Transaction Fraud Detection

Executive Summary

The project aims to build a transaction fraud detection model with the dataset Card Transaction Data. For the project, data cleaning and imputation are conducted. Then, feature engineering is conducted to build the feature pool for feature selection. After selecting the best features, multiple machine learning and deep learning models are tried, including Logistic Regression, Boost Tree, Decision Tree, Random Forest, and Neural Network. After tuning the hyperparameters, the performance of the models is evaluated to choose the final model to be deployed. The final model can achieve a training accuracy of 0.759, a testing accuracy rate of 0.726 and an oot accuracy rate of 0.531. In the end, the score cutoff point is recommended to achieve maximum overall savings.

Data Description

This dataset is formed with transaction data. It includes information of numerous transactions in the year of 2010. It has 96,753 rows and 10 fields. 8 out of 10 fields have no null data, and two fields have null values.

Field Summary Table

1. Numerical Table

Field Name	% Population	Min	Max	Mean	Stdev	% Zero
Date	100	2010-01-01	2010-12-31	/	/	0
Amount	100	0.01	3,102,046	427.8857	10,006.14	0

2. Categorical Table

Field Name	% Population	# Unique Values	Most Common Value
Recnum	100	96,753	N/A
Cardnum	100	1,645	5142148452
Merchnum	96.51	13,092	930090121224
Merch description	100	13,126	GSA-FSS-ADV
Merch state	98.76	228	TN
Merch zip	95.19	4,568	38118
Transtype	100	4	P
Fraud	100	2	0

Statistics Table

	Field Name	Field Type	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
0	Recnum	categorical	96753	100.00%	0	96753	1
1	Cardnum	categorical	96753	100.00%	0	1645	5142148452
2	Date	categorical	96753	100.00%	0	365	2/28/10 0:00
3	Merchnum	categorical	93378	96.51%	0	13091	9.3009E+11
4	Merch description	categorical	96753	100.00%	0	13126	GSA-FSS-ADV
5	Merch state	categorical	95558	98.76%	0	227	TN
6	Merch zip	categorical	92097	95.19%	0	4567	38118
7	Transtype	categorical	96753	100.00%	0	4	P
8	Amount	categorical	96753	100.00%	0	34909	3.62
9	Fraud	categorical	96753	100.00%	0	2	0

	Field Name	Field Type	# Records Have Values	% Populated	# Zeros	Min	Max	Most Common
0	Date	datetime	96753	100.00%	0	1/1/10	12/31/10	2/28/10

Data Cleaning and Imputation Logic

After we checked the dataset, we noticed the existence of 3375 missing values are in Merchnum, 1195 missing values are in Merch state, and 4656 missing values are in Merch zip. To prepare the dataset for modeling, the missing values are imputed following the logic below:

For *Merch num*:

Create a dictionary of Merch description and Merch num, and replace the missing values in Merch num with the most corresponding value in the dictionary. For the transaction whose Merch description is ‘Retail Credit Adjustment’ or ‘Retail Debit Adjustment’, fill the missing values with ‘unknown’. Otherwise, fill the missing values with the maximum Merch num + 1.

For *Merch state*:

Create a dictionary of Merch description and Merch state and a dictionary of Merch zip and Merch state, and fill the missing values with the most commonly seen Merch state according to the Merch zip and Merch state. For those with a value that is not a US state, replace them with the value ‘foreign’. Otherwise, fill the missing values with the value ‘unknown’.

For *Merch zip*:

Create a dictionary of Merchnum and Merch zip and a dictionary of Merch description and Merch zip. Fill the missing values with the most commonly seen Merch zip according to the Merchnum and Merch description. For the transaction whose Merch description is ‘Retail Credit

Adjustment' or 'Retail Debit Adjustment', fill the missing values with 'unknown'. Otherwise, fill in the missing value with 'unknown'.

Variable Creation

Description of Variables	# Variables Created
Original fields from the dataset excluding 'Recnum' and 'fraud'	8
Date of week target encoded (average fraud percentage of that day)	1
Dow_risk : The probability of being a fraud considering that day in a week	1
New entities combining/concatenating different original fields	10
Day Since Variables: Number of days since the last transaction of that entity has been seen.	15
Frequency and Amount Variable: Average, Maximum, Median, Total, Actual/Average, Actual/Maximum, Actual/Median, and Actual/Total amount for the same entity our the past {0,1,3,7,14,30,60} days	945
Velocity Change Variables: Number of transactions with one entity in the past {0,1} days divided by the average daily number of transactions with the same entity over the past {7,14,30,60} days.	120
Velocity / Days Since Variables: For each entity, for the past {0,1} days over past {7,14,30,60} days, velocity variables divided by day since variables.	120
Variability Average, Median, and Max amount difference between one record of one entity and the former record of the same entity over the past {0,1,3,7,14,30,60} day	315
Acceleration Number of transactions with one entity in the past {0,1} days divided by the number of transactions with the same entity over the past {7,14,30,60} days over the power of days.	120
Cardnum Gas/Online Frequency Variables: The Number of transactions with 'gas' or '.com' in the Merch description for each Card number in the past {0,1,3,7} days.	8
Risk of Fraud within the Zip The frequency of fraudulent transactions occurred over the previous {0,1,3,7} days.	4

Feature Selection

To find the variables for modeling, a series of exploratory forward selection and backward selection with different models and different combination of values for the parameters *num_filter* and *num_wrapper* was conducted, including:

Backward Selection (LGBM (n_estimators=10, num_leaves=4), num_filter=100, num_wrapper=25)

Forward Selection (Ramdom Foreast (n_estimators=5),num_filter=100, num_wrapper=25)

