Credit Card Fraud Detection Utilizing Supervised Machine Learning

1. Executive Summary

The aim of this project was to precisely detect fraudulent credit card transactions with the purpose of reducing the financial risk associated with fraud and maximizing savings. To identify fraudulent transactions, a machine learning model was trained and tested on the dataset and assessed for accuracy based on previously identified frauds. The final model will eliminate 56.42% of fraud by rejecting only 3% of transactions, resulting in $21,000,000 in savings per year. Ultimately this model can be used to reject the fewest amount of credit card transactions possible while still optimizing for savings.

2. Description of Data

The data was composed of company credit card transactions from a US government organization. The dataset had **10 fields**, **96,753 records** and covered the year **2010**. Of the fields, 2 were numerical and 8 were categorical.

(1) Numerical Table

Field Name	% Pop.	Min	Max	Mean	Std. Dev.	% Zero
Amount	100.00	0.01	3,102,045.53	427.89	10,006.14	0.00
Date	100.00	2010-01-01	2010-12-31	N/A	N/A	0.00

(2) Categorical Table

Field Name	%Populated	# Blank	# Zeros	# Unique Values	Most Common Value
Recnum	100.00	0	0	96,753	1
Cardnum	100.00	0	0	1,645	5142148452
Merchnum	96.51	3,375	231	13,091	930090121224
Merch Description	100.00	0	3	13,126	GSA-FSS-ADV
Merch State	98.76	1,195	0	227	TN
Merch Zip	95.19	4,656	0	4,567	38118
Transtype	100.00	0	0	4	P
Fraud	100.00	0	0	2	0.0

3. Data Cleaning

In terms of data cleaning, an extreme outlier was removed from Amount, and the Transtype field was filtered to only include 'P' transactions, which were the majority of the data. This prevented unusual values from skewing the data results. Additionally, Date was converted to datetime format.

There were also substantial NA values in the Merchnum, Merch State and Merch Zip fields. For Merchnum, missing values were imputed with the corresponding Merch Description for that record. For Merch State missing values were first imputed by mapping the zip code of the record to the correct state and filling in the NA with that value. Remaining missing values were then imputed with the corresponding Merchnum or Merch Description of the record. For Merch Zip missing values were imputed with the corresponding Merchnum or corresponding Merch Description of the record. For all three fields, all adjustment transaction records were set to 'unknown' and any remaining NA values after imputation were also set to 'unknown'.

4. Variable Creation

Additional variables were created to catch credit card transaction fraud. Transaction fraud includes fraud during the account usage process and involves an existing account. Credit card transaction fraud can stem from a lost or stolen credit card, a skimmed credit card, or an online hack. The additional variables were created to catch any unusual or out of the ordinary behavior in credit card usage that could indicate the credit card information was being used by a fraudster.

During the variable creation step, a total of 1,383 variables were created. This included two Benford's Law variables, which measured the unusualness of the first digit of the Merchnumber and Cardnumber fields. Amount variables calculated the average, max, median and total amount spent for each entity over a particular time period, and an Amount Bin variable was also created to group the transaction amounts into 5 bins to make the amount field more useful for analysis.

Two new variables were also created that were specific to Project 2. A Gas Station variable indicated whether a transaction took place at a gas station, which is the most common location for credit card transaction fraud. A New Purchase Location variable was also created that indicated if the next purchase made on the same Cardnumber was in a different Zip Code or State, which could track stolen cards taken to a different location.

Description of Variables	# Variables Created
Original fields from the dataset (record and fraud label)	2
Risk Variable: The likelihood of fraud for any day of the week	1
Benford's Law Variables: Measure of unusualness according to Benford's Law of the first digit of Merchnumber and Cardnumber	2
Days Since Variables: # of days since a record with that entity has been seen	11

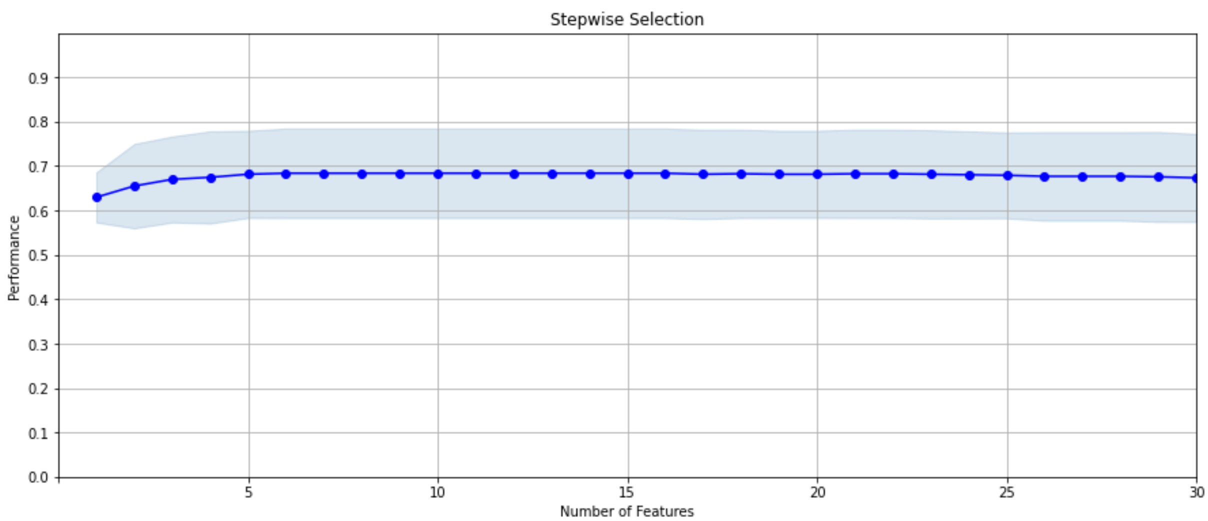
Amount Variables: Calculated average, max, median and total Amount spent for each entity over {0,1,3,7,14,30} days	594
Velocity Change: # of records with the same entity within the last 0-1 days over the # of records with the same entity over 7, 14 and 30 days	66
Velocity Days Since: Velocity Change over the # of days since a record with that entity has been seen	66
Cross-Entity Uniqueness: # of distinct records for an entity that are present for each group of a different entity	53
Grouped Relative Velocity: # of records with the same entity within the last 0-1 days over the # of records with the same entity over 7, 14 and 30 days grouped by Cardnumber	8
Variability: Calculated average, max and median difference in Amount spent for each entity over {0,1,3,7,14,30} days	198
Amount Bin: Amount values grouped into 5 bins to make them more useful for analysis	1
Velocity Change Squared: # of records with the same entity within the last 0-1 days over the # of records with the same entity over 7, 14 and 30 days squared and normalized by the longer time period	66
Unique Count Variables: # of unique records of an entity within a rolling time window of {1,3,7,14,30,60} days for each value of a different entity	312
***New* Gas Station Variable**: Indicates whether a purchase took place at a gas station. Created because gas stations are the most common place for credit card fraud	1
***New* Next Purchase Location Variables**: Indicates whether the next purchase made on the same Cardnumber was in a different Zip Code or State. Created to track potentially stolen cards taken to a different location.	2
Total Variables (After Dedup)	**1,383**

