Report

Identity Theft Fraud Analysis

Name: - Aditya Chavan

Email: - chavana@usc.edu

1) Description of Data

The data is about **product applications**. It comprises **applicant information** for each application. There are **10 attributes** and **one million entries** from the year **2017**.

1) Numerical Attributes

Field Name	% Populated	Min	Max	Mean	Median	Stddev	% Zero
Date	100	01-01-2017	12-31-2017	-	-	-	0.00
DOB	100						0.00

2) Categorical Attributes

Field Name	Description	%Populated	# Unique Values	Most Common Value
Record	Unique record Number	100.00	1,000,000	N/A
SSN	SSN of the applicant	100.00	835,819	999999999
First Name	First Name of the applicant	100.00	78,136	EAMSTRMT
Last Name	Last Name of the applicant	100.00	177,001	ERJSAXA
Address	Address of the applicant	100.00	828,774	123, MAIN ST
Zip	Zip code of the address	100.00	26,370	68138
Homephone	Contact number of the applicant	100.00	28,244	9999999999
Fraud Label	Is the application a fraud	100.00	2	0

2) Data Cleaning

The data does not contain any missing values it is a completely populated data set. There need not be any steps taken here.

There are some frivolous values in the data. These values will affect the model and mess with the prediction. These values were replaced with a negative of the corresponding record number.

Data Field	Frivolous Value	Number of Records
ssn	999999999	16,935
address	123 MAIN ST	1,079
dob	19070626	126,568
homephone	9999999999	78,512

In the zip code some zip codes are not 5 digit long. So leading zeros are added to the zip code. For example, “1204” is replaced with “01204”.

3) Variable Creation

The variables are not enough to predict fraudulent activity. The occurrence of values with respect to time needs to be considered. In simple words how often can we see a value in the data.

A table below shows how original variables were converted to more variables.

Entities were created first then the occurrence of entities was concerted to variables (Days since variables, Velocity variables, Relative velocity variables). Entities were grouped together (Counting entities, Maximum Indicator). Then some variables revolving around age of the applicant were created. Fraud label is the target variable which is binary in nature.

Description of Variables	Number of created variables
Original Variables (Date, SSN, Firstname, Lastname, Address, Zip, DOB, Homephone)	6
Age of the Applicant	1
Entities (Name, Full address, Name_DOB, Name_Fulladdress, Name_Homephone, Fulladdress_DOB, Fulladdress_Homephone, DOB_Homephone, Homephone_Name_DOB)	9
Days Since Variables (Days Since the same entity has been seen before)	23
Velocity Variables (Number of records with the same entity in the last specific number of days)	138
Relative Velocity Variables (Recent number of records with the same entity)	184
Counting Entities (Unique number of entities grouped by entities)	1753
Maximum Indicator (Maximum count of an entity grouped by other entities)	92
Age Indicator (Min, Median, Mode of Age when applicants apply)	69
Target Variable (The variable which the model predicts or compares to while training)	1

4) Feature Selection

If we use all the created variables the model will be inefficient. Also, so many variables are not required to predict fraud. We used the sequential feature selector for the task. Sequential feature selection (SFS) algorithms belong to a group of greedy search algorithms that aim to reduce an original set of features with d dimensions to a subspace of features with k dimensions where k is smaller than d. These algorithms automatically choose the most important subset of features relevant to a given problem. Feature selection has two primary objectives: firstly, to increase computational efficiency and reduce the generalization error of the model by eliminating irrelevant features and noise. Secondly, if an embedded feature selection technique such as LASSO is not possible, a wrapper approach like SFS is beneficial. In summary, SFS methods iteratively include or exclude one feature at a time based on the classifier's performance, until a subset of features with the desired size k is obtained.

I have selected 45 variables the table with their Univariate KS is provided below.