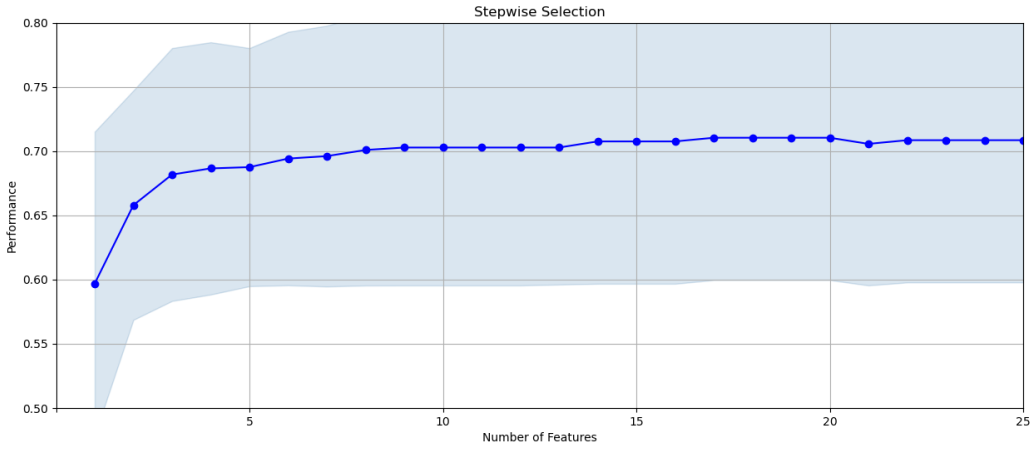
Forward Selection (LGBM (n_estimators=20, num_leaves=4), num_filter=100, num_wrapper=25)

Forward Selection (LGBM (n_estimators=20, num_leaves=4), num_filter=200, num_wrapper=25)

Forward Selection (LGBM (n_estimators=20, num_leaves=4), num_filter=300, num_wrapper=25)

Forward Selection (LGBM (n_estimators=20, num_leaves=4), num_filter=400, num_wrapper=25)

Forward LGBM with filter number of 300 and 25 wrappers was chosen for the final feature selection model, given its performance as presented below.



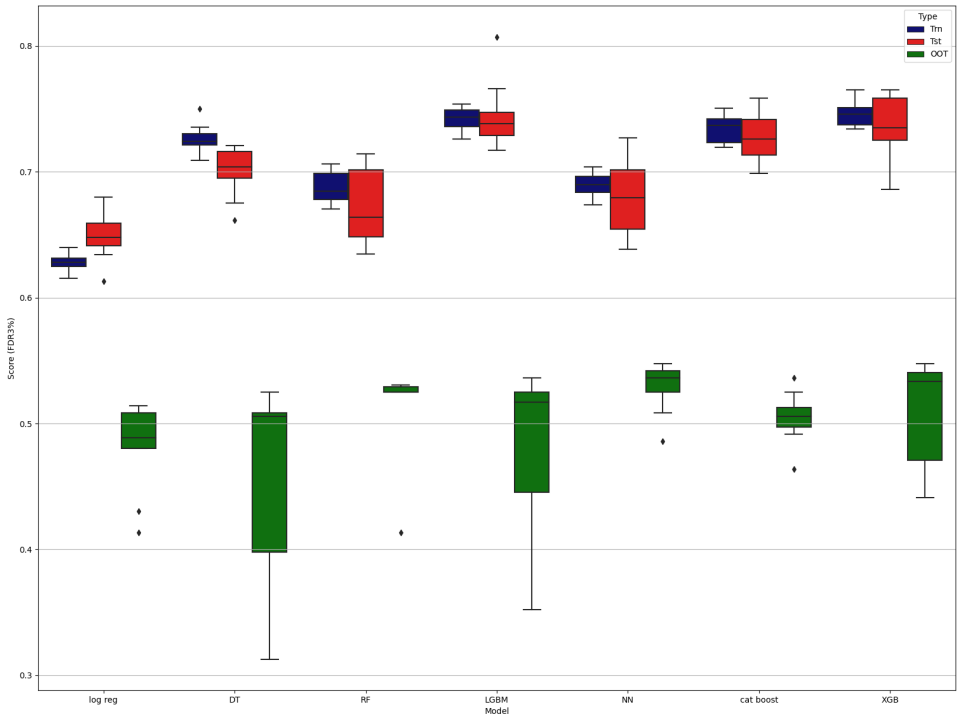
The list of final variables sorted by their univariate KS's score is attached below.

Wrapper Order	Variable	Filter score
1	cardnum_merchnum_merchstate_total_14	0.630058883
2	cardnum_zip3_max_14	0.629514577
3	cardnum_merchnum_merchzip_avg_14	0.51812201
4	cardnum_merchnum_merchdes_avg_7	0.519505431
5	Merch_description_max_0	0.530587568
6	cardnum_merchnum_avg_14	0.518386438
7	cardnum_merchnum_merchdes_total_14	0.612648557
8	Merch_zip_max_0	0.515097616
9	cardnum_merchnum_avg_7	0.524280981
10	cardnum_merchnum_merchstate_avg_7	0.524270154
11	cardnum_merchnum_zip3_avg_14	0.518397265
12	cardnum_merchnum_merchstate_avg_14	0.518364784
13	cardnum_merchnum_merchdes_avg_14	0.515386604
14	cardnum_merchnum_zip3_avg_7	0.524291807
15	cardnum_merchnum_merchzip_avg_7	0.523336698
16	cardnum_merchdes_avg_7	0.516608334
17	cardnum_merchnum_avg_30	0.520958054
18	cardnum_merchnum_merchzip_avg_30	0.521967299
19	cardnum_merchnum_merchstate_avg_30	0.520947227
20	cardnum_merchnum_zip3_avg_30	0.520925573
21	merchnum_merchdes_max_0	0.530685631
22	merchnum_zip3_max_0	0.533066084
23	merchnum_merchstate_max_0	0.533033603
24	Merchnum_max_0	0.533022776
25	merchnum_merchzip_max_0	0.530031577

Preliminary Models Exploration

After 25 best features were selected based on their ranked importance, multiple models including Logistic Regression, Decision Tree, Random Forest, Boosted Tree (XGB, LGBM), and Neural Network were tested. Different combinations of parameters associated with different models were tried, and the models' performances, which were set to be the average FDR@3%, were recorded as below.