5. Feature Selection

Feature Selection is an important step before model exploration to reduce dimensionality that will make nonlinear models harder to fit. Feature selection generally involves two components: a filter and a wrapper. The filter sorts the variables in order of their importance in predicting the y variable to reduce the list of candidate variables. Then a wrapper is used to run many models that add or remove variables at each step to find the optimal order and number of variables to include.

For Project 2, the Benford's Law variables were first removed as they led to some unnatural behavior when they were included in the wrapper. Additionally, OOT data was included in the feature selection steps for this project because some seasonal differences were detected in the OOT data that led to poor OOT performance in the model exploration step when it was excluded from feature selection.

A univariate KS filter was first used to rank the variables by importance in predicting the y variable, Fraud, and reduce the number of variables to 300. Multiple wrappers were tested, including LGBM and Random Forest with both forward and backward selection methods. Ultimately, an LGBM wrapper utilizing forward selection had the most consistent performance and was chosen to generate a finalized list of 20 variables that achieved a peak performance of 0.69 FDR. The performance of the LGBM forward selection wrapper (Filter Number = 300, Wrapper Number = 30) and the list of the final 20 variables ranked in wrapper order are below.



Wrapper Order	Variable Name	Univariate KS
1	Merchnum_max_30	0.48
2	Card_Merchnum_desc_total_30	0.66
3	Merchnum_desc_med_60	0.39
4	Card_Merchdesc_total_30	0.66
5	Merchnum_med_7	0.45
6	Card_Merchdesc_total_60	0.65
7	Card_Merchnum_desc_total_60	0.65
8	Merchnum_desc_total_30	0.51

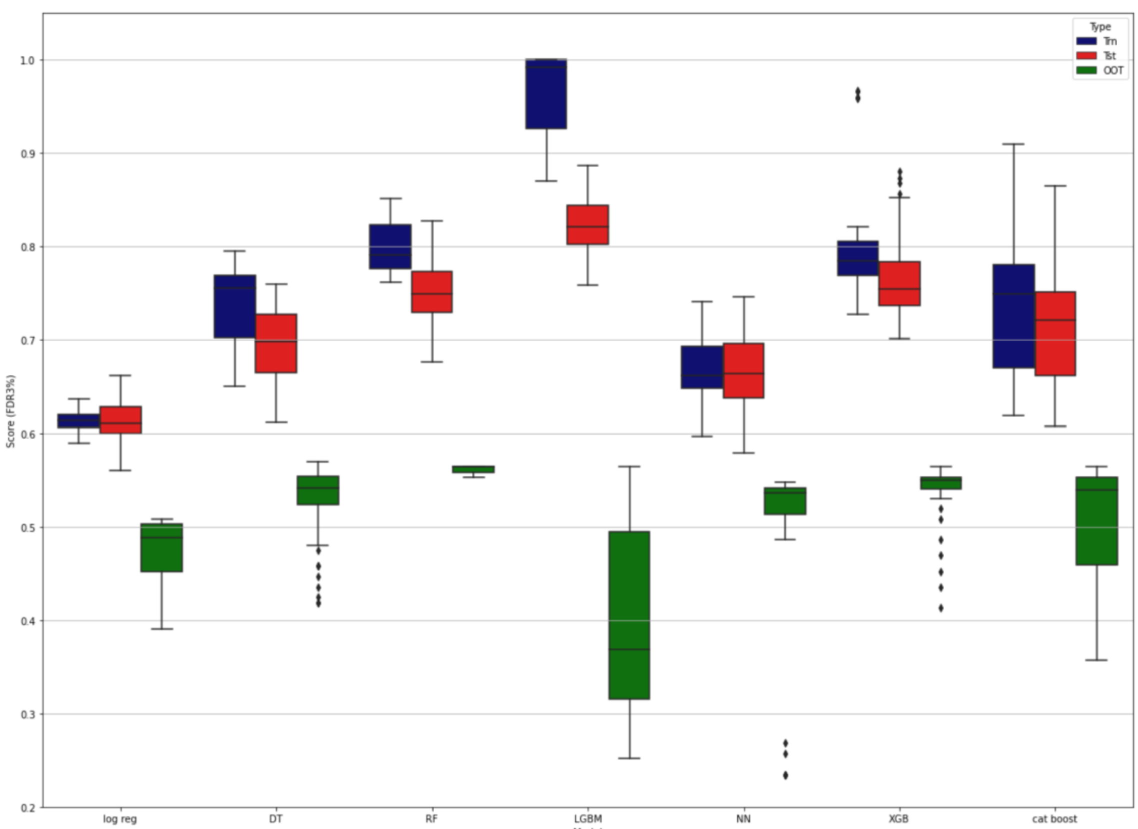
9	merch_zip_max_30	0.48
10	Merchnum_desc_total_60	0.46
11	merch_zip_med_7	0.45
12	Merchnum_desc_med_30	0.42
13	merch_zip_med_30	0.41
14	Merchnum_total_30	0.43
15	Merchnum_med_30	0.41
16	card_merch_total_60	0.64
17	Cardnum_total_60	0.45
18	Merchnum_desc_med_7	0.45
19	zip3_total_amount_0_by_60	0.39
20	merch_zip_total_30	0.43