Variable	Univariate KS
max_count_by_address_30	0.359215
max_count_by_ssn_dob_7	0.228401
max_count_by_homephone_3	0.224757
zip5_count_1	0.221239
max_count_by_fulladdress_30	0.359914
max_count_by_name_30	0.222191
max_count_by_homephone_7	0.232235
max_count_by_ssn_dob_30	0.240836
fulladdress_count_0_by_30	0.290722
ssn_firstname_day_since	0.226428
max_count_by_homephone_30	0.215931
fulladdress_day_since	0.333269
address_unique_count_for_ssn_zip5_60	0.289724
max_count_by_fulladdress_homephone_30	0.249724
address_count_30	0.332648
max_count_by_address_7	0.343335
address_day_since	0.33414
max_count_by_fulladdress_3	0.329538
max_count_by_address_3	0.329445
address_count_14	0.322436
fulladdress_count_14	0.321953
max_count_by_address_1	0.315332
max_count_by_fulladdress_1	0.315253
address_count_7	0.301735
fulladdress_count_7	0.301666
address_unique_count_for_name_homephone_60	0.292438
address_count_0_by_30	0.291922
address_unique_count_for_homephone_name_dob_60	0.29141
fulladdress_unique_count_for_ssn_homephone_60	0.289991
address_unique_count_for_ssn_name_60	0.289679
fulladdress_unique_count_for_name_homephone_60	0.289535
address_unique_count_for_ssn_homephone_60	0.289166
fulladdress_unique_count_for_homephone_name_dob_60	0.288483
fulladdress_unique_count_for_dob_homephone_60	0.288443
address_unique_count_for_ssn_firstname_60	0.288127
address_unique_count_for_ssn_name_dob_60	0.287645
address_unique_count_for_dob_homephone_60	0.287556
address_unique_count_for_ssn_lastname_60	0.287444
address_unique_count_for_name_60	0.287411
fulladdress_unique_count_for_ssn_name_60	0.286799
fulladdress_unique_count_for_ssn_lastname_60	0.286776
fulladdress_unique_count_for_ssn_60	0.286764
fulladdress_unique_count_for_ssn_firstname_60	0.286763

address_unique_count_for_ssn_60	0.285913
address_unique_count_for_name_dob_60	0.285912

5) Preliminary Models Exploration

							Mean			Std Dev		
Model	Number of Variables	N_estimators	max_depth	min_samples_sp	min_samples_le	max_features	Train	Test	OOT	Train	Test	OOT
	20	3	2	1000	500	8	0.479	0.478	0.465	0.003	0.005	0.001
	20	5	10	1000	500	10	0.522	0.523	0.5	0.005	0.004	0.002
	20	5	10	100	50	8	0.528	0.526	0.503	0.0006	0.001	0.001
	15	5	10	100	50	8	0.528	0.524	0.504	0.001	0.002	0.001
Random Forest Classifier	15	5	15	500	100	8	0.529	0.526	0.503	0.002	0.006	0.002

							Mean			Std Dev		
Model	Number of Variables	Activation	Solver	Hidden Layer Si	Learning Rate	alpha	Train	Test	OOT	Train	Test	OOT
	15	relu	adam	2	constant	0.0001	0.505	0.503	0.486	0.017	0.013	0.009
	15	logistic	adam	2	constant	0.0001	0.504	0.503	0.486	0.0008	0.002	0.0009
	15	logistic	adam	2	adaptive	0.0001	0.504	0.505	0.486	0.001	0.003	0.0009
	15	relu	adam	4	adaptive	0.0001	0.513	0.509	0.492	0.016	0.014	0.01
Neural Networks	20						0.517	0.52	0.497	0.004	0.007	0.004

							Mean			Std Dev		
Model	Number of Variables	max_depth	n_estimators	learning_rate	min_samples_sp	min_sample_lea	Train	Test	OOT	Train	Test	OOT
	20	2	5	0.1	2	1	0.495	0.497	0.477	0.003	0.007	0
	20	2	5	0.2	4	2	0.491	0.488	0.473	0.001	0.002	0.0006
	20	5	5	0.2	4	2	0.491	0.488	0.461	0.018	0.023	0.025
	20	5	10	0.05	2	1	0.52	0.521	0.499	0.006	0.007	0.004
Gradient Boosted Classifier	20	3	10	0.05	2	1	0.516	0.518	0.495	0.004	0.006	0.005

						Mean			Std Dev		
Model	Number of Variables	max_iter	penalty	solver	l_ratio	Train	Test	OOT	Train	Test	OOT
	13	20	l2	lbfgs	none	0.479	0.481	0.465	0.005	0.008	0.005
	13	50	l1	saga	none	0.483	0.484	0.469	0.004	0.009	0.005
	20	50	l1	saga	none	0.482	0.487	0.469	0.002	0.014	0.005
	20	50	elastic_net	saga	0.8	0.478	0.481	0.465	0.006	0.003	0.004
Logistic Regression	20	100	l1	saga	none	0.484	0.476	0.467	0.007	0.004	0.006

