Project 1 Report

1. Data Description

This is the synthetic applications data showing identity information for each applicant during the Year 2017. It has 10 fields and 1,000,000 records.

The summary tables for numerical and categorical columns are displayed in Table 1 and Table 2 respectively.

Table 1 Numerical Table

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

Table 2 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

2. Data Cleaning

We cleaned frivolous field values, which are inadvertently identical values probably caused by the automatic filling in those missing fields. As many of our variables keep track of the number of times a specific field value has been seen in the past, with a high frequency indicating potential identity fraud, those improperly common values will cause our linkage variables to make wrong counts and thus artificially increase the risk level of an application. Therefore, we cleaned the frivolous field value by replacing it with the unique record number which will not link to any previous value for that field.

3. Variable Creation

Identity fraud typically takes on three forms, that are, identity theft, identity manipulation and synthetic identity. Identity theft is when the fraudster steals the victims' legitimate identities. Identity manipulation is when the fraudster slightly modifies his identity or someone else's. Synthetic identity is when the fraudster makes up all the identities. We are looking for identity theft in this project, and designed variables to discover unusual things about the events or unusual connections between the events. The total variables created are listed in Table 3.

Table 3 Variable Creation

Description of variables	# Variables created
Age When Apply: the difference in the year between application date and date of birth	1
Date of Week Target Encoded: average fraud percentage of that day	1
Days Since: # days since an application with that attribute has been seen	23
Velocity: # records with the same attribute over the past 0,1,3,7,14,30 days	138
Relative Velocity: ratio of velocity over the past 0, 1 day to average velocity over the past 3,7,14,30 days	184
Number of Unique: # unique attributes for another attribute over the past 0,1,3,7,14,30 days	3542
Maximum Indicator: maximum velocity of applications with the same attribute over the past 1,3,7,30 days	92
Age Indicator: mean, max and min of age when apply of applications with the same attribute	69

4. Feature Selection

We removed Maximum Indicator variables before feature engineering because we later found out that they were generated over the entire dataset, which leads to a target leak.

Then, we began feature selection by firstly quickly filtering variables. We calculated the KS score for each explanatory variable, with higher score meaning the higher importance in predicting the fraud probability, and chose the 200 variables with the highest KS scores. To further narrow down the list, we utilized the wrapper method. We employed Light GBM to complete the forward selection process and were left with 20 variables finally. The 20 ultimate variables with their corresponding KS scores are presented in Table 4.

Table 4 Feature Selection Final Results

Wrapper Order	Variable	Filter Score
1	fulladdress_day_since	0.33326854
2	name_dob_count_30	0.22749727
3	address_unique_count_for_name_homephone_60	0.29243796
4	fulladdress_unique_count_for_dob_homephone_3	0.26435919
5	address_unique_count_for_homephone_name_dob_30	0.28398921
6	address_unique_count_for_ssn_name_dob_14	0.27689388

7	address_day_since	0.33413994
8	address_count_14	0.32243628
9	address_count_7	0.30173528
10	address_count_0_by_30	0.29192219
11	address_unique_count_for_homephone_name_dob_60	0.29140979
12	fulladdress_count_0_by_30	0.29072213
13	address_unique_count_for_ssn_zip5_60	0.28972362
14	address_unique_count_for_ssn_name_60	0.28967921
15	address_unique_count_for_ssn_firstname_60	0.28812727
16	address_unique_count_for_ssn_name_dob_60	0.28764489
17	address_unique_count_for_dob_homephone_60	0.28755587
18	address_unique_count_for_ssn_homephone_60	0.2891664
19	address_unique_count_for_ssn_lastname_60	0.2874436
20	address_unique_count_for_ssn_60	0.28591335

5. Preliminary Models Exploration

To establish a benchmark, we created a baseline classification model using logistic regression. Afterwards, we explored various nonlinear models, including decision tree, random forest, XGBoost, LightGBM, and Neural Network, to see how the model performance (Fraud Detection Rate at 3%) would vary with different hyperparameters. We also experimented with different numbers of variables, such as 10, 18, and 20. Table 5 illustrates the results of all the trials.