6. Preliminary Model Exploration

Using the finalized list of variables from the wrapper, many different models were run to identify the one with optimal performance. The chart below displays each type of model that was run and their output. All models were run separately on the training set, the test set and the OOT set. For each output, the model was run 5 times and the performance across each run was averaged to give a singular FDR score.

The simplest model, Logistic Regression, was run first to get a baseline performance to compare the nonlinear models to. For all models, the hyperparameters were tuned to optimize performance. A summary of the performance of each model can be found below as well as a visualization of the training, testing and OOT performance for each of the top performing models of each type.

Model		Dataset	Parameters						Average FDR at 3%		
	Iteration	NVARS	max_iter	penalty	c	solver			Train	Test	OOT
Logistic Regression	1	10	1000	l2	1	lbfgs			0.624	0.592	0.478
	2	10	100	l1	0.5	liblinear			0.613	0.611	0.468
	3	10	500	l1	0.5	liblinear			0.619	0.613	0.476
	4	10	500	l2	3	lbfgs			0.609	0.616	**0.482**
	Iteration	NVARS	criterion	max_depth	min_samples_split	min_samples_leaf	splitter		Train	Test	OOT
Decision Tree	1	10	gini	None	1000	10	best		0.771	0.700	0.554
	2	10	entropy	None	1000	10	best		0.734	0.681	0.521
	3	10	gini	None	1000	20	best		0.755	0.702	0.549
	4	10	gini	None	750	20	best		0.762	0.728	0.542
	5	10	gini	None	750	5	best		0.771	0.726	**0.555**
	Iteration	NVARS	n_estimators	criterion	max_depth	min_samples_split	min_samples_leaf	max_features	Train	Test	OOT
Random Forest	1	**10**	**100**	**gini**	**none**	**1000**	**20**	**8**	**0.775**	**0.751**	**0.564**
	2	10	100	gini	none	1000	10	10	0.771	0.726	0.555
	3	10	150	gini	none	500	20	8	0.829	0.765	0.558
	4	10	100	gini	none	500	10	8	0.837	0.791	0.559
	5	10	50	gini	none	1000	20	8	0.800	0.720	0.564
	Iteration	NVARS	n_estimators	num_leaves	max_depth	boosting type	min_data_in_leaf		Train	Test	OOT
LightGBM (Boost)	1	10	200	100	20	GOSS	1000		0.920	0.839	0.512
	2	10	200	100	none	gbdt	1000		0.976	0.853	0.381
	3	10	200	1000	none	gbdt	1000		0.977	0.843	0.385
	4	10	100	1000	none	GOSS	1000		0.883	0.839	**0.543**
	5	10	500	100	none	GOSS	500		0.922	0.851	0.502
	Iteration	NVARS	hidden_layer_size	activation	alpha	learning_rate	solver	max_iter	Train	Test	OOT
Neural Net (NN)	1	10	20	relu	0.1	constant	adam	50	0.653	0.666	0.542
	2	10	50	relu	0.1	constant	adam	100	0.655	0.660	**0.542**
	3	10	50	relu	0.1	adaptive	adam	100	0.661	0.637	0.536
	4	10	100	relu	0.01	adaptive	lbfgs	200	0.731	0.704	0.512
	5	10	20	relu	0.01	adaptive	lbfgs	200	0.712	0.700	0.516
	6	10	20	relu	0.01	constant	adam	200	0.685	0.690	0.509
	Iteration	NVARS	n_estimators	max_depth	tree_method	min_child_weight	subsample	booster	Train	Test	OOT
XGBoost	1	10	100	10	auto	100	1	gbtree	0.771	0.739	0.550
	2	10	200	10	auto	100	1	gbtree	0.789	0.761	0.553
	3	10	300	5	auto	100	1	gbtree	0.811	0.767	**0.554**
	4	10	300	5	exact	100	0.8	gbtree	0.775	0.747	0.540
	5	10	200	10	exact	10	0.8	gbtree	0.962	0.866	0.451
	Iteration	NVARS	max_depth	iterations	bootstrap_type	learning_rate			Train	Test	OOT
CatBoost	1	10	10	200	Bayesian	0.1			0.902	0.843	0.449
	2	10	15	50	Bayesian	0.1			0.777	0.732	0.551
	3	10	15	50	Bernoulli	0.01			0.671	0.641	0.456
	4	10	16	25	Bayesian	0.1			0.74	0.717	0.542
	5	10	10	100	Bayesian	0.05			0.767	0.737	**0.556**



7. Final Model Performance

The final selected model was a Random Forest model with the following parameters: 10 Variables, 100 Estimators, Gini Criterion, No Max Depth, 1,000 Minimum Sample Split, 20 Minimum Sample Leaf and 8 Max Features. At 3% this model had the following FDRs for each set: 73.52% Training, 72.60% Testing, and 56.42% OOT. The 56.42% FDR for the OOT set was the highest achieved of all the models and indicates that rejecting only 3% of transactions will eliminate 56.42% of fraud. Below are detailed tables of model performance for the final Random Forest model for the Training, Test and OOT sets.