Model			Parameter							Average FDR at 3%		
	iteration	mber of Variab	penalty	C	solver		l1_ratio			Train	Test	OOT
	1	15	l2	1	lbfgs		0.4			0.640225	0.626423	0.494972
	2	15	l1	1	liblinear		None			0.637466	0.631766	0.486592
	3	15	l1	0.1	liblinear		None			0.631864	0.643852	0.451955
	4	20	l1	1	saga		0.5			0.635085	0.64176	0.468715
Logistic	5	20	l2	0.8	lbfgs		0.8			0.645212	0.60809	0.488268
Regression	6	20	l2	1	lbfgs		None			0.634775	0.621996	0.493855
	iteration	mber of Variab	Criterion	max_features	min_samples_split	min_samples_leaf	splitter		max_depth	Train	Test	OOT
	1	15	gini	log2	100	20	random		5	0.338347	0.333916	0.184916
	2	15	entropy	None	40	20	best		5	0.723693	0.700732	0.507263
	3	15	entropy	None	40	5	best		10	0.851658	0.765516	0.321788
	4	20	gini	None	40	35	best		5	0.707965	0.678303	0.501676
	5	20	gini	None	50	25	best		5	0.694939	0.67442	0.461453
Decision Tree	6	20	entropy	None	50	30	best		10	0.831464	0.754162	0.363687
	iteration	mber of Variab	n_estimators	criterion	max_depth	min_samples_split	min_samples_leaf		max_features	Train	Test	OOT
	1	15	300	gini	2	50	500		8	0.64239	0.655316	0.495531
	2	15	200	gini	5	50	500		log2	0.644458	0.64237	0.452514
	3	15	200	entropy	3	80	200		log2	0.686794	0.692089	0.527374
	4	20	100	gini	15	20	200		log2	0.758528	0.726189	0.531285
Random	5	20	200	entropy	15	50	400		None	0.665656	0.658555	0.441341
Forest	6	20	100	entropy	15	30	200		log2	0.754732	0.737136	0.530726
	iteration	mber of Variab	num_leaves	max_depth	learning_rate	boosting_type	n_estimators	min_child_samples	_child_weight=0	Train	Test	OOT
	1	15	10	5	0.01	gbdt	50	10	0.001	0.744439	0.733936	0.473743
	2	15	15	15	0.001	gbdt	50	20	0.001	0.734406	0.707374	0.356425
	3	15	10	10	0.01	gbdt	50	20	0.002	0.73976	0.758845	0.52067
	4	20	10	15	0.03	gbdt	50	10	0.001	0.801032	0.768889	0.514525
	5	20	15	20	0.03	gbdt	200	10	0.001	0.880118	0.792411	0.378212
LightGBM	6	20	20	20	0.01	gbdt	80	25	0.001	0.871639	0.776832	0.432122
	iteration	mber of Variab	dden_layer_siz	activation	alpha	learning_rate	learning_rate_init		max_iter	Train	Test	OOT
	1	15	2	relu	0.0001	adaptive	0.001		200	0.57898	0.570946	0.405028
	2	15	5,5	relu	0.0001	adaptive	0.001		200	0.679508	0.670107	0.510056
	3	15	5,5,5	relu	0.001	constant	0.01		200	0.697394	0.672518	0.50838
	4	20	5,5	relu	0.0001	adaptive	0.01		100	0.695838	0.693295	0.501676
Neural	5	20	5,5,5	identity	0.001	adaptive	0.001		200	0.62959	0.616579	0.407263
Network	6	20	5,5	relu	0.001	constant	0.001		100	0.67235	0.673692	0.502793
	iteration	mber of Variab	bootstrap_type	max_depth	iterations	l2_leaf_reg	learning_rate		random_state	Train	Test	OOT
	1	15	Bayesian	6	1000	3	0.01		None	0.79907	0.768674	0.441899
	2	15	Bayesian	6	500	12	0.01		None	0.73659	0.726852	0.510056
	3	15	Bernoulli	6	500	3	0.01		None	0.749808	0.714132	0.473743
	4	20	Bayesian	5	500	5	0.01		None	0.727247	0.731414	0.435754
	5	20	MVS	7	500	12	0.02		4	0.862971	0.79281	0.42514
CatBoost	6	20	Bayesian	7	500	15	0.01		4	0.736503	0.737737	0.502793
	iteration	mber of Variab	booster	max_depth	tree_method	min_child_weight	colsample_bytree		n_estimator	Train	Test	OOT
	1	15	gbtree	6	approx	1	1		100	0.944679	0.846824	0.411732
	2	15	gbtree	6	exact	100	1		80	0.74346	0.726354	0.507821
	3	15	dart	6	auto	100	0.8		80	0.74448	0.724806	0.507263
	4	20	gbtree	6	approx	100	1		100	0.733083	0.702899	0.483799
	5	20	gbtree	7	auto	100	0.8		100	0.743161	0.724468	0.503352
XGBoost	6	20	dart	7	auto	200	1		300	0.652562	0.652164	0.385475



Final Model Performance

The final model selected was random forest with *n_estimators*=200, *criterion*='gini', *max_depth*=15, *min_samples_split*=15, *min_samples_leaf*=20, and *max_features*='log2'. It can achieve an average FDR at 3% of 0.758528, 0.726189, 0.531285 for training, testing and OOT respectively.

