Project 1 Documentation

Data Description

The dataset is about Credit Card Application Data. The data contains synthetic personal identifying information (PII) on each application. The data covers the time of year 2017, and there are 10 fields and 1,000,000 records. It contains both numerical and categorical field, such as date of application, date of birth, first name, and last name. All the fields are shown in the following summary tables.

Numerical Table

Field Name	% Populated	Min	Max	Mean	Stdev	% Zero
date	100.00	2017-01-01	2017-12-31	/	/	0.00
dob	100.00	1900-01-01	2016-10-31	/	/	0.00

Categorical Table

Field Name	% 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

Data Cleaning

Having gained some insights from the data summaries, there appears to be frivolous values, typical value filled for missing fields, in the data. After examining the situation and consulting industry experts, it is understood that companies fill in missing fields with a dummy placeholder and these frivolous values should be transformed before conducting further investigation.
The dummy placeholder for this dataset:

- date of birth (dob) - “1907-06-26”
- social security number (ssn) - “9999999999”
- address - “123 MAIN ST”
- home phone - “9999999999”

All of the frivolous values are replaced with a unique string that includes the record number of that application and number of leading zeros that help to format the final value. After transforming the frivolous fields, an incoming application will not appear to be overly risky when it has missing fields that are replaced by dummy placeholders.

Variable Creation

There three mode of identity fraud, synthetic identity, identity manipulation, identity theft. Synthetic identity fraud is when people apply for a product using a synthesized or made-up identity. Identity manipulation is when people use someone else's true identity with slight modifications to apply for a product. Identity theft is when people apply for a product using a stolen identity that's not their true identity. The case being investigated in this project is identity theft as companies would like to catch potential identity fraud that would cause damage to their profit.

In order to detect potential frauds, looking at fields independently are not sufficient. Variables are computed statistically from the fields in order to detect anomaly, for instance, the age of the applicants are computed from their date of birth and the applicants who are relatively older in age are more likely to be fraudulent. On the other hand, same values in certain fields may be used with multiple applications and this may indicate a fraudulent application. Thus, fields are linked together to produce variables that represent the frequency and velocity that certain entity values have appeared on past applications. Examples of these frequency and velocity variables are the number of application records that have the same field value or combined field values for the past day, week, or month. Age indicators are also created to examine the maximum, average, and minimum applicant age associated with the entity values. Lastly, a target encoded variable, day of week, is created to replace the categorical field, date of application, to understand the average fraud percentage on each weekday. There are a total of 3,958 created variables and their descriptions are shown in the following chart.

Description of Variables	# Variables Created
Age when apply (age of the applicant at application)	1
Date of week target encoded (average fraud percentage of that day)	1
Days since Variables: # days since an application with that entity has been seen.	23
Velocity: # records with the same entity over the last {0, 1, 3, 7, 14, 30} days	138
Relative velocity: ratio of the short-term velocity over the last {0, 1} days to a longer-term averaged velocity over the last {3, 7, 14, 30} days	184
Number of unique: # records with the same entity over the past (0, 1, 3, 7, 14, 30} days	3542
Age indicator: max, mean, min applicant age associated with the fields	69

Feature Selection

While using all of the 3,958 variables may generate an outstanding model, it would increase the complexity of the model excessively and make it hard to fit nonlinear models. Therefore, it is important to reduce the dimensionality, number of independent variables, and select the best few candidate variables to reduce the difficulty of modelling. The type of feature selection process used in this project contains two steps: filter and wrapper. The filter considers all the candidate independent variables univariately and sort them by their importance for predicting the dependent variable, fraud label for each application. After filtering the all the candidate independent variables down to a few hundreds, a wrapper looks for good subsets of variables by their multivariate importance, taking into account the correlation between variables. In this project, 3,958 candidate variables are filtered to 224 variables and result in 25 final sorted variables after a wrapper. The final 25 variables and their sorted filter score, importance for predicting the dependent variable, are listed in the following table.