Table 5 Preliminary Model Exploration Results

Model		Parameter						Average FDR at 3%			
Logistic Regression	Number of variables	max_iter						Train	Test	OOT	
1	10	20						0.489	0.486	0.473	
2	18	20						0.478	0.477	0.463	
3	20	20						0.484	0.478	0.468	
Decision Tree	Number of variables	max_depth	min_sample_split	min_sample_leaf	max_features			Train	Test	OOT	
1	10	5	50	30	5			0.503	0.508	0.488	underfitting
2	10	13	45	25	7			0.530	0.526	0.504	
3	10	15	40	20	10			0.535	0.520	0.502	overfitting
4	10	10	45	25	7			0.528	0.528	0.506	best for # variable =10
5	18	15	30	16	6			0.534	0.521	0.504	
6	20	20	40	20	7			0.537	0.525	0.504	
Random Forest	Number of variables	max_depth	min_sample_split	min_sample_leaf	max_features	n_estimators		Train	Test	OOT	
1	10	10	45	25	5	5		0.497	0.498	0.478	underfitting
2	10	10	50	30	7	10		0.533	0.518	0.505	
3	10	10	60	40	7	20		0.531	0.522	0.505	
4	10	10	65	45	8	50		0.529	0.527	0.506	
5	10	10	65	45	8	100		0.529	0.526	0.506	overfitting
6	10	10	45	25	7	70		0.530	0.529	0.506	best for # variable =10
7	18	15	30	16	6	40		0.537	0.522	0.505	
8	20	25	35	15	8	55		0.541	0.522	0.503	
XGBoost	Number of variables	max_depth			n_estimators			Train	Test	OOT	
2	10	2			50			0.522	0.526	0.501	underfitting
3	10	4			75			0.530	0.527	0.506	
4	10	4			100			0.532	0.523	0.505	overfitting
5	10	4			50			0.529	0.526	0.507	best for # variable =10
6	18	5			700			0.545	0.516	0.502	
7	20	5			800			0.542	0.524	0.502	
LightGBM	Number of variables	num_leaves			n_estimators			Train	Test	OOT	
1	10	4			50			0.520	0.523	0.499	underfitting
2	10	8			100			0.527	0.528	0.507	
3	10	9			500			0.528	0.521	0.505	overfitting
4	10	8			200			0.528	0.528	0.507	best for # variable =10
5	18	6			50			0.532	0.521	0.507	
6	20	6			500			0.530	0.527	0.508	
Neural Network	Number of variables	hidden_layer_size	activation	alpha	learning_rate	solver	learning_rate_init	Train	Test	OOT	
1	10	(5,)	relu	0.0001	constant	adam	0.001	0.512	0.512	0.494	underfitting
2	10	(10,10)	relu	0.0001	constant	adam	0.001	0.529	0.522	0.504	overfitting
3	10	(20,20,20)	relu	0.0001	constant	adam	0.001	0.525	0.524	0.505	best for # variable =10
4	10	(5,)	logistic	0.0001	constant	adam	0.001	0.520	0.523	0.500	
5	10	(5,)	relu	0.0001	adaptive	adam	0.001	0.518	0.512	0.496	
6	10	(5,)	relu	0.0001	constant	lbfgs	0.001	0.522	0.522	0.500	
7	10	(10,10)	logistic	0.01	adaptive	lbfgs	0.001	0.520	0.516	0.497	
8	10	(20,20,20)	logistic	0.0001	adaptive	lbfgs	0.01	0.520	0.519	0.497	
9	18	(20,20,20)	logistic	0.0001	constant	adam	0.0001	0.523	0.516	0.497	
10	20	(5,)	relu	0.0001	adaptive	lbfgs	0.0001	0.523	0.528	0.503	

6. Summary of Final Model Results

We selected XGBoost model as our final classification model because it has one of the best performances and is easy to explain. Based on the preliminary model results, we set max_depth = 4 and n_estimators = 50. We chose 10 variables for the ultimate model, as seen in Table 6.

