Homework 6. Name: I-Chieh (Hazel) Huang

I. Data Description

The data is a synthetic application data. It contains applications for credit cards and cellphones with PII from 2017-01-01 to 2017-12-31. There are 10 fields and 1,000,000 application data. It contains 2 numerical fields and 8 categorical fields.

Numerical Table:

Field Name	% Populated	Min	Max	Mean	Stdev	% Zero
date	100	NA	NA	NA	NA	0
dob	100	NA	NA	NA	NA	0

Categorical Table:

Field Name	% Populated	# Unique Values	Most Common Value
record	100	1,000,000	NA
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	28244	9999999999
fraud_label	100	2	0

II. Data Cleaning

Sometimes, company put in a default data when dealing with null values in a data set. It helps data analyst to easily identify and clean the data afterwards. When summarizing the data, we can find out some odd patterns in it. In this data, SSN has a most common value of 999999999. Address has a most common value of 123 MAIN ST. Homephone has a most common value of 9999999999. DOB has a most common value 0f 19070626. After consulting with an expert, we found out that these values are the default value for the null data. To clean these data, we change all of these values into their corresponding record numbers.

III. Variable Creation

Fraudsters have different ways to commit identity fraud. There are three modes of identity fraud, identity theft, identity manipulation and synthetic identity. Identity theft, which is this project focused on, is stealing other's identity. Identity manipulation is changing a portion of information in a particular identity. Synthetic identity is a fake identity created by combining synthetic credentials not associated with real person. In the table below, the maximum counts variables later found out to be invalid since they look into the future. Therefore, it should be removed. However, due to weak computer power, I had no choice but continued including these wrong variables for this project.

Description of the variables	# Variables created
Applicant's age when applying	1
Day of week target encoding with smoothing	1
Days since a record with the same attribute was seen	23
Velocity: The number of records with the same entity over 0, 1, 3, 7, 14, 30 days	138
Velocity Change Variables: This set of variables is the number of each attribute in past 0 or 1 day out of the total number in 3, 7, 14, 30 days	184
Unique Values Variables: The number of unique values for one entity to another entity in the past 0, 1, 3, 7, 14, 30, 60 days	3542
Maximum Counts Variables (Should be removed) The maximum value of each entity in the past 1, 3, 7, 30 days	92
Age Indicator Variables: The maximum value, mean value and minimum value of age when apply in each entity	69

IV. Feature Selection

Because of the curse of dimensionality, we need to do feature selection. Also, computer power needs to be stronger with high dimensions. That's why we do feature selection that only necessary variables are left. From the variables we created earlier, we first get rid of the replicated variables. Then, we do a filter, a univariate and fast method to get the variables with high filter scores. Lastly, we do a wrapper, a multivariate selection that could remove correlations. We get 18 variables left.

variable	filter score
max_count_by_address_30	0.359215465
max_count_by_ssn_dob_7	0.228400837
max_count_by_homephone_3	0.224757436
zip5_count_1	0.221239028
max_count_by_fulladdress_30	0.359913969
max_count_by_name_30	0.222190696
max_count_by_homephone_7	0.232235291
max_count_by_ssn_dob_30	0.24083569
fulladdress_count_0_by_30	0.290722131
ssn_firstname_day_since	0.226427511
max_count_by_homephone_30	0.21593074
fulladdress_day_since	0.333268536
address_unique_count_for_ssn_zip5_60	0.289723617
max_count_by_fulladdress_homephone_30	0.249723749
address_count_30	0.332648157
max_count_by_address_7	0.343335432
address_day_since	0.334139944
max_count_by_fulladdress_3	0.329537708

