS	Parameters						Average FDR at 3%			
									Train	Test	OOT	
Logistic Regression	1	10	max_iter: 20	penalty: NA	c: NA	solver: NA	l1_ratio: NA		0.4889	0.4858	0.4730	
	2	10	20	l2	1	lbfgs	None		0.4878	0.4886	0.4733	
	3	15	20	l2	1	saga	None		0.4797	0.4783	0.4653	
	4	10	20	l1	0.5	saga	None		0.4889	0.4795	0.4711	
	5	15	20	elasticnet	0.5	saga	0.4		0.4840	0.4770	0.4679	
Single Decision Tree	Iteration	NVARS	max_depth		min_samples_split		min_samples_leaf		Train	Test	OOT	
	1	15	5		50		30		0.4765	0.4809	0.4530	UNDER-FITTING
	2	10	10		40		24		0.5289	0.5238	0.5041	
	3	10	20		30		14		0.5358	0.5221	0.5017	
	4	15	25		20		8		0.5434	0.5173	0.5008	OVER-FITTING
	5	15	30		5		4		0.5412	0.5221	0.4985	OVER-FITTING
Random Forest	Iteration	NVARS	max_depth	min_samples_split	min_samples_leaf	max_features	bootstrap	n_estimators	Train	Test	OOT	
	1	10	2	50	30	4	TRUE	3	0.4402	0.4365	0.4120	UNDER-FITTING
	2	10	10	40	24	5	TRUE	15	0.5250	0.5219	0.5036	
	3	10	20	30	14	6	TRUE	40	0.5276	0.5256	0.5021	
	4	15	20	20	10	12	TRUE	70	0.5417	0.5211	0.5013	OVER-FITTING
	5	15	30	5	4	15	TRUE	100	0.5431	0.5213	0.5013	OVER-FITTING
Nueral Net (NN)	Iteration	NVARS	hidden_layer_size	activation	alpha	learning_rate	solver	learning_rate_init	Train	Test	OOT	
	1	10	5	logistic	0.1	constant	adam	0.01	0.5023	0.5026	0.4834	
	2	15	5	relu	0.1	adaptive	lbfgs	0.01	0.5232	0.5212	0.5011	
	3	15	20, 20, 20	logistic	0.01	constant	sgd	0.001	0.4785	0.4678	0.4421	UNDER-FITTING
	4	10	20, 20, 20	relu	0.001	adaptive	lbfgs	0.001	0.5282	0.5250	0.5048	
	5	15	10, 10	relu	0.001	constant	lbfgs	0.0001	0.5304	0.5191	0.5063	OVER-FITTING
LightGBM (Boost)	Iteration	NVARS	num_leaves				n_estimators		Train	Test	OOT	
	1	10	2				20		0.509	0.518	0.489	
	2	15	4				100		0.529	0.520	0.504	
	3	10	6				300		0.527	0.527	0.507	
	4	15	8				700		0.532	0.523	0.509	OVER-FITTING
	5	10	10				1000		0.532	0.532	0.506	
XGBoost	Iteration	NVARS	max_depth	n_estimators	tree_method	subsample	eta	eval_metrics	Train	Test	OOT	
	1	15	2	20	auto	1	0.3	logloss	0.5162	0.5183	0.4939	
	2	10	3	100	exact	0.8	0.2	logloss	0.5273	0.5300	0.5069	
	3	15	4	300	approx	0.8	0.3	logloss	0.5364	0.5237	0.5037	
	4	10	10	700	auto	0.8	0.2	logloss	0.5439	0.5109	0.4942	OVER-FITTING
	5	15	30	100	auto	1	0.3	logloss	0.5273	0.5241	0.5124	
CatBoost	Iteration	NVARS	bootstrap_type	max_depth	iterations	l2_leaf_reg	verbose	random_state	Train	Test	OOT	
	1	10	Bayesian	2	5	3	0	none	0.4652	0.4701	0.4536	UNDER-FITTING
	2	15	MVS	5	10	6	0	8	0.5017	0.4989	0.4785	
	3	10	Bayesian	8	45	8	0	10	0.5182	0.5180	0.4955	
	4	15	Bayesian	10	100	12	0	8	0.5234	0.5217	0.4985	
	5	10	MVS	15	30	14	0	3	0.5214	0.5214	0.4988	

VI. Summary of Results

a. **Final Model:** LightGBM – a gradient boosting framework that uses tree-based learning algorithms.