By taking the average performance of 30 runs, the accurate measures of the average FDR at 3% for the model were: trn 0.755622, tst 0.722669, oot 0.526816.

The detail of the model performances is listed below.

Train

TRN	# Records	# Goods	# Bads	Fraud Rate								
	59,010	58,395	615	0.0104								
	Bin Statistics					Cumulative Statistcis						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Cumulative Goods	% Cumulative Bads(FDR)	KS	FDR
1	590	278	312	47.12%	52.88%	590	278	312	0.48%	50.73%	50.26	0.89
2	590	495	95	83.90%	16.10%	1180	773	407	1.32%	66.18%	64.86	1.90
3	590	528	62	89.49%	10.51%	1770	1301	469	2.23%	76.26%	74.03	2.77
4	590	569	21	96.44%	3.56%	2360	1870	490	3.20%	79.67%	76.47	3.82
5	590	558	32	94.58%	5.42%	2950	2428	522	4.16%	84.88%	80.72	4.65
6	591	572	19	96.79%	3.21%	3541	3000	541	5.14%	87.97%	82.83	5.55
7	590	583	7	98.81%	1.19%	4131	3583	548	6.14%	89.11%	82.97	6.54
8	590	582	8	98.64%	1.36%	4721	4165	556	7.13%	90.41%	83.27	7.49
9	590	587	3	99.49%	0.51%	5311	4752	559	8.14%	90.89%	82.76	8.50
10	590	583	7	98.81%	1.19%	5901	5335	566	9.14%	92.03%	82.90	9.43
11	590	586	4	99.32%	0.68%	6491	5921	570	10.14%	92.68%	82.54	10.39
12	590	584	6	98.98%	1.02%	7081	6505	576	11.14%	93.66%	82.52	11.29
13	590	585	5	99.15%	0.85%	7671	7090	581	12.14%	94.47%	82.33	12.20
14	590	588	2	99.66%	0.34%	8261	7678	583	13.15%	94.80%	81.65	13.17
15	591	587	4	99.32%	0.68%	8852	8265	587	14.15%	95.45%	81.29	14.08
16	590	588	2	99.66%	0.34%	9442	8853	589	15.16%	95.77%	80.61	15.03
17	590	582	8	98.64%	1.36%	10032	9435	597	16.16%	97.07%	80.92	15.80
18	590	590	0	100.00%	0.00%	10622	10025	597	17.17%	97.07%	79.91	16.79
19	590	587	3	99.49%	0.51%	11212	10612	600	18.17%	97.56%	79.39	17.69
20	590	588	2	99.66%	0.34%	11802	11200	602	19.18%	97.89%	78.71	18.60

Test

TST	# Records	# Goods	# Bads	Fraud Rate								
	25,290	25,025	265	0.0105								
	Bin Statistics					Cumulative Statistcis						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Cumulative Goods	% Cumulative Bads(FDR)	KS	FDR
1	253	121	132	47.83%	52.17%	253	121	132	0.48%	49.81%	49.33	0.92
2	253	217	36	85.77%	14.23%	506	338	168	1.35%	63.40%	62.05	2.01
3	253	226	27	89.33%	10.67%	759	564	195	2.25%	73.58%	71.33	2.89
4	253	242	11	95.65%	4.35%	1012	806	206	3.22%	77.74%	74.52	3.91
5	252	243	9	96.43%	3.57%	1264	1049	215	4.19%	81.13%	76.94	4.88
6	253	250	3	98.81%	1.19%	1517	1299	218	5.19%	82.26%	77.07	5.96
7	253	253	0	100.00%	0.00%	1770	1552	218	6.20%	82.26%	76.06	7.12
8	253	250	3	98.81%	1.19%	2023	1802	221	7.20%	83.40%	76.20	8.15
9	253	250	3	98.81%	1.19%	2276	2052	224	8.20%	84.53%	76.33	9.16
10	253	250	3	98.81%	1.19%	2529	2302	227	9.20%	85.66%	76.46	10.14
11	253	251	2	99.21%	0.79%	2782	2553	229	10.20%	86.42%	76.21	11.15
12	253	251	2	99.21%	0.79%	3035	2804	231	11.20%	87.17%	75.97	12.14
13	253	250	3	98.81%	1.19%	3288	3054	234	12.20%	88.30%	76.10	13.05
14	253	252	1	99.60%	0.40%	3541	3306	235	13.21%	88.68%	75.47	14.07
15	253	251	2	99.21%	0.79%	3794	3557	237	14.21%	89.43%	75.22	15.01
16	252	252	0	100.00%	0.00%	4046	3809	237	15.22%	89.43%	74.21	16.07
17	253	252	1	99.60%	0.40%	4299	4061	238	16.23%	89.81%	73.58	17.06
18	253	252	1	99.60%	0.40%	4552	4313	239	17.23%	90.19%	72.95	18.05
19	253	251	2	99.21%	0.79%	4805	4564	241	18.24%	90.94%	72.71	18.94
20	253	252	1	99.60%	0.40%	5058	4816	242	19.24%	91.32%	72.08	19.90