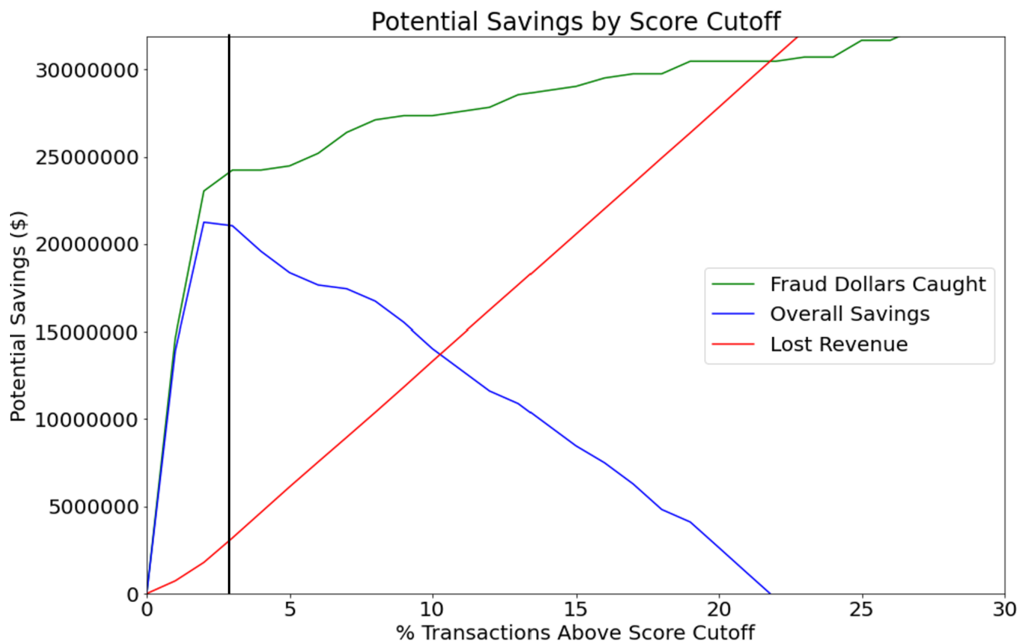
Training	# Records	# Goods	# Bads	Fraud Rate								
	17,674	17,052	622	0.0364767								
Bin Statistics						Cumulative Statistics						
bin	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
0	0	0	0	0.00%	0.00%	0	0	0	0.00%	0.00%	0.00	0.00
1	589	272	317	46.18%	53.82%	589	272	317	0.47%	50.88%	50.42	0.86
2	589	478	111	81.15%	18.85%	1,178	750	428	1.29%	68.70%	67.41	1.75
3	589	559	30	94.91%	5.09%	1,767	1,309	458	2.25%	73.52%	71.27	2.86
4	589	562	27	95.42%	4.58%	2,356	1,871	485	3.21%	77.85%	74.64	3.86
5	590	544	46	92.20%	7.80%	2,946	2,415	531	4.14%	85.23%	81.09	4.55
6	589	571	18	96.94%	3.06%	3,535	2,986	549	5.12%	88.12%	83.00	5.44
7	589	570	19	96.77%	3.23%	4,124	3,556	568	6.10%	91.17%	85.07	6.26
8	589	576	13	97.79%	2.21%	4,713	4,132	581	7.09%	93.26%	86.17	7.11
9	589	585	4	99.32%	0.68%	5,302	4,717	585	8.09%	93.90%	85.81	8.06
10	589	579	10	98.30%	1.70%	5,891	5,296	595	9.09%	95.51%	86.42	8.90
11	589	580	9	98.47%	1.53%	6,480	5,876	604	10.08%	96.95%	86.87	9.73
12	589	587	2	99.66%	0.34%	7,069	6,463	606	11.09%	97.27%	86.18	10.67
13	590	587	3	99.49%	0.51%	7,659	7,050	609	12.09%	97.75%	85.66	11.58
14	589	588	1	99.83%	0.17%	8,248	7,638	610	13.10%	97.91%	84.81	12.52
15	589	587	2	99.66%	0.34%	8,837	8,225	612	14.11%	98.23%	84.12	13.44
16	589	586	3	99.49%	0.51%	9,426	8,811	615	15.12%	98.72%	83.60	14.33
17	589	588	1	99.83%	0.17%	10,015	9,399	616	16.12%	98.88%	82.75	15.26
18	589	589	0	100.00%	0.00%	10,604	9,988	616	17.14%	98.88%	81.74	16.21
19	589	588	1	99.83%	0.17%	11,193	10,576	617	18.14%	99.04%	80.89	17.14
20	589	587	2	99.66%	0.34%	11,782	11,163	619	19.15%	99.36%	80.21	18.03
21	590	589	1	99.83%	0.17%	12,372	11,752	620	20.16%	99.52%	79.36	18.95
22	589	589	0	100.00%	0.00%	12,961	12,341	620	21.17%	99.52%	78.35	19.90
23	589	589	0	100.00%	0.00%	13,550	12,930	620	22.18%	99.52%	77.34	20.85
24	589	588	1	99.83%	0.17%	14,139	13,518	621	23.19%	99.68%	76.49	21.77
25	589	589	0	100.00%	0.00%	14,728	14,107	621	24.20%	99.68%	75.48	22.72
26	589	589	0	100.00%	0.00%	15,317	14,696	621	25.21%	99.68%	74.47	23.67
27	589	589	0	100.00%	0.00%	15,906	15,285	621	26.22%	99.68%	73.46	24.61
28	589	588	1	99.83%	0.17%	16,495	15,873	622	27.23%	99.84%	72.61	25.52
29	589	589	0	100.00%	0.00%	17,084	16,462	622	28.24%	99.84%	71.60	26.47
30	590	590	0	100.00%	0.00%	17,674	17,052	622	29.25%	99.84%	70.59	27.41