wrapper order	variable	filter score
1	fulladdress_day_since	0.3333
2	ssn_dob_day_since	0.2286
3	fulladdress_unique_count_for_ssn_name_30	0.2819
4	zip5_unique_count_for_fulladdress_dob_1	0.2191
5	fulladdress_count_7	0.3017
6	ssn_firstname_count_30	0.2260
7	fulladdress_unique_count_for_name_homephone_60	0.2895
8	name_dob_day_since	0.2281
9	fulladdress_unique_count_for_ssn_homephone_30	0.2841
10	address_unique_count_for_ssn_lastname_30	0.2818
11	address_day_since	0.3341
12	address_count_30	0.3326
13	address_count_14	0.3224
14	address_count_0_by_30	0.2919
15	fulladdress_count_0_by_30	0.2907
16	fulladdress_unique_count_for_homephone_name_dob_60	0.2885
17	fulladdress_unique_count_for_dob_homephone_60	0.2884
18	address_unique_count_for_dob_homephone_60	0.2876
19	address_unique_count_for_name_dob_60	0.2859
20	fulladdress_unique_count_for_ssn_name_dob_60	0.2847
21	fulladdress_unique_count_for_ssn_dob_60	0.2847
22	fulladdress_unique_count_for_name_60	0.2845
23	address_unique_count_for_homephone_name_dob_30	0.2840
24	address_unique_count_for_ssn_dob_60	0.2838
25	fulladdress_unique_count_for_name_homephone_30	0.2836

Preliminary Model Exploration

Before start modelling with the selected variables after wrapper, variables are standardized using z-scaling to smoothen the distribution and remove extreme outliers. The original dataset is then split into three subsets, training, testing, and out of time data. With the selected variables, different model algorithms and hyperparameters are used to build preliminary models. Logistic regression models are initially built as a baseline for the other nonlinear models. Decision tree, random forest, three boosted trees (GBC, LGBM, XGB), and neural network models are built to explore model performance on different hyperparameters. Each choice of model and hyperparameters is performed 5 times and the average fraud detection rates (FDR) at 3 percent for the training, testing, and out of time serve as the measure of goodness. The following table represents performance of each chosen model and hyperparameters.