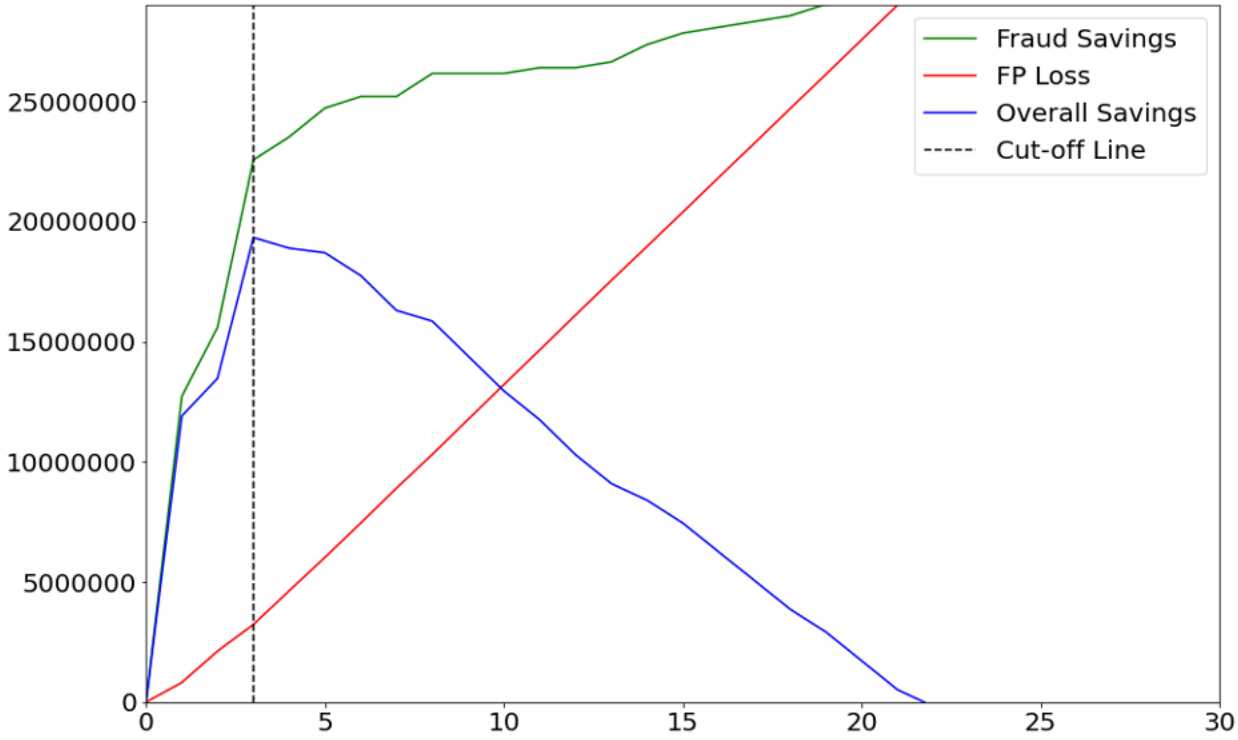
OOT

OOT	# Records	# Goods	# Bads	Fraud Rate								
	12,097	11,918	179	0.0148								
	Bin Statistics					Cumulative Statistcis						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Cumulative Goods	% Cumulative Bads(FDR)	KS	FDR
1	121	68	53	56.20%	43.80%	121	68	53	0.57%	29.61%	29.04	1.28
2	121	109	12	90.08%	9.92%	242	177	65	1.49%	36.31%	34.83	2.72
3	121	92	29	76.03%	23.97%	363	269	94	2.26%	52.51%	50.26	2.86
4	121	117	4	96.69%	3.31%	484	386	98	3.24%	54.75%	51.51	3.94
5	121	116	5	95.87%	4.13%	605	502	103	4.21%	57.54%	53.33	4.87
6	121	119	2	98.35%	1.65%	726	621	105	5.21%	58.66%	53.45	5.91
7	121	121	0	100.00%	0.00%	847	742	105	6.23%	58.66%	52.43	7.07
8	121	117	4	96.69%	3.31%	968	859	109	7.21%	60.89%	53.69	7.88
9	121	121	0	100.00%	0.00%	1089	980	109	8.22%	60.89%	52.67	8.99
10	121	121	0	100.00%	0.00%	1210	1101	109	9.24%	60.89%	51.66	10.10
11	121	120	1	99.17%	0.83%	1331	1221	110	10.25%	61.45%	51.21	11.10
12	121	121	0	100.00%	0.00%	1452	1342	110	11.26%	61.45%	50.19	12.20
13	121	120	1	99.17%	0.83%	1573	1462	111	12.27%	62.01%	49.74	13.17
14	121	118	3	97.52%	2.48%	1694	1580	114	13.26%	63.69%	50.43	13.86
15	121	119	2	98.35%	1.65%	1815	1699	116	14.26%	64.80%	50.55	14.65
16	121	120	1	99.17%	0.83%	1936	1819	117	15.26%	65.36%	50.10	15.55
17	120	119	1	99.17%	0.83%	2056	1938	118	16.26%	65.92%	49.66	16.42
18	121	120	1	99.17%	0.83%	2177	2058	119	17.27%	66.48%	49.21	17.29
19	121	119	2	98.35%	1.65%	2298	2177	121	18.27%	67.60%	49.33	17.99
20	121	120	1	99.17%	0.83%	2419	2297	122	19.27%	68.16%	48.88	18.83

Financial Curves and Recommended Cutoff

Assume $400 gain for every fraud that's caught, $20 loss for every false positive (a good that's flagged as a bad). The amount of fraud saving, lost revenue, and overall savings versus

different cut-off threshold is visualized below. Since the overall savings is maximized at cut-off threshold set to be 3%, the client should set a score cutoff at 3%.