b. **Number of Variables (NVARS)**: 12

c. **Hyperparameters:** num_leaves=15, n_estimators=400, max_depth=0, learning_rate=0.1, objective=None, min_split_gain=0.0, min_child_weight=0.001

d. **List of final variables:**

No.	Variables
1	fulladdress_day_since
2	name_dob_count_30
3	address_unique_count_for_name_homephone_60
4	fulladdress_unique_count_for_dob_homephone_3
5	address_unique_count_for_homephone_name_dob_30
6	address_unique_count_for_ssn_name_dob_14
7	address_day_since
8	address_count_14
9	address_count_7
10	address_count_0_by_30
11	address_unique_count_for_homephone_name_dob_60
12	fulladdress_count_0_by_30

e. **Description of results:** We have divided the data into 3 sets – training, testing and out-of-time (OOT), where each set yields bin and cumulative statistics, which produces result based on individual population bins, and running results cumulated. The fraud rate is derived from amount of fraud records, among total records.

By declining top 3% of the applicants, we are able to eliminate ~50% of fraud in the rejected applications, with:

- Fraud rate of 0.0144 for training set
- Fraud rate of 0.0143 for testing set
- Fraud rate of 0.0143 for oot set

Model = LGBMClassifier
(num_leaves=15, n_estimators=400, max_depth=0, learning_rate=0.1, objective=None, min_split_gain=0.0, min_child_weight=0.001)

Training	# Records	# Goods	# Bads	Fraud Rate
	583,454	575,017	8,437	0.014460437

Population Bin %	Bin Statistics					Cumulative Statistics						
	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
1	5835	1553	4282	26.62%	73.38%	5835	1553	4282	0.27%	50.75%	50.48	0.36
2	5834	5678	156	97.33%	2.67%	11669	7231	4438	1.26%	52.60%	51.34	1.63
3	5835	5770	65	98.89%	1.11%	17504	13001	4503	2.26%	53.37%	51.11	2.89
4	5834	5784	50	99.14%	0.86%	23338	18785	4553	3.27%	53.96%	50.70	4.13
5	5835	5783	52	99.11%	0.89%	29173	24568	4605	4.27%	54.58%	50.31	5.34
6	5834	5793	41	99.30%	0.70%	35007	30361	4646	5.28%	55.07%	49.79	6.53
7	5835	5791	44	99.25%	0.75%	40842	36152	4690	6.29%	55.59%	49.30	7.71
8	5834	5796	38	99.35%	0.65%	46676	41948	4728	7.30%	56.04%	48.74	8.87
9	5835	5804	31	99.47%	0.53%	52511	47752	4759	8.30%	56.41%	48.10	10.03
10	5834	5803	31	99.47%	0.53%	58345	53555	4790	9.31%	56.77%	47.46	11.18
11	5835	5795	40	99.31%	0.69%	64180	59350	4830	10.32%	57.25%	46.93	12.29
12	5834	5793	41	99.30%	0.70%	70014	65143	4871	11.33%	57.73%	46.40	13.37
13	5835	5783	52	99.11%	0.89%	75849	70926	4923	12.33%	58.35%	46.02	14.41
14	5835	5788	47	99.19%	0.81%	81684	76714	4970	13.34%	58.91%	45.57	15.44
15	5834	5797	37	99.37%	0.63%	87518	82511	5007	14.35%	59.35%	45.00	16.48
16	5835	5793	42	99.28%	0.72%	93353	88304	5049	15.36%	59.84%	44.49	17.49
17	5834	5790	44	99.25%	0.75%	99187	94094	5093	16.36%	60.37%	44.00	18.48
18	5835	5787	48	99.18%	0.82%	105022	99881	5141	17.37%	60.93%	43.56	19.43
19	5834	5796	38	99.35%	0.65%	110856	105677	5179	18.38%	61.38%	43.01	20.40
20	5835	5803	32	99.45%	0.55%	116691	111480	5211	19.39%	61.76%	42.38	21.39

Model = LGBMClassifier
(num_leaves=15, n_estimators=400, max_depth=0, learning_rate=0.1, objective=None, min_split_gain=0.0, min_child_weight=0.001)