Table 6 List of Variables for Final Model

Number	Variable
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

For the training, testing and OOT datasets, the final XGBoost model results in one experiment are illustrated in Table 7, 8 and 9 respectively, and the corresponding average FDRs at 3% after running the model for 30 times are 52.83%, 52.79% and 50.61%, which means the model generally catches 52.83%, 52.79% and 50.61% of all the frauds in 3% of the population respectively, suggesting the effectiveness of this classification model. Additionally, this model is deemed as successful since training, testing and OOT all have good results, while training is slightly better than testing. It indeed balances the complexity and generalization and tries to reach the best performance.

Table 7 Final XGBoost Model Results for Training Datasets

Training	# Records	# Goods	# Bads	Fraud Rate
	583454	575145	8309	0.014241054

	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Cumulative Goods	% Cumulative Bads (FDR)	KS	FPR
1	5835	1660	4175	28.45%	71.55%	5835	1660	4175	0.29%	50.25%	49.96	0.40
2	5834	5687	147	97.48%	2.52%	11669	7347	4322	1.28%	52.02%	50.74	1.70
3	5835	5775	60	98.97%	1.03%	17504	13122	4382	2.28%	52.74%	50.46	2.99
4	5834	5804	30	99.49%	0.51%	23338	18926	4412	3.29%	53.10%	49.81	4.29
5	5835	5788	47	99.19%	0.81%	29173	24714	4459	4.30%	53.66%	49.37	5.54
6	5834	5799	35	99.40%	0.60%	35007	30513	4494	5.31%	54.09%	48.78	6.79
7	5835	5783	52	99.11%	0.89%	40842	36296	4546	6.31%	54.71%	48.40	7.98
8	5834	5788	46	99.21%	0.79%	46676	42084	4592	7.32%	55.27%	47.95	9.16
9	5835	5802	33	99.43%	0.57%	52511	47886	4625	8.33%	55.66%	47.34	10.35
10	5834	5791	43	99.26%	0.74%	58345	53677	4668	9.33%	56.18%	46.85	11.50
11	5835	5785	50	99.14%	0.86%	64180	59462	4718	10.34%	56.78%	46.44	12.60
12	5834	5782	52	99.11%	0.89%	70014	65244	4770	11.34%	57.41%	46.06	13.68
13	5835	5801	34	99.42%	0.58%	75849	71045	4804	12.35%	57.82%	45.46	14.79
14	5835	5790	45	99.23%	0.77%	81684	76835	4849	13.36%	58.36%	45.00	15.85
15	5834	5795	39	99.33%	0.67%	87518	82630	4888	14.37%	58.83%	44.46	16.90
16	5835	5789	46	99.21%	0.79%	93353	88419	4934	15.37%	59.38%	44.01	17.92
17	5834	5793	41	99.30%	0.70%	99187	94212	4975	16.38%	59.87%	43.49	18.94
18	5835	5796	39	99.33%	0.67%	105022	100008	5014	17.39%	60.34%	42.96	19.95
19	5834	5792	42	99.28%	0.72%	110856	105800	5056	18.40%	60.85%	42.45	20.93
20	5835	5800	35	99.40%	0.60%	116691	111600	5091	19.40%	61.27%	41.87	21.92