					Mean			Std Dev		
Model	Number of Variables	max_depth	min_samples	min_samples_leaf	Train	Test	OOT	Train	Test	OOT
	20	2	1000	500	0.46	0.459	0.443	0.001	0.005	0
	20	10	1000	500	0.527	0.52	0.503	0.004	0.008	0.006
	20	15	100	50	0.536	0.517	0.503	0.002	0.006	0.007
	20	15	100	50	0.529	0.525	0.502	0.001	0.002	0.002
Decision Tree	20	20	800	350	0.522	0.528	0.467	0.002	0.012	0.003

6) Summary of Results

Training	# Records	# Goods	# Bads	# FraudRate						
	583454	574999	8455	0.01449129						
	Bin Statistics					Cumulative Statistics						
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	1601	4234	27.43787	72.56213	5835	1601	4234	0.278435	50.07688	49.79844	0.378129
2	5834	5628	206	96.46897	3.531025	11669	7229	4440	1.25722	52.51331	51.25609	1.628153
3	5835	5748	87	98.509	1.491003	17504	12977	4527	2.256873	53.54228	51.28541	2.866578
4	5834	5766	68	98.83442	1.165581	23338	18743	4595	3.259658	54.34654	51.08688	4.078999
5	5835	5776	59	98.98886	1.01114	29173	24519	4654	4.264181	55.04435	50.78017	5.268371
6	5834	5788	46	99.21152	0.788481	35007	30307	4700	5.270792	55.58841	50.31762	6.448298
7	5835	5788	47	99.19452	0.805484	40842	36095	4747	6.277402	56.14429	49.86689	7.60375
8	5834	5804	30	99.48577	0.514227	46676	41899	4777	7.286795	56.49911	49.21232	8.770986
9	5835	5797	38	99.34876	0.651243	52511	47696	4815	8.294971	56.94855	48.65358	9.905711
10	5834	5791	43	99.26294	0.737059	58345	53487	4858	9.302103	57.45713	48.15502	11.01009
11	5835	5789	46	99.21165	0.788346	64180	59276	4904	10.30889	58.00118	47.6923	12.08728
12	5834	5793	41	99.29722	0.702777	70014	65069	4945	11.31637	58.4861	47.16974	13.15854
13	5835	5791	44	99.24593	0.75407	75849	70860	4989	12.3235	59.00651	46.68301	14.20325
14	5835	5783	52	99.10883	0.891174	81684	76643	5041	13.32924	59.62153	46.29229	15.20393
15	5834	5792	42	99.28008	0.719918	87518	82435	5083	14.33655	60.11827	45.78173	16.21778
16	5835	5806	29	99.503	0.497001	93353	88241	5112	15.34629	60.46127	45.11498	17.26154
17	5834	5791	43	99.26294	0.737059	99187	94032	5155	16.35342	60.96984	44.61642	18.24093
18	5835	5795	40	99.31448	0.685518	105022	99827	5195	17.36125	61.44293	44.08169	19.21598
19	5834	5776	58	99.00583	0.994172	110856	105603	5253	18.36577	62.12892	43.76315	20.10337
20	5835	5802	33	99.43445	0.565553	116691	111405	5286	19.37482	62.51922	43.1444	21.07548