Testing	# Records	# Goods	# Bads	Fraud Rate								
	7,575	7,331	244	0.0332833								
Bin Statistics						Cumulative Statistics						
bin	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
0	0	0	0	0.00%	0.00%	0	0	0	0.00%	0.00%	0.00	0.00
1	252	115	137	45.63%	54.37%	252	115	137	0.46%	53.31%	52.85	0.84
2	253	211	42	83.40%	16.60%	505	326	179	1.30%	69.65%	68.35	1.82
3	252	247	5	98.02%	1.98%	757	573	184	2.29%	71.60%	69.30	3.11
4	253	237	16	93.68%	6.32%	1,010	810	200	3.24%	77.82%	74.58	4.05
5	252	237	15	94.05%	5.95%	1,262	1,047	215	4.19%	83.66%	79.47	4.87
6	253	248	5	98.02%	1.98%	1,515	1,295	220	5.18%	85.60%	80.42	5.89
7	252	251	1	99.60%	0.40%	1,767	1,546	221	6.19%	85.99%	79.81	7.00
8	253	248	5	98.02%	1.98%	2,020	1,794	226	7.18%	87.94%	80.76	7.94
9	252	249	3	98.81%	1.19%	2,272	2,043	229	8.17%	89.11%	80.93	8.92
10	253	253	0	100.00%	0.00%	2,525	2,296	229	9.19%	89.11%	79.92	10.03
11	252	249	3	98.81%	1.19%	2,777	2,545	232	10.18%	90.27%	80.09	10.97
12	253	252	1	99.60%	0.40%	3,030	2,797	233	11.19%	90.66%	79.47	12.00
13	252	251	1	99.60%	0.40%	3,282	3,048	234	12.20%	91.05%	78.85	13.03
14	253	249	4	98.42%	1.58%	3,535	3,297	238	13.19%	92.61%	79.41	13.85
15	252	250	2	99.21%	0.79%	3,787	3,547	240	14.19%	93.39%	79.19	14.78
16	253	253	0	100.00%	0.00%	4,040	3,800	240	15.20%	93.39%	78.18	15.83
17	252	252	0	100.00%	0.00%	4,292	4,052	240	16.21%	93.39%	77.17	16.88
18	253	253	0	100.00%	0.00%	4,545	4,305	240	17.23%	93.39%	76.16	17.94
19	252	252	0	100.00%	0.00%	4,797	4,557	240	18.23%	93.39%	75.15	18.99
20	253	252	1	99.60%	0.40%	5,050	4,809	241	19.24%	93.77%	74.53	19.95
21	252	251	1	99.60%	0.40%	5,302	5,060	242	20.25%	94.16%	73.92	20.91
22	253	253	0	100.00%	0.00%	5,555	5,313	242	21.26%	94.16%	72.90	21.95
23	252	252	0	100.00%	0.00%	5,807	5,565	242	22.27%	94.16%	71.90	23.00
24	253	253	0	100.00%	0.00%	6,060	5,818	242	23.28%	94.16%	70.88	24.04
25	252	251	1	99.60%	0.40%	6,312	6,069	243	24.28%	94.55%	70.27	24.98
26	253	253	0	100.00%	0.00%	6,565	6,322	243	25.30%	94.55%	69.26	26.02
27	252	252	0	100.00%	0.00%	6,817	6,574	243	26.30%	94.55%	68.25	27.05
28	253	253	0	100.00%	0.00%	7,070	6,827	243	27.32%	94.55%	67.24	28.09
29	252	251	1	99.60%	0.40%	7,322	7,078	244	28.32%	94.94%	66.62	29.01
30	253	253	0	100.00%	0.00%	7,575	7,331	244	29.33%	94.94%	65.61	30.05