Summary

The project aims to build a transaction fraud detection model with the dataset Card Transaction Data. For the project, data cleaning and imputation are conducted. Then, feature engineering is conducted to build the feature pool for feature selection. Different forward and backward selection models are tried, and the KS score for each feature is computed for the best forward selection model for feature selection. After selecting the best features, multiple machine learning and deep learning models are tried, including Logistic Regression, Boost Tree, Decision Tree, Random Forest, and Neural Network. After tuning the hyperparameters, the performance of the models is evaluated to choose the final model to be deployed. The final model is chosen as a random forest with top 25 features, n_estimator=100, criterion=gini, max_depth=15, min_samples_leaf=200, and max_deatures=log2. The model could achieve a Training accuracy of 0.759, a Testing accuracy rate of 0.726, and a OOT of 0.531. The model can capture 53.1% of all the fraud with FDR at 3%. According to the financial curve, we recommend a score cutoff at 3%, where a maximum overall savings is achieved and is equal to 19,332,000.

Appendix – DQR

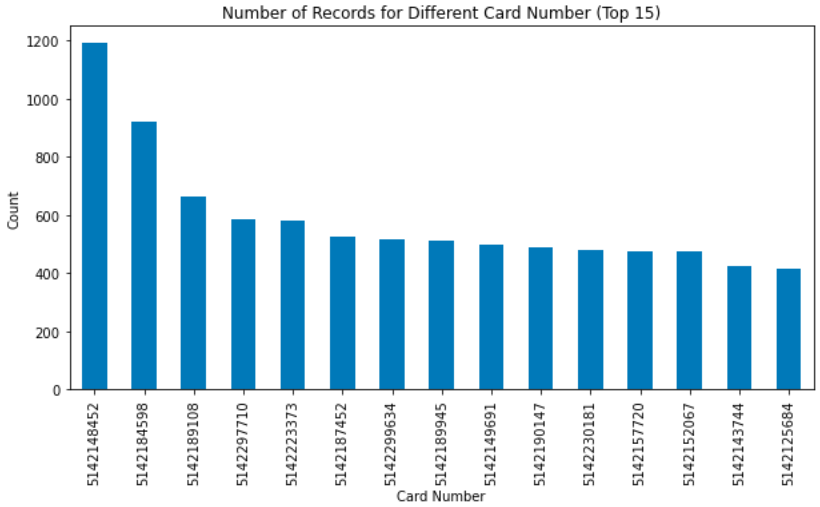
The dataset has 10 different fields, which are *Recnum, Cardnum, Date, Merchnum, Merch description, Merch state, Merch zip, Transtype, Amount, Fraud* respectively. The description of and the visualization of each field can be found below.

1. *Recnum*

The field *Recnum* can be regarded as an index. It is ordinal unique positive integer for each record, from 1 to 96,753.

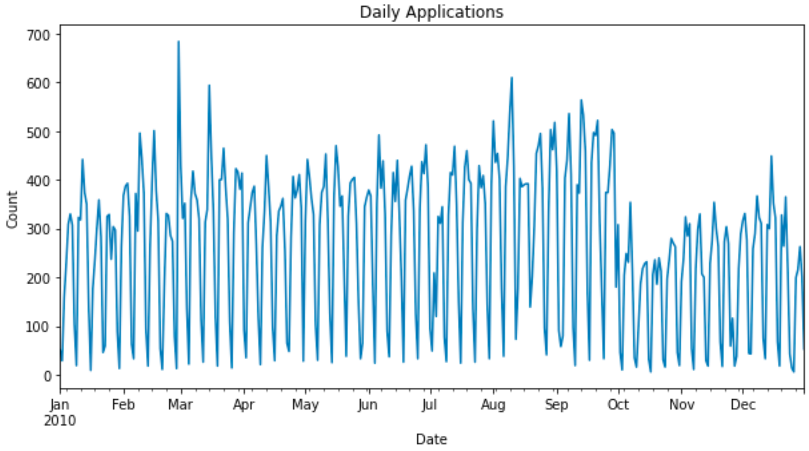
2. *Cardnum*

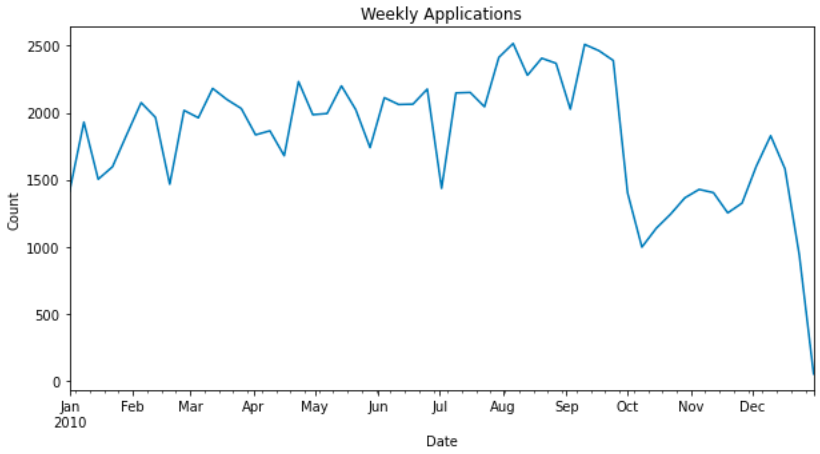
The field *Cardnum* stores the card number for each transaction. It has 1645 unique values. The most frequently seen value for *Cardnum* is 5142148452, which shows up 1192 times in the data set. Below is the visualization for the top 15 most seen values for *Cardnum*.