Test	# Records	# Goods	# Bads	# FraudRate						
	166493	164101	2386	0.014330933						
	Bin Statistics					Cumulative Statistics						
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	770	1731	30.78768	69.21232	2501	770	1731	0.312372	48.73311	48.42074	0.44483
2	2500	2438	62	97.52	2.48	5001	3208	1793	1.301415	50.4786	49.17719	1.78918
3	2501	2476	25	99.0004	0.9996	7502	5684	1818	2.305873	51.18243	48.87656	3.126513
4	2500	2477	23	99.08	0.92	10002	8161	1841	3.310737	51.82995	48.51922	4.432917
5	2501	2488	13	99.48021	0.519792	12503	10649	1854	4.320064	52.19595	47.87588	5.743797
6	2500	2487	13	99.48	0.52	15003	13136	1867	5.328984	52.56194	47.23295	7.035886
7	2501	2479	22	99.12035	0.879648	17504	15615	1889	6.33466	53.18131	46.84665	8.266278
8	2500	2479	21	99.16	0.84	20004	18094	1910	7.340335	53.77252	46.43219	9.473298
9	2501	2488	13	99.48021	0.519792	22505	20582	1923	8.349662	54.13851	45.78885	10.70307
10	2500	2486	14	99.44	0.56	25005	23068	1937	9.358177	54.53266	45.17448	11.90914
11	2501	2482	19	99.2403	0.759696	27506	25550	1956	10.36507	55.06757	44.7025	13.06237
12	2500	2478	22	99.12	0.88	30006	28028	1978	11.37034	55.68694	44.3166	14.16987
13	2501	2477	24	99.04038	0.959616	32507	30505	2002	12.3752	56.36261	43.98741	15.23726
14	2500	2486	14	99.44	0.56	35007	32991	2016	13.38372	56.75676	43.37304	16.36458
15	2501	2483	18	99.28029	0.719712	37508	35474	2034	14.39102	57.26351	42.8725	17.44051
16	2500	2485	15	99.4	0.6	40008	37959	2049	15.39913	57.68581	42.28668	18.52562
17	2501	2477	24	99.04038	0.959616	42509	40436	2073	16.40399	58.36149	41.9575	19.50603
18	2501	2484	17	99.32027	0.679728	45010	42920	2090	17.41169	58.84009	41.4284	20.53589
19	2500	2473	27	98.92	1.08	47510	45393	2117	18.41494	59.60023	41.18529	21.44214
20	2501	2486	15	99.40024	0.59976	50011	47879	2132	19.42345	60.02252	40.59907	22.45732

OOT	# Records	# Goods	# Bads	# FraudRate						
	250053	246501	3552	0.014204989						
	Bin Statistics					Cumulative Statistics						
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	517	1148	31.05105	68.94895	1665	517	1148	0.315038	48.114	47.79896	0.450348
2	1665	1628	37	97.77778	2.222222	3330	2145	1185	1.307074	49.66471	48.35764	1.810127
3	1665	1642	23	98.61862	1.381381	4995	3787	1208	2.307641	50.62867	48.32103	3.134934
4	1665	1651	14	99.15916	0.840841	6660	5438	1222	3.313692	51.21542	47.90173	4.450082
5	1665	1656	9	99.45946	0.540541	8325	7094	1231	4.322789	51.59262	47.26983	5.762794
6	1665	1656	9	99.45946	0.540541	9990	8750	1240	5.331887	51.96982	46.63794	7.056452
7	1665	1657	8	99.51952	0.48048	11655	10407	1248	6.341594	52.30511	45.96352	8.338942
8	1664	1652	12	99.27885	0.721154	13319	12059	1260	7.348254	52.80805	45.45979	9.570635
9	1665	1652	13	99.21922	0.780781	14984	13711	1273	8.354915	53.35289	44.99798	10.77062
10	1665	1651	14	99.15916	0.840841	16649	15362	1287	9.360966	53.93965	44.57868	11.93629
11	1665	1655	10	99.3994	0.600601	18314	17017	1297	10.36945	54.35876	43.98931	13.12028
12	1665	1658	7	99.57958	0.42042	19979	18675	1304	11.37977	54.65214	43.27237	14.32132
13	1665	1658	7	99.57958	0.42042	21644	20333	1311	12.39009	54.94552	42.55543	15.50953
14	1665	1649	16	99.03904	0.960961	23309	21982	1327	13.39492	55.61609	42.22117	16.56518
15	1665	1652	13	99.21922	0.780781	24974	23634	1340	14.40158	56.16094	41.75936	17.63731
16	1665	1656	9	99.45946	0.540541	26639	25290	1349	15.41068	56.53814	41.12746	18.74722
17	1665	1649	16	99.03904	0.960961	28304	26939	1365	16.41551	57.20872	40.79321	19.73553
18	1665	1651	14	99.15916	0.840841	29969	28590	1379	17.42156	57.79547	40.37391	20.73241
19	1665	1653	12	99.27928	0.720721	31634	30243	1391	18.42883	58.29841	39.86958	21.74191
20	1665	1659	6	99.63964	0.36036	33299	31902	1397	19.43976	58.54987	39.11012	22.83608

The tables above show model performance. FDR stands for Fraud detection rate this is basicaly the percentage of frauds detected higher the better. FPR stands for false positive rate this means the percentage of good applications identified as fraudulent by the algorithm lower the FPR number the better it is.