Model		Parameter								Average FDR at 3%		
Logistic Regression	Iteration	NVARS	Penalty	C		Solver		l1_ratio	Training/Testing Split	Train	Test	OOT
	1	6	l2	1		lbfgs		None	0.3	0.477	0.477	0.462
	2	10	l2	1		lbfgs		None	0.3	0.492	0.480	0.474
	3	15	l2	1		lbfgs		None	0.3	0.479	0.483	0.466
	4	20	l2	1		lbfgs		None	0.3	0.483	0.481	0.468
Decision Tree	Iteration	NVARS	Criterion	Max_depth		Min_samples_split		Min_samples_leaf		Train	Test	OOT
	1	6	gini	20		20		10		0.534	0.515	0.502
	2	10	gini	10		50		30		0.531	0.520	0.505
	3	10	gini	7		40		20		0.526	0.520	0.502
	4	10	entropy	10		50		30		0.530	0.525	0.504
	5	15	gini	1		10		5		0.249	0.241	0.229
	6	20	gini	10		20		10		0.530	0.524	0.507
	7	20	entropy	10		20		10		0.532	0.522	0.506
Random Forest	Iteration	NVARS	Criterion	n_estimators	Max_depth	Min_samples_split		Min_samples_leaf		Train	Test	OOT
	1	6	gini	50	10	40		20		0.525	0.523	0.502
	2	10	gini	50	10	40		20		0.531	0.523	0.505
	3	10	entropy	50	10	40		20		0.534	0.522	0.505
	4	15	gini	300	2	40		20		0.506	0.504	0.481
	5	20	gini	200	200	2		1		0.545	0.520	0.500
	6	20	entropy	100	10	40		20		0.530	0.528	0.504
Gradient Boosting Classifier	Iteration	NVARS	Criterion	n_estimators	Max_depth	Min_samples_split		Min_samples_leaf		Train	Test	OOT
	1	6	friedman_mse	50	2	40		20		0.515	0.512	0.492
	2	10	friedman_mse	100	3	40		20		0.529	0.522	0.503
	3	10	squared_error	100	4	40		20		0.528	0.526	0.506
	4	15	friedman_mse	600	10	40		20		0.544	0.518	0.498
	5	20	friedman_mse	100	3	2		1		0.530	0.523	0.506
	6	20	squared_error	100	4	40		20		0.529	0.526	0.505
LightGBM	Iteration	NVARS	Max_depth	n_estimators		Num_leaves		Learning Rate		Train	Test	OOT
	1	6	6	50		6		0.1		0.525	0.523	0.502
	2	10	6	400		6		0.1		0.530	0.523	0.508
	3	10	6	400		30		0.1		0.530	0.528	0.505
	4	15	1	50		30		0.1		0.513	0.511	0.488
	5	20	5	100		50		0.1		0.530	0.526	0.509
	6	20	20	500		100		0.1		0.534	0.524	0.504
XGBoost	Iteration	NVARS	Booster	Tree_method	Max_depth	Min_child_weight		n_estimators		Train	Test	OOT
	1	6	gbtree	hist	2	2		10		0.507	0.507	0.485
	2	10	gbtree	auto	6	1		100		0.536	0.521	0.503
	3	10	gbtree	approx	5	10		100		0.530	0.525	0.506
	4	10	dart	approx	5	10		100		0.531	0.521	0.506
	5	15	gbtree	exact	15	2		40		0.540	0.520	0.503
	6	20	gbtree	auto	5	20		200		0.533	0.526	0.508
Neural Network	Iteration	NVARS	hidden_layer_sizes	activation		alpha	learning_rate	solver	learning_rate_init	Train	Test	OOT
	1	6	(20,20,20)	logistic		0.001	constant	lbfgs	0.005	0.416	0.416	0.397
	2	10	(5)	relu		0.001	constant	adam	0.01	0.523	0.521	0.502
	3	10	(100)	logistic		0.001	adaptive	adam	0.01	0.524	0.520	0.499
	4	15	(5)	relu		0.001	adaptive	adam	0.01	0.526	0.520	0.502
	5	15	(20,20,20)	logistic		0.001	constant	adam	0.001	0.523	0.519	0.498
	6	20	(20,20,20)	relu		0.001	constant	adam	0.001	0.528	0.523	0.508

Result Summary

After exploring models and hyperparameters, findings are that as the model complexity, such as number of variables and depth of trees, increases, measure of goodness for the model also increases. However, once models become overly complex, the models start to overfitting and perform not as well on predicting the outcome of the training and out of time sets. On the opposite end, too little model complexity would lead to underfitting and result in models that are not performing less efficiently on all three subsets of data.

By comparing all the models built in the model exploration section, a final model is determined. The final model is a boosted tree using the LightGBM algorithm, which is a boosting ensemble learning method that combines all the simple trees built iteratively to minimize training errors. The hyperparameters used are maximum depth of 6, maximum tree leaves of 6 for base learners, 400 boosted trees to fit, and a learning rate of 0.1. The following table shows the list of ten variables and their filter score used in the final model.

Variable Number	Variable Name	filter score
1	fulladdress_day_since	0.3333
2	ssn_dob_day_since	0.2286
3	fulladdress_unique_count_for_ssn_name_30	0.2819
4	zip5_unique_count_for_fulladdress_dob_1	0.2191
5	fulladdress_count_7	0.3017
6	ssn_firstname_count_30	0.2260
7	fulladdress_unique_count_for_name_homephone_60	0.2895
8	name_dob_day_since	0.2281
9	fulladdress_unique_count_for_ssn_homephone_30	0.2841
10	address_unique_count_for_ssn_lastname_30	0.2818

The following three tables show model performance on the training, testing, and out of time sets. This model can achieve a fraud detection rate at 3 percent of 52.9%, 52.6%, and 50.7% for training, testing, and out of time perspectives. This indicates that the model is able to reject only three percent of the applications and catch 50.7% of the fraud.