Testing	# Records	# Goods	# Bads	Fraud Rate
	250,053	246,483	3,570	0.014276973

Population Bin %	Bin Statistics					Cumulative Statistics						
	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
1	2501	727	1774	29.07%	70.93%	2501	727	1774	0.29%	49.69%	49.40	0.41
2	2500	2445	55	97.80%	2.20%	5001	3172	1829	1.29%	51.23%	49.95	1.73
3	2501	2462	39	98.44%	1.56%	7502	5634	1868	2.29%	52.32%	50.04	3.02
4	2500	2488	12	99.52%	0.48%	10002	8122	1880	3.30%	52.66%	49.37	4.32
5	2501	2487	14	99.44%	0.56%	12503	10609	1894	4.30%	53.05%	48.75	5.60
6	2500	2488	12	99.52%	0.48%	15003	13097	1906	5.31%	53.39%	48.08	6.87
7	2501	2479	22	99.12%	0.88%	17504	15576	1928	6.32%	54.01%	47.69	8.08
8	2500	2486	14	99.44%	0.56%	20004	18062	1942	7.33%	54.40%	47.07	9.30
9	2501	2480	21	99.16%	0.84%	22505	20542	1963	8.33%	54.99%	46.65	10.46
10	2500	2487	13	99.48%	0.52%	25005	23029	1976	9.34%	55.35%	46.01	11.65
11	2501	2492	9	99.64%	0.36%	27506	25521	1985	10.35%	55.60%	45.25	12.86
12	2500	2478	22	99.12%	0.88%	30006	27999	2007	11.36%	56.22%	44.86	13.95
13	2501	2486	15	99.40%	0.60%	32507	30485	2022	12.37%	56.64%	44.27	15.08
14	2500	2484	16	99.36%	0.64%	35007	32969	2038	13.38%	57.09%	43.71	16.18
15	2501	2482	19	99.24%	0.76%	37508	35451	2057	14.38%	57.62%	43.24	17.23
16	2500	2485	15	99.40%	0.60%	40008	37936	2072	15.39%	58.04%	42.65	18.31
17	2501	2486	15	99.40%	0.60%	42509	40422	2087	16.40%	58.46%	42.06	19.37
18	2501	2477	24	99.04%	0.96%	45010	42899	2111	17.40%	59.13%	41.73	20.32
19	2500	2487	13	99.48%	0.52%	47510	45386	2124	18.41%	59.50%	41.08	21.37
20	2501	2478	23	99.08%	0.92%	50011	47864	2147	19.42%	60.14%	40.72	22.29

Model = LGBMClassifier
(num_leaves=15, n_estimators=400, max_depth=0, learning_rate=0.1, objective=None, min_split_gain=0.0, min_child_weight=0.001)

OOT	# Records	# Goods	# Bads	Fraud Rate
	166,493	164,107	2,386	0.014330933

Population Bin %	Bin Statistics					Cumulative Statistics						
	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
1	1665	509	1156	30.57%	69.43%	1665	509	1156	0.31%	48.45%	48.14	0.44
2	1665	1638	27	98.38%	1.62%	3330	2147	1183	1.31%	49.58%	48.27	1.81
3	1665	1647	18	98.92%	1.08%	4995	3794	1201	2.31%	50.34%	48.02	3.16
4	1665	1647	18	98.92%	1.08%	6660	5441	1219	3.32%	51.09%	47.77	4.46
5	1665	1653	12	99.28%	0.72%	8325	7094	1231	4.32%	51.59%	47.27	5.76
6	1665	1654	11	99.34%	0.66%	9990	8748	1242	5.33%	52.05%	46.72	7.04
7	1665	1659	6	99.64%	0.36%	11655	10407	1248	6.34%	52.31%	45.96	8.34
8	1664	1655	9	99.46%	0.54%	13319	12062	1257	7.35%	52.68%	45.33	9.60
9	1665	1658	7	99.58%	0.42%	14984	13720	1264	8.36%	52.98%	44.62	10.85
10	1665	1657	8	99.52%	0.48%	16649	15377	1272	9.37%	53.31%	43.94	12.09
11	1665	1650	15	99.10%	0.90%	18314	17027	1287	10.38%	53.94%	43.56	13.23
12	1665	1653	12	99.28%	0.72%	19979	18680	1299	11.38%	54.44%	43.06	14.38
13	1665	1655	10	99.40%	0.60%	21644	20335	1309	12.39%	54.86%	42.47	15.53
14	1665	1652	13	99.22%	0.78%	23309	21987	1322	13.40%	55.41%	42.01	16.63
15	1665	1654	11	99.34%	0.66%	24974	23641	1333	14.41%	55.87%	41.46	17.74
16	1665	1652	13	99.22%	0.78%	26639	25293	1346	15.41%	56.41%	41.00	18.79
17	1665	1655	10	99.40%	0.60%	28304	26948	1356	16.42%	56.83%	40.41	19.87
18	1665	1653	12	99.28%	0.72%	29969	28601	1368	17.43%	57.33%	39.91	20.91
19	1665	1655	10	99.40%	0.60%	31634	30256	1378	18.44%	57.75%	39.32	21.96
20	1665	1645	20	98.80%	1.20%	33299	31901	1398	19.44%	58.59%	39.15	22.82