3. Date

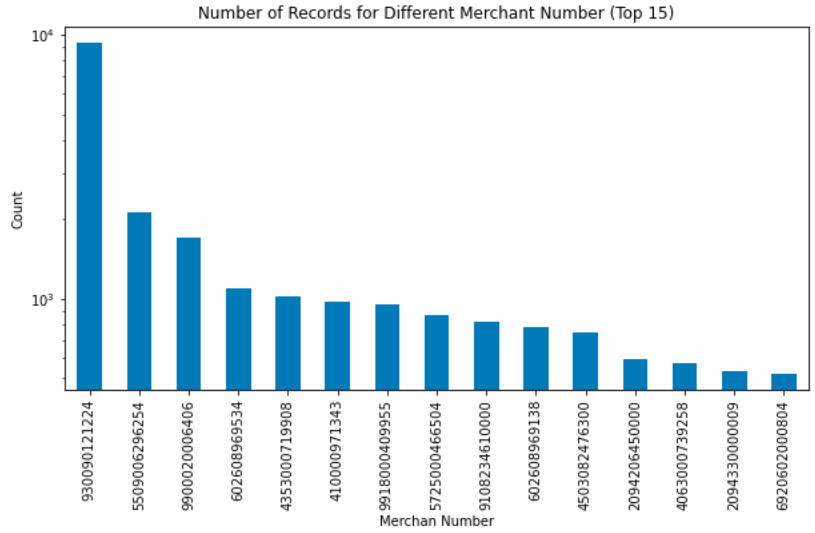
The field *Date* stores the date each transaction takes place. It ranges from 2010-01-01 to 2010-12-31. Below are the visualizations of the distribution of applications.





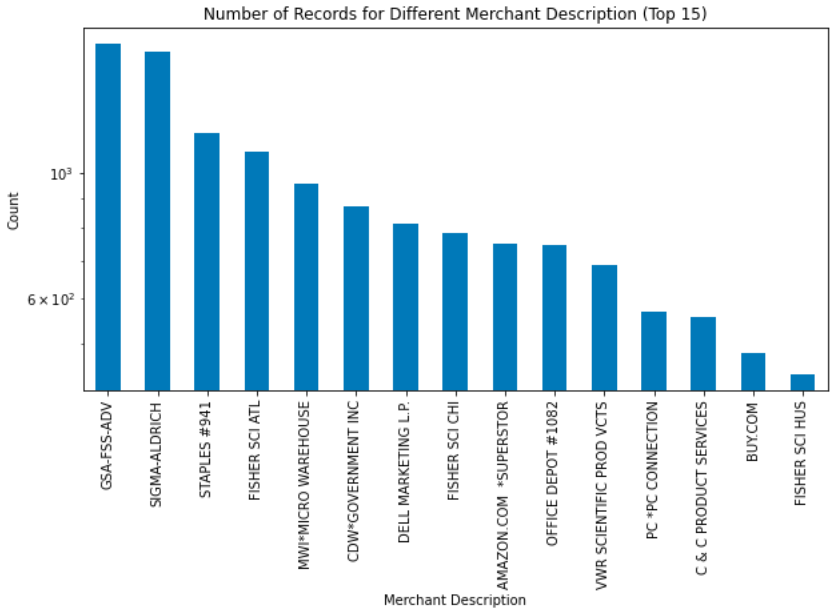
4. *Merchnum*

The field *Merchnum* stores the merchant number for each transaction. It has 13092 unique values. The most frequently seen value for *Merchnum* is 930090121224, which shows up 9310 times in the data set. In addition, it includes 3375 records of ‘’, and they are replaced as np.nan value. Below is the visualization for the top 15 most seen values for *Merchnum* without the nan’s.



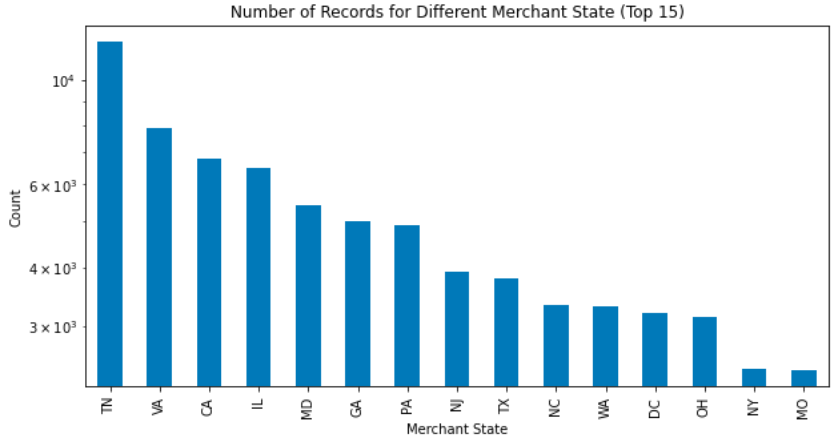
5. *Merch description*

The field *Merch description* stores the description for the merchants involved in each transaction. It has 13126 unique values. The most frequently seen value for *Merch description* is GSA-FSS-ADV, which shows up 1688 times in the data set. Below is the visualization for the top 15 most seen values for *Merch description.*