V. Preliminary Model Explanation

Model	Parameters									Average FDR at 3%			
Logistic Regression	Iteration	n_variables	penalty	C	solver	l1_ratio	max_iter			Train	Test	OOT	
	1	10	l2	1	lbfgs	0.5	20			48.85%	48.74%	47.37%	
	2	10	l1	0.5	liblinear	0.5	20			48.97%	48.45%	47.38%	
	3	10	l1	0.3	saga	0.8	10			48.97%	48.49%	47.44%	
	4	10	elasticnet	0.7	saga	1	20			48.87%	48.02%	47.18%	overfitting
	5	15	l2	0.5	liblinear	0.1	30			48.75%	49.06%	47.44%	
Decision Tree	Iteration	n_variables	criterion	max_depth	min_samples_split	min_samples_leaf	max_features			Train	Test	OOT	
	1	10	gini	5	50	30	3			47.45%	47.99%	45.13%	
	2	10	gini	10	40	25	8			52.89%	52.26%	50.42%	
	3	10	gini	15	30	20	5			53.45%	52.08%	50.14%	overfitting
	4	10	gini	20	20	10	8			53.75%	52.09%	50.12%	overfitting
	5	15	gini	30	10	5	10			54.19%	51.72%	49.99%	overfitting
Random Forest	Iteration	n_variables	criterion	n_estimators	max_depth	min_samples_split	min_samples_leaf	max_features		Train	Test	OOT	
	1	10	entropy	5	5	50	30	3		51.80%	51.52%	49.39%	
	2	10	entropy	20	10	40	20	5		53.27%	52.07%	50.42%	
	3	10	entropy	50	15	30	15	8		53.87%	51.99%	50.21%	overfitting
	4	10	gini	80	20	20	10	8		54.02%	52.07%	50.08%	overfitting
	5	15	gini	100	30	10	5	10		54.04%	52.11%	50.08%	overfitting
LightGBM	Iteration	n_variables	boosting_type	max_depth	num_leaves	n_estimators	colsample_bytree	subsample	learning_rate	Train	Test	OOT	
	1	10	gbdt	3	100	20	0.2	0.2	0.1	50.14%	49.94%	47.66%	
	2	10	gbdt	10	200	50	0.5	0.5	0.05	53.04%	52.22%	50.72%	
	3	10	GOSS	10	100	50	0.5	0.5	0.01	52.70%	52.68%	50.47%	
	4	10	GOSS	20	300	70	0.8	0.5	0.01	53.30%	52.15%	50.49%	overfitting
	5	15	GOSS	40	500	100	0.8	0.8	0.01	53.02%	52.80%	50.65%	
Neural Network	Iteration	n_variables	hidden_layer_sizes	activation	alpha	learning_rate	solver	learning_rate_init		Train	Test	OOT	
	1	10	5	logistic	0.1	constant	adam	0.01		50.12%	49.54%	47.87%	
	2	10	10	logistic	0.05	constant	lbfgs	0.005		52.55%	52.27%	50.24%	
	3	10	(10,10,10)	relu	0.01	adaptive	lbfgs	0.001		52.71%	52.67%	50.49%	
	4	10	(10,10,10)	relu	0.001	constant	adam	0.0005		52.62%	52.92%	50.58%	
	5	15	(20,20,20)	relu	0.0001	constant	adam	0.0001		52.81%	52.31%	50.48%	overfitting
GBC	Iteration	n_variables	learning_rate	subsample	max_depth	n_estimators	min_samples_split	min_samples_leaf		Train	Test	OOT	
	1	10	0.1	1	2	5	2	1		49.64%	48.71%	47.55%	
	2	10	0.3	0.2	5	15	8	3		34.73%	33.99%	32.19%	
	3	15	0.05	0.8	20	20	3	1		54.13%	51.61%	49.66%	overfitting
	4	15	0.3	0.5	10	15	3	1		42.10%	40.94%	39.15%	
	5	15	0.1	0.8	10	10	5	2		53.19%	52.52%	50.19%	
Catboost	Iteration	n_variables	verbose	max_depth	iterations	l2_leaf_reg	learning_rate	bootstrap_type		Train	Test	OOT	
	1	10	0	2	5	1	0.01	Bayesian		49.60%	49.54%	47.74%	
	2	10	2	5	20	3	0.1	Bayesian		52.04%	51.50%	49.87%	
	3	10	5	8	20	5	0.5	Bayesian		52.58%	52.33%	50.15%	
	4	10	8	10	30	10	0.05	Bernoulli		51.62%	50.99%	49.25%	
	5	15	10	5	20	15	0.1	Bernoulli		50.78%	50.47%	48.42%	
XGBoost	Iteration	n_variables	booster	max_depth	n_estimators	min_child_weight	colsample_bytree	subsample	eta	Train	Test	OOT	
	1	10	gbtree	2	5	1	1	1	0.2	49.53%	49.15%	47.61%	
	2	10	gbtree	5	20	10	0.8	0.8	0.2	52.74%	52.50%	50.33%	
	3	10	gbtree	20	50	20	0.5	0.7	0.4	52.84%	52.94%	50.61%	
	4	10	dart	10	30	10	0.7	0.7	0.3	52.98%	52.47%	50.67%	overfitting
	5	15	dart	20	80	20	0.8	0.8	0.3	53.33%	52.18%	50.50%	overfitting