Table 8 Final XGBoost Model Results for Testing Datasets

Testing	# Records		# Goods		# Bads		Fraud Rate					
	250053		246355		3698		0.014788865					
	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Cumulative Goods	% Cumulative Bads (FDR)	KS	FPR
1	2501	646	1855	25.83%	74.17%	2501	646	1855	0.26%	50.16%	49.90	0.35
2	2500	2434	66	97.36%	2.64%	5001	3080	1921	1.25%	51.95%	50.70	1.60
3	2501	2468	33	98.68%	1.32%	7502	5548	1954	2.25%	52.84%	50.59	2.84
4	2500	2474	26	98.96%	1.04%	10002	8022	1980	3.26%	53.54%	50.29	4.05
5	2501	2484	17	99.32%	0.68%	12503	10506	1997	4.26%	54.00%	49.74	5.26
6	2500	2484	16	99.36%	0.64%	15003	12990	2013	5.27%	54.43%	49.16	6.45
7	2501	2489	12	99.52%	0.48%	17504	15479	2025	6.28%	54.76%	48.48	7.64
8	2500	2487	13	99.48%	0.52%	20004	17966	2038	7.29%	55.11%	47.82	8.82
9	2501	2483	18	99.28%	0.72%	22505	20449	2056	8.30%	55.60%	47.30	9.95
10	2500	2485	15	99.40%	0.60%	25005	22934	2071	9.31%	56.00%	46.69	11.07
11	2501	2479	22	99.12%	0.88%	27506	25413	2093	10.32%	56.60%	46.28	12.14
12	2500	2483	17	99.32%	0.68%	30006	27896	2110	11.32%	57.06%	45.73	13.22
13	2501	2479	22	99.12%	0.88%	32507	30375	2132	12.33%	57.65%	45.32	14.25
14	2500	2476	24	99.04%	0.96%	35007	32851	2156	13.33%	58.30%	44.97	15.24
15	2501	2484	17	99.32%	0.68%	37508	35335	2173	14.34%	58.76%	44.42	16.26
16	2500	2488	12	99.52%	0.48%	40008	37823	2185	15.35%	59.09%	43.73	17.31
17	2501	2482	19	99.24%	0.76%	42509	40305	2204	16.36%	59.60%	43.24	18.29
18	2501	2479	22	99.12%	0.88%	45010	42784	2226	17.37%	60.19%	42.83	19.22
19	2500	2480	20	99.20%	0.80%	47510	45264	2246	18.37%	60.74%	42.36	20.15
20	2501	2480	21	99.16%	0.84%	50011	47744	2267	19.38%	61.30%	41.92	21.06

Table 9 Final XGBoost Model Results for OOT Datasets

OOT	# Records		# Goods		# Bads		Fraud Rate					
	166493		164107		2386		0.014330933					
	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative 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	1643	22	98.68%	1.32%	3330	2153	1177	1.31%	49.33%	48.02	1.83
3	1665	1637	28	98.32%	1.68%	4995	3790	1205	2.31%	50.50%	48.19	3.15
4	1665	1648	17	98.98%	1.02%	6660	5438	1222	3.31%	51.22%	47.90	4.45
5	1665	1647	18	98.92%	1.08%	8325	7085	1240	4.32%	51.97%	47.65	5.71
6	1665	1654	11	99.34%	0.66%	9990	8739	1251	5.33%	52.43%	47.11	6.99
7	1665	1655	10	99.40%	0.60%	11655	10394	1261	6.33%	52.85%	46.52	8.24
8	1664	1656	8	99.52%	0.48%	13319	12050	1269	7.34%	53.19%	45.84	9.50
9	1665	1658	7	99.58%	0.42%	14984	13708	1276	8.35%	53.48%	45.13	10.74
10	1665	1651	14	99.16%	0.84%	16649	15359	1290	9.36%	54.07%	44.71	11.91
11	1665	1652	13	99.22%	0.78%	18314	17011	1303	10.37%	54.61%	44.24	13.06
12	1665	1655	10	99.40%	0.60%	19979	18666	1313	11.37%	55.03%	43.66	14.22
13	1665	1657	8	99.52%	0.48%	21644	20323	1321	12.38%	55.36%	42.98	15.38
14	1665	1653	12	99.28%	0.72%	23309	21976	1333	13.39%	55.87%	42.48	16.49
15	1665	1653	12	99.28%	0.72%	24974	23629	1345	14.40%	56.37%	41.97	17.57
16	1665	1649	16	99.04%	0.96%	26639	25278	1361	15.40%	57.04%	41.64	18.57
17	1665	1657	8	99.52%	0.48%	28304	26935	1369	16.41%	57.38%	40.96	19.67
18	1665	1652	13	99.22%	0.78%	29969	28587	1382	17.42%	57.92%	40.50	20.69
19	1665	1654	11	99.34%	0.66%	31634	30241	1393	18.43%	58.38%	39.95	21.71
20	1665	1650	15	99.10%	0.90%	33299	31891	1408	19.43%	59.01%	39.58	22.65