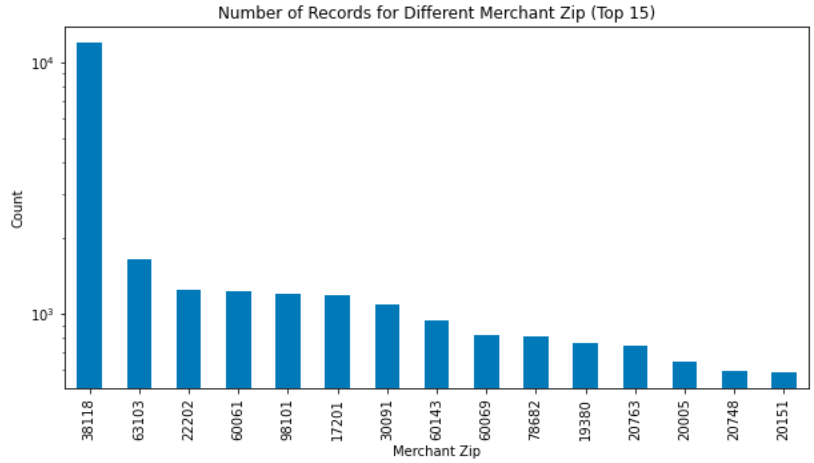
6. *Merch state*

The field *Merch state* stores the state where the merchants are located for the merchants involved in each transaction. It has 228 unique values. The most frequently seen value for *Merch state* is TN, which shows up 12035 times in the data set. Below is the visualization for the top 15 most seen values for *Merch state.*



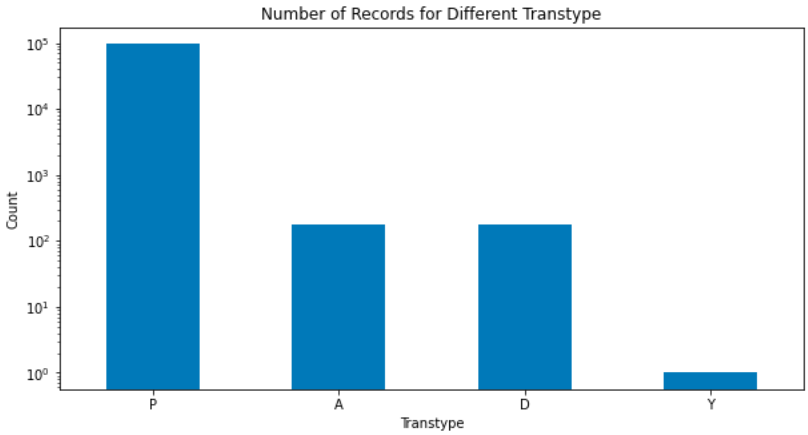
7. *Merch zip*

The field *Merch zip* stores the zipcode for the merchants involved in each transaction. It has 4568 unique values. The most frequently seen value for *Merch zip* is 38118, which shows up 11868 times in the data set. In addition, there are 8098 records of value that is not in the standard form of zip code, i.e. 5 digit number, and those records are replaced with np.nan values. Below is the visualization for the top 15 most seen values for *Merch zip.*



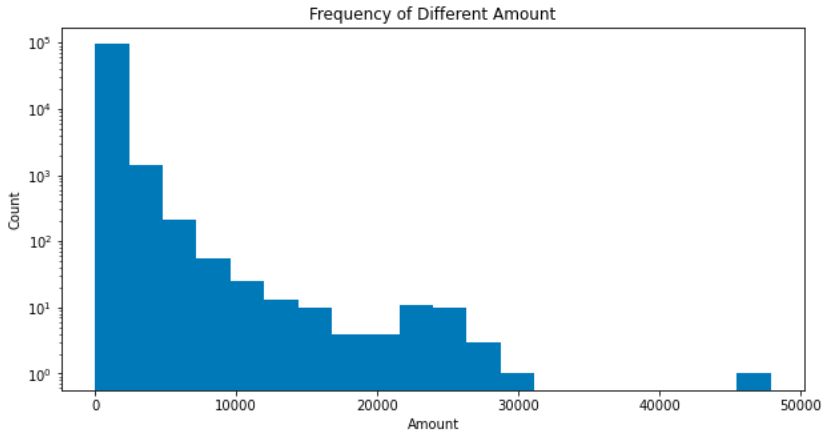
8. *Transtype*

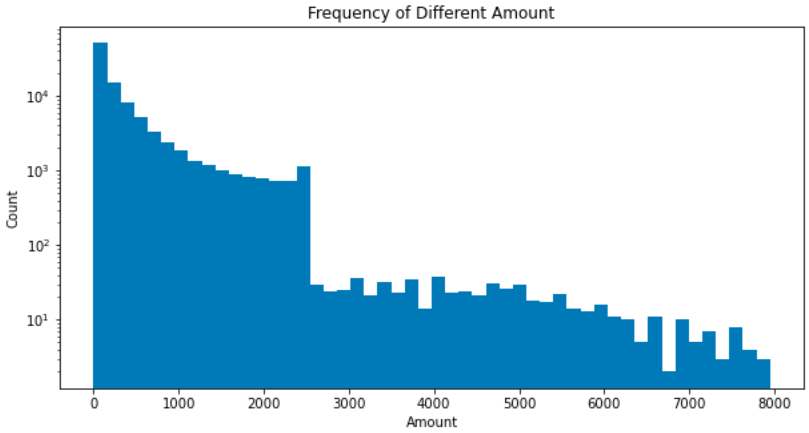
The field *Transtype* stores the transaction type for each transaction. It has 4 unique values. The most frequently seen value for *Transtype* is P, which shows up 96398 times in the data set. Below is the visualization for the top 15 most seen values for *Transtype.*



9. *Amount*

The field *Amount* stores the volume of each transaction. It ranges from 0.01to 3,102,046.53. After removing the maximum value 3,102,046.53, the distribution can be visualized as below. One thing to be noted is that there is a significant drop in count around amount equals $2500.





10. *Fraud*

The field *Fraud* stores the tag indicating whether a transaction is regarded as a fraud. It uses 1 to denote a fraud and 0 to denote a normal transaction. In the data set, each application is given a *Fraud*. The most common value is 0, which has 95694 records, and the value 1 has 1059 records.