OOT	# Records	# Goods	# Bads	Fraud Rate								
	3,671	3,533	138	0.0390603								
Bin Statistics						Cumulative Statistics						
bin	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
0	0	0	0	0.00%	0.00%	0	0	0	0.00%	0.00%	0.00	0.00
1	122	61	61	50.00%	50.00%	122	61	61	0.51%	34.08%	33.57	1.00
2	123	88	35	71.54%	28.46%	245	149	96	1.24%	53.63%	52.40	1.55
3	122	117	5	95.90%	4.10%	367	266	101	2.21%	56.42%	54.22	2.63
4	122	122	0	100.00%	0.00%	489	388	101	3.22%	56.42%	53.21	3.84
5	123	122	1	99.19%	0.81%	612	510	102	4.23%	56.98%	52.75	5.00
6	122	119	3	97.54%	2.46%	734	629	105	5.22%	58.66%	53.44	5.99
7	123	118	5	95.93%	4.07%	857	747	110	6.20%	61.45%	55.26	6.79
8	122	119	3	97.54%	2.46%	979	866	113	7.18%	63.13%	55.95	7.66
9	122	121	1	99.18%	0.82%	1,101	987	114	8.19%	63.69%	55.50	8.66
10	123	123	0	100.00%	0.00%	1,224	1,110	114	9.21%	63.69%	54.48	9.74
11	122	121	1	99.18%	0.82%	1,346	1,231	115	10.21%	64.25%	54.04	10.70
12	122	121	1	99.18%	0.82%	1,468	1,352	116	11.21%	64.80%	53.59	11.66
13	123	120	3	97.56%	2.44%	1,591	1,472	119	12.21%	66.48%	54.27	12.37
14	122	121	1	99.18%	0.82%	1,713	1,593	120	13.21%	67.04%	53.83	13.28
15	122	121	1	99.18%	0.82%	1,835	1,714	121	14.22%	67.60%	53.38	14.17
16	123	121	2	98.37%	1.63%	1,958	1,835	123	15.22%	68.72%	53.50	14.92
17	122	121	1	99.18%	0.82%	2,080	1,956	124	16.22%	69.27%	53.05	15.77
18	122	122	0	100.00%	0.00%	2,202	2,078	124	17.23%	69.27%	52.04	16.76
19	123	120	3	97.56%	2.44%	2,325	2,198	127	18.23%	70.95%	52.72	17.31
20	122	122	0	100.00%	0.00%	2,447	2,320	127	19.24%	70.95%	51.71	18.27
21	123	123	0	100.00%	0.00%	2,570	2,443	127	20.26%	70.95%	50.69	19.24
22	122	122	0	100.00%	0.00%	2,692	2,565	127	21.27%	70.95%	49.68	20.20
23	122	121	1	99.18%	0.82%	2,814	2,686	128	22.28%	71.51%	49.23	20.98
24	123	123	0	100.00%	0.00%	2,937	2,809	128	23.30%	71.51%	48.21	21.95
25	122	118	4	96.72%	3.28%	3,059	2,927	132	24.28%	73.74%	49.47	22.17
26	122	122	0	100.00%	0.00%	3,181	3,049	132	25.29%	73.74%	48.45	23.10
27	123	120	3	97.56%	2.44%	3,304	3,169	135	26.28%	75.42%	49.14	23.47
28	122	120	2	98.36%	1.64%	3,426	3,289	137	27.28%	76.54%	49.26	24.01
29	122	121	1	99.18%	0.82%	3,548	3,410	138	28.28%	77.09%	48.81	24.71
30	123	123	0	100.00%	0.00%	3,671	3,533	138	29.30%	77.09%	47.79	25.60

8. Financial Curve & Recommended Cutoff

The graph below represents the potential savings by score cutoff. A score cutoff of 3% has been chosen, represented by the black line in the graph below. This cutoff was chosen because it denies the fewest amount of credit card transactions possible while still leading to strong overall savings. It is far left enough to capture the peak savings and far right enough to capture the sharpest fraud increase. Overall, a 3% cutoff will lead to approximately $21,000,000 in savings per year.



9. Summary of Results

To get the final model results, the data was taken through the entire fraud analytics pipeline. First the data was cleaned, using exclusions, imputing null values, and removing an outlier. A total of 1,383 variables were then created to identify unusual behavior in the credit card transactions that could indicate fraud. The variables were then reduced to a total of 20 using feature selection. A univariate KS filter sorted the variables in order of their importance in predicting the y variable, Fraud. After the filter reduced the variables to 300, an LGBM wrapper utilizing forward selection further reduced and ranked the top 20 final variables to be used during model exploration.

After testing multiple machine learning models with different hyperparameter combinations, the model that achieved the highest FDR on the OOT data was a Random Forest model. The final model achieved a 56.42% FDR on the OOT set, which indicates that rejecting only 3% of transactions will eliminate 56.42% of fraud. The 3% cutoff was chosen by analyzing the optimal balance of savings and fraud dollars caught for this model, and ultimately with a 3% rejection rate there will be approximately $21,000,000 in yearly savings.

A potential next step in the project that could be worth investigating is looking into the seasonality of the data. There was a large discrepancy in performance between the OOT and training/testing sets, indicating a seasonal factor may be affecting the final 2 months of data contained in OOT. While including the OOT data in feature selection helped to mitigate some of this discrepancy, it could be worth investigating further how to account for this seasonality in the data. Additionally, research into why there is a seasonal change would also be helpful when building the model.

10. Appendix - Data Quality Report

1. Data Description

The data is composed of company credit card transactions. The company is a US government organization. The dataset has **10 fields**, **96,753 records** and covers the year **2010**. Of the fields, 2 are numerical and 8 are categorical.