Training	# Records		# Goods		# Bads		Fraud Rate					
	583454		575126		8328		0.0143					
	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulatve Bads	% Cumulative Goods	% Cumulative Bads (FDR)	KS	FPR
1	5835	1639	4196	28.09%	71.91%	5835	1639	4196	0.28%	50.38%	50.10	0.39
2	5834	5691	143	97.55%	2.45%	11669	7330	4339	1.27%	52.10%	50.83	1.69
3	5835	5763	72	98.77%	1.23%	17504	13093	4411	2.28%	52.97%	50.69	2.97
4	5834	5789	45	99.23%	0.77%	23338	18882	4456	3.28%	53.51%	50.22	4.24
5	5835	5803	32	99.45%	0.55%	29173	24685	4488	4.29%	53.89%	49.60	5.50
6	5834	5789	45	99.23%	0.77%	35007	30474	4533	5.30%	54.43%	49.13	6.72
7	5835	5791	44	99.25%	0.75%	40842	36265	4577	6.31%	54.96%	48.65	7.92
8	5834	5785	49	99.16%	0.84%	46676	42050	4626	7.31%	55.55%	48.24	9.09
9	5835	5798	37	99.37%	0.63%	52511	47848	4663	8.32%	55.99%	47.67	10.26
10	5834	5796	38	99.35%	0.65%	58345	53644	4701	9.33%	56.45%	47.12	11.41
11	5835	5792	43	99.26%	0.74%	64180	59436	4744	10.33%	56.96%	46.63	12.53
12	5834	5795	39	99.33%	0.67%	70014	65231	4783	11.34%	57.43%	46.09	13.64
13	5835	5781	54	99.07%	0.93%	75849	71012	4837	12.35%	58.08%	45.73	14.68
14	5835	5793	42	99.28%	0.72%	81684	76805	4879	13.35%	58.59%	45.23	15.74
15	5834	5792	42	99.28%	0.72%	87518	82597	4921	14.36%	59.09%	44.73	16.78
16	5835	5792	43	99.26%	0.74%	93353	88389	4964	15.37%	59.61%	44.24	17.81
17	5834	5790	44	99.25%	0.75%	99187	94179	5008	16.38%	60.13%	43.76	18.81
18	5835	5781	54	99.07%	0.93%	105022	99960	5062	17.38%	60.78%	43.40	19.75
19	5834	5800	34	99.42%	0.58%	110856	105760	5096	18.39%	61.19%	42.80	20.75
20	5835	5788	47	99.19%	0.81%	116691	111548	5143	19.40%	61.76%	42.36	21.69