VI. Result Summary

I chose neural network as my final model for this project. The FDR for training, testing and out of trend data are listed as following table. The model is not overfitting, and it catches 50.7% of fraud by only rejecting 3% of the applications.

trn	tst	Oot
0.528138	0.524907	0.507125

Train:

bin	#recs	#g	#b	%g	%b	tot	cg	cb	%cg	FDR	KS	FPR
0	0	0	0	0	0	0	0	0	0	0	0	0
1	5835	1566	4269	26.83805	73.16195	5835	1566	4269	0.272367	50.25901	49.98664	0.366831
2	5834	5707	127	97.82311	2.176894	11669	7273	4396	1.264958	51.75418	50.48922	1.654459
3	5835	5745	90	98.45758	1.542416	17504	13018	4486	2.264158	52.81375	50.54959	2.901917
4	5834	5795	39	99.3315	0.668495	23338	18813	4525	3.272054	53.2729	50.00084	4.157569
5	5835	5795	40	99.31448	0.685518	29173	24608	4565	4.27995	53.74382	49.46387	5.390581
6	5834	5789	45	99.22866	0.77134	35007	30397	4610	5.286803	54.2736	48.9868	6.593709
7	5835	5788	47	99.19452	0.805484	40842	36185	4657	6.293481	54.82694	48.53346	7.770024
8	5834	5793	41	99.29722	0.702777	46676	41978	4698	7.30103	55.30963	48.0086	8.935292
9	5835	5788	47	99.19452	0.805484	52511	47766	4745	8.307708	55.86296	47.55525	10.0666
10	5834	5794	40	99.31436	0.685636	58345	53560	4785	9.315431	56.33388	47.01845	11.19331
11	5835	5787	48	99.17738	0.822622	64180	59347	4833	10.32194	56.89899	46.57705	12.27954
12	5834	5798	36	99.38293	0.617072	70014	65145	4869	11.33035	57.32282	45.99246	13.37954
13	5835	5792	43	99.26307	0.736932	75849	70937	4912	12.33773	57.82906	45.49133	14.44157
14	5835	5793	42	99.28021	0.719794	81684	76730	4954	13.34528	58.32352	44.97825	15.48849
15	5834	5793	41	99.29722	0.702777	87518	82523	4995	14.35282	58.80622	44.45339	16.52112
16	5835	5791	44	99.24593	0.75407	93353	88314	5039	15.36003	59.32423	43.9642	17.5261
17	5834	5795	39	99.3315	0.668495	99187	94109	5078	16.36792	59.78338	43.41546	18.53269
18	5835	5794	41	99.29734	0.702656	105022	99903	5119	17.37564	60.26607	42.89043	19.51612
19	5834	5794	40	99.31436	0.685636	110856	105697	5159	18.38337	60.73699	42.35363	20.48789

Test:

bin	#recs	#g	#b	%g	%b	tot	cg	cb	%cg	FDR	KS	FPR
0	0	0	0	0	0	0	0	0	0	0	0	0
1	2501	747	1754	29.86805	70.13195	2501	747	1754	0.302993	49.92884	49.62584	0.425884
2	2500	2451	49	98.04	1.96	5001	3198	1803	1.297153	51.32365	50.0265	1.77371
3	2501	2460	41	98.36066	1.639344	7502	5658	1844	2.294962	52.49075	50.19579	3.06833
4	2500	2482	18	99.28	0.72	10002	8140	1862	3.301695	53.00313	49.70144	4.371643
5	2501	2482	19	99.2403	0.759696	12503	10622	1881	4.308429	53.54398	49.23555	5.646996
6	2500	2481	19	99.24	0.76	15003	13103	1900	5.314756	54.08483	48.77007	6.896316
7	2501	2489	12	99.52019	0.479808	17504	15592	1912	6.324329	54.42642	48.10209	8.154812
8	2500	2477	23	99.08	0.92	20004	18069	1935	7.329034	55.08113	47.75209	9.337984
9	2501	2485	16	99.36026	0.639744	22505	20554	1951	8.336984	55.53658	47.19959	10.53511
10	2500	2483	17	99.32	0.68	25005	23037	1968	9.344123	56.0205	46.67637	11.70579
11	2501	2485	16	99.36026	0.639744	27506	25522	1984	10.35207	56.47595	46.12387	12.86391
12	2500	2483	17	99.32	0.68	30006	28005	2001	11.35921	56.95986	45.60065	13.9955
13	2501	2481	20	99.20032	0.79968	32507	30486	2021	12.36554	57.52918	45.16364	15.08461
14	2500	2483	17	99.32	0.68	35007	32969	2038	13.37268	58.01309	44.64042	16.17713
15	2501	2474	27	98.92043	1.079568	37508	35443	2065	14.37617	58.78167	44.4055	17.16368
16	2500	2483	17	99.32	0.68	40008	37926	2082	15.3833	59.26558	43.88228	18.21614
17	2501	2485	16	99.36026	0.639744	42509	40411	2098	16.39125	59.72104	43.32978	19.26168
18	2501	2479	22	99.12035	0.879648	45010	42890	2120	17.39677	60.34728	42.95051	20.23113
19	2500	2488	12	99.52	0.48	47510	45378	2132	18.40594	60.68887	42.28293	21.28424

OOT:

bin	#recs	#g	#b	%g	%b	tot	cg	cb	%cg	FDR	KS	FPR
0	0	0	0	0	0	0	0	0	0	0	0	0
1	1665	514	1151	30.87087	69.12913	1665	514	1151	0.31321	48.23973	47.92652	0.446568
2	1665	1629	36	97.83784	2.162162	3330	2143	1187	1.305855	49.74853	48.44268	1.805392
3	1665	1642	23	98.61862	1.381381	4995	3785	1210	2.306422	50.71249	48.40607	3.128099
4	1665	1643	22	98.67868	1.321321	6660	5428	1232	3.307598	51.63453	48.32694	4.405844
5	1665	1653	12	99.27928	0.720721	8325	7081	1244	4.314868	52.13747	47.8226	5.692122
6	1665	1657	8	99.51952	0.48048	9990	8738	1252	5.324575	52.47276	47.14818	6.979233
7	1665	1658	7	99.57958	0.42042	11655	10396	1259	6.334891	52.76614	46.43124	8.257347
8	1664	1652	12	99.27885	0.721154	13319	12048	1271	7.341552	53.26907	45.92752	9.47915
9	1665	1649	16	99.03904	0.960961	14984	13697	1287	8.346384	53.93965	45.59326	10.64258
10	1665	1656	9	99.45946	0.540541	16649	15353	1296	9.355481	54.31685	44.96137	11.84645
11	1665	1649	16	99.03904	0.960961	18314	17002	1312	10.36031	54.98743	44.62711	12.95884
12	1665	1655	10	99.3994	0.600601	19979	18657	1322	11.3688	55.40654	44.03774	14.11271
13	1665	1651	14	99.15916	0.840841	21644	20308	1336	12.37485	55.99329	43.61844	15.2006
14	1665	1647	18	98.91892	1.081081	23309	21955	1354	13.37847	56.74769	43.36923	16.21492
15	1665	1655	10	99.3994	0.600601	24974	23610	1364	14.38695	57.16681	42.77985	17.30938
16	1665	1653	12	99.27928	0.720721	26639	25263	1376	15.39422	57.66974	42.27552	18.35974
17	1665	1655	10	99.3994	0.600601	28304	26918	1386	16.40271	58.08885	41.68614	19.42136
18	1665	1652	13	99.21922	0.780781	29969	28570	1399	17.40937	58.6337	41.22432	20.42173
19	1665	1648	17	98.97898	1.021021	31634	30218	1416	18.4136	59.34619	40.93259	21.3404