2. Summary Tables

(3) Numerical Table

Field Name	% Pop.	Min	Max	Mean	Std. Dev.	% Zero
Amount	100.00	0.01	3,102,045.53	427.89	10,006.14	0.00
Date	100.00	2010-01-01	2010-12-31	N/A	N/A	0.00

(4) Categorical Table

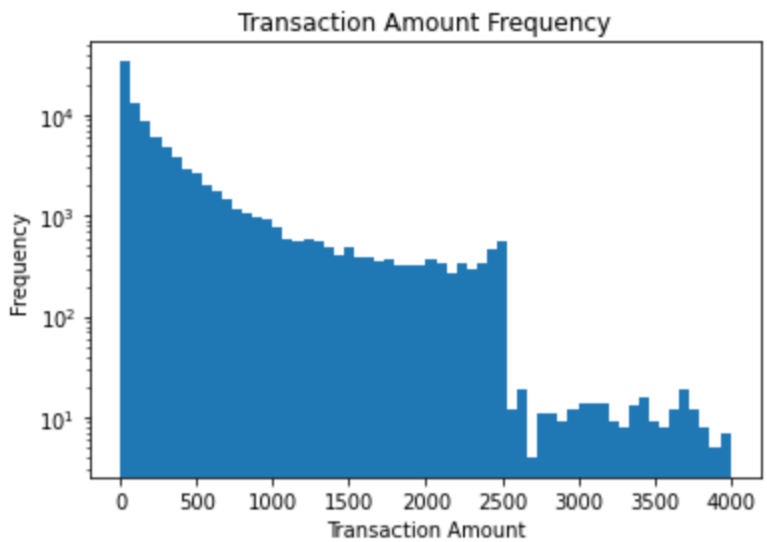
Field Name	%Populated	# Blank	# Zeros	# Unique Values	Most Common Value
Recnum	100.00	0	0	96,753	1
Cardnum	100.00	0	0	1,645	5142148452
Merchnum	96.51	3,375	231	13,091	930090121224
Merch Description	100.00	0	3	13,126	GSA-FSS-ADV
Merch State	98.76	1,195	0	227	TN
Merch Zip	95.19	4,656	0	4,567	38118
Transtype	100.00	0	0	4	P
Fraud	100.00	0	0	2	0.0

3. Visualization of Fields

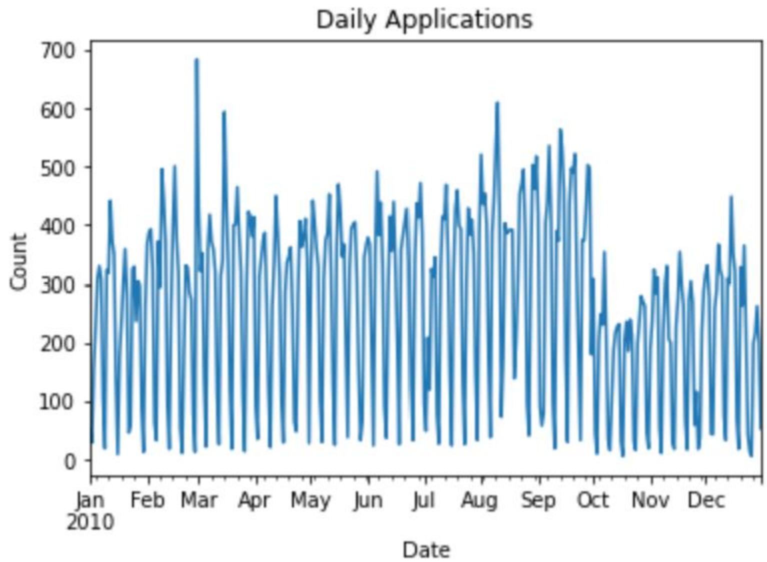
(1) Field Name: Record

Description: Ordinal unique positive integer for each credit card transaction, from 1 to 96,753.

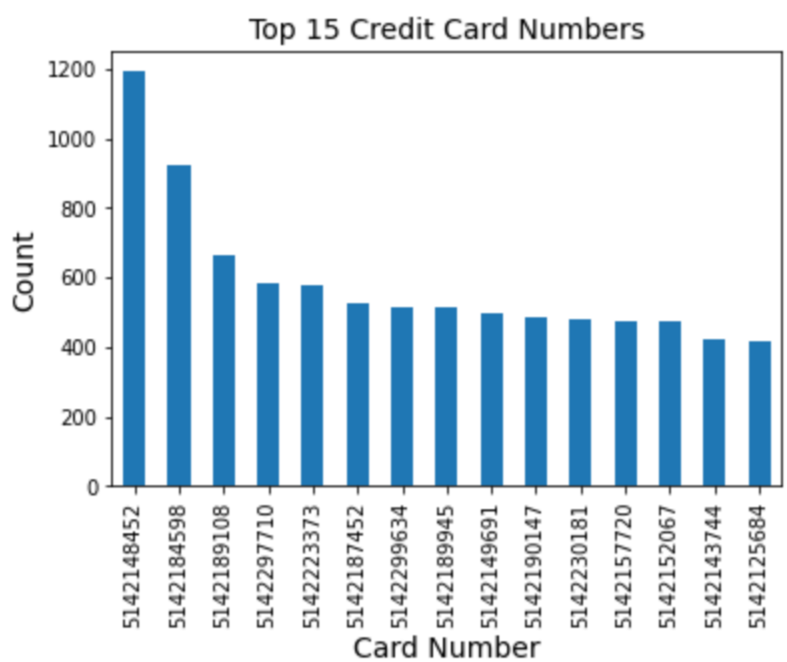
(2) Field Name: Amount
Description: The amount spent in each transaction, in dollars.
Low value transactions are more common, and there is a steep drop off in frequency at $2,500.