Testing	# Records		# Goods		# Bads		Fraud Rate					
	250053		246374		3679		0.0147					
	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulatve Bads	% Cumulative Goods	% Cumulative Bads (FDR)	KS	FPR
1	2501	664	1837	26.55%	73.45%	2501	664	1837	0.27%	49.93%	49.66	0.36
2	2500	2455	45	98.20%	1.80%	5001	3119	1882	1.27%	51.16%	49.89	1.66
3	2501	2457	44	98.24%	1.76%	7502	5576	1926	2.26%	52.35%	50.09	2.90
4	2500	2483	17	99.32%	0.68%	10002	8059	1943	3.27%	52.81%	49.54	4.15
5	2501	2484	17	99.32%	0.68%	12503	10543	1960	4.28%	53.28%	49.00	5.38
6	2500	2473	27	98.92%	1.08%	15003	13016	1987	5.28%	54.01%	48.73	6.55
7	2501	2476	25	99.00%	1.00%	17504	15492	2012	6.29%	54.69%	48.40	7.70
8	2500	2471	29	98.84%	1.16%	20004	17963	2041	7.29%	55.48%	48.19	8.80
9	2501	2487	14	99.44%	0.56%	22505	20450	2055	8.30%	55.86%	47.56	9.95
10	2500	2481	19	99.24%	0.76%	25005	22931	2074	9.31%	56.37%	47.07	11.06
11	2501	2485	16	99.36%	0.64%	27506	25416	2090	10.32%	56.81%	46.49	12.16
12	2500	2483	17	99.32%	0.68%	30006	27899	2107	11.32%	57.27%	45.95	13.24
13	2501	2483	18	99.28%	0.72%	32507	30382	2125	12.33%	57.76%	45.43	14.30
14	2500	2476	24	99.04%	0.96%	35007	32858	2149	13.34%	58.41%	45.08	15.29
15	2501	2478	23	99.08%	0.92%	37508	35336	2172	14.34%	59.04%	44.70	16.27
16	2500	2480	20	99.20%	0.80%	40008	37816	2192	15.35%	59.58%	44.23	17.25
17	2501	2483	18	99.28%	0.72%	42509	40299	2210	16.36%	60.07%	43.71	18.23
18	2501	2489	12	99.52%	0.48%	45010	42788	2222	17.37%	60.40%	43.03	19.26
19	2500	2490	10	99.60%	0.40%	47510	45278	2232	18.38%	60.67%	42.29	20.29
20	2501	2488	13	99.48%	0.52%	50011	47766	2245	19.39%	61.02%	41.63	21.28

OOT	# Records		# Goods		# Bads		Fraud Rate					
	166493		164107		2386		0.0143					
	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulatve Bads	% Cumulative Goods	% Cumulative Bads (FDR)	KS	FPR
1	1665	510	1155	30.63%	69.37%	1665	510	1155	0.31%	48.41%	48.10	0.44
2	1665	1638	27	98.38%	1.62%	3330	2148	1182	1.31%	49.54%	48.23	1.82
3	1665	1639	26	98.44%	1.56%	4995	3787	1208	2.31%	50.63%	48.32	3.13
4	1665	1653	12	99.28%	0.72%	6660	5440	1220	3.31%	51.13%	47.82	4.46
5	1665	1649	16	99.04%	0.96%	8325	7089	1236	4.32%	51.80%	47.48	5.74
6	1665	1654	11	99.34%	0.66%	9990	8743	1247	5.33%	52.26%	46.94	7.01
7	1665	1655	10	99.40%	0.60%	11655	10398	1257	6.34%	52.68%	46.35	8.27
8	1664	1655	9	99.46%	0.54%	13319	12053	1266	7.34%	53.06%	45.71	9.52
9	1665	1656	9	99.46%	0.54%	14984	13709	1275	8.35%	53.44%	45.08	10.75
10	1665	1648	17	98.98%	1.02%	16649	15357	1292	9.36%	54.15%	44.79	11.89
11	1665	1652	13	99.22%	0.78%	18314	17009	1305	10.36%	54.69%	44.33	13.03
12	1665	1653	12	99.28%	0.72%	19979	18662	1317	11.37%	55.20%	43.83	14.17
13	1665	1655	10	99.40%	0.60%	21644	20317	1327	12.38%	55.62%	43.24	15.31
14	1665	1652	13	99.22%	0.78%	23309	21969	1340	13.39%	56.16%	42.77	16.39
15	1665	1655	10	99.40%	0.60%	24974	23624	1350	14.40%	56.58%	42.18	17.50
16	1665	1650	15	99.10%	0.90%	26639	25274	1365	15.40%	57.21%	41.81	18.52
17	1665	1650	15	99.10%	0.90%	28304	26924	1380	16.41%	57.84%	41.43	19.51
18	1665	1657	8	99.52%	0.48%	29969	28581	1388	17.42%	58.17%	40.76	20.59
19	1665	1653	12	99.28%	0.72%	31634	30234	1400	18.42%	58.68%	40.25	21.60
20	1665	1655	10	99.40%	0.60%	33299	31889	1410	19.43%	59.09%	39.66	22.62