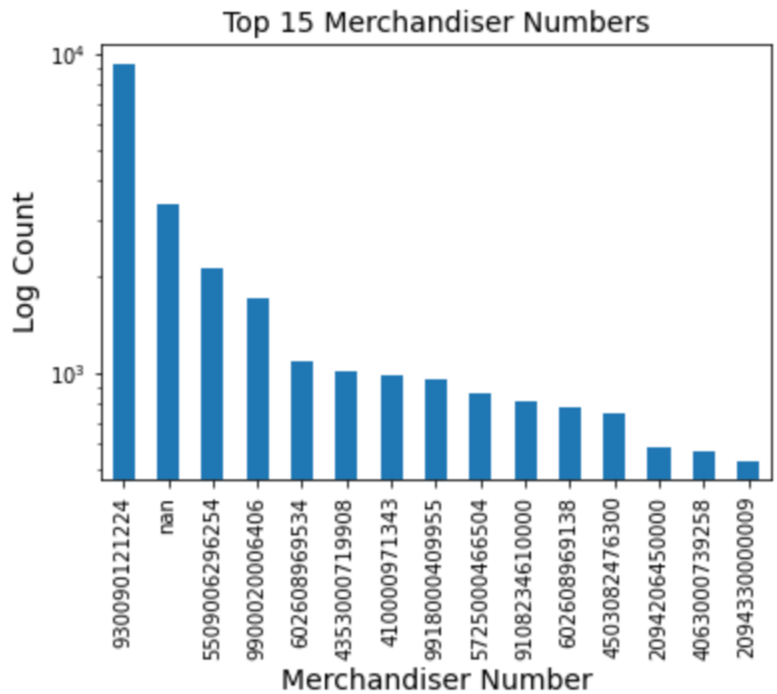
(3) Field Name: Date
Description: The date of each credit card transaction.
The distribution shows the daily number of applications submitted for the date range of 1/1/2010 – 12/31/2010.



(4) Field Name: Cardnum
Description: The credit card number used in the transaction.
The most common credit card number is 5142148452, which has a count of 1,192.



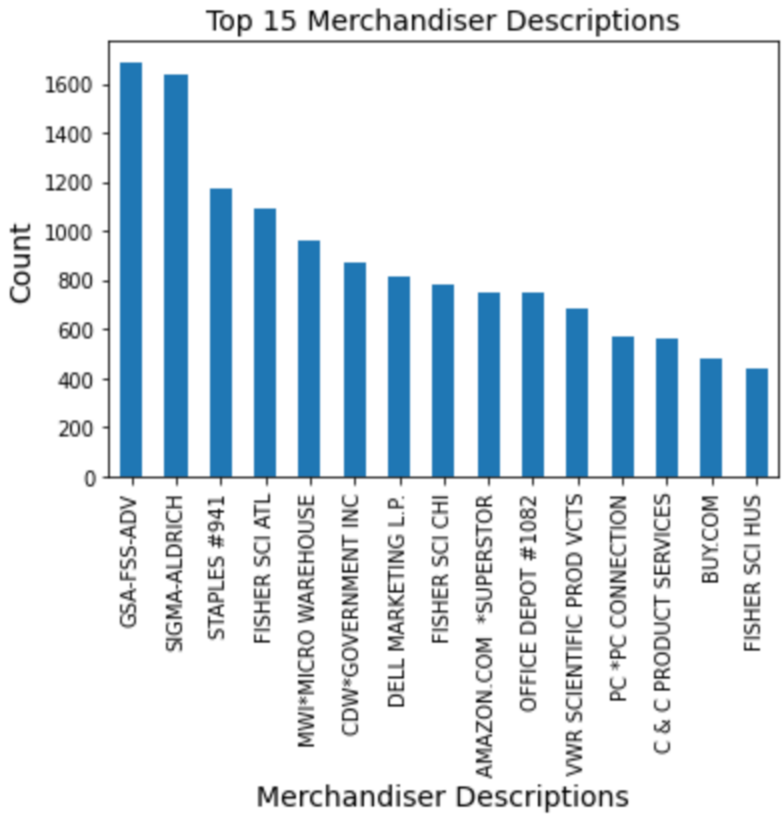
(5) Field Name: Merchnum
Description: The unique identifier of the merchandiser where the credit card transaction took place.
The most common merchandiser number is 930090121224 with a count of 9,310. This field also has a large number of N/A values at 3,375.



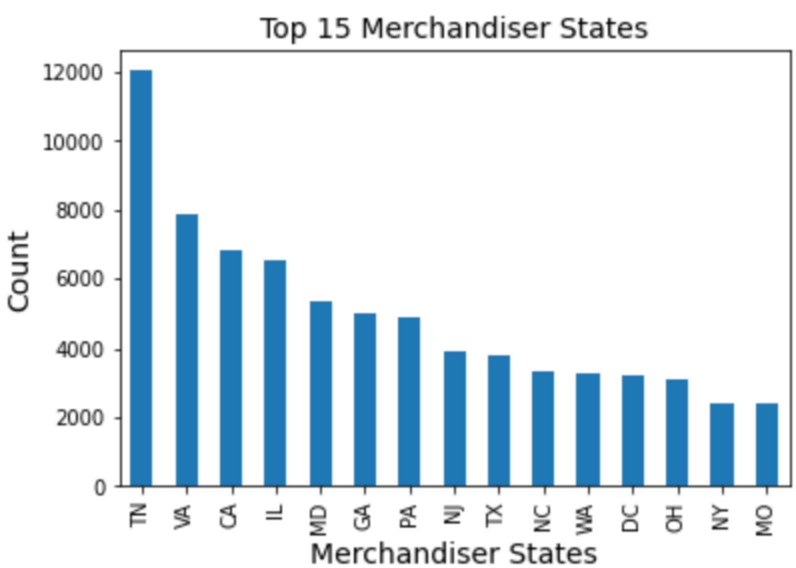
(6) Field Name: Merch Description

<u>Description:</u> A description of the merchandiser where the transaction took place. Values primarily include the name of the merchandiser, while some others include dates and codes.

The most common merchandiser is GSA-FSS-ADV with a count of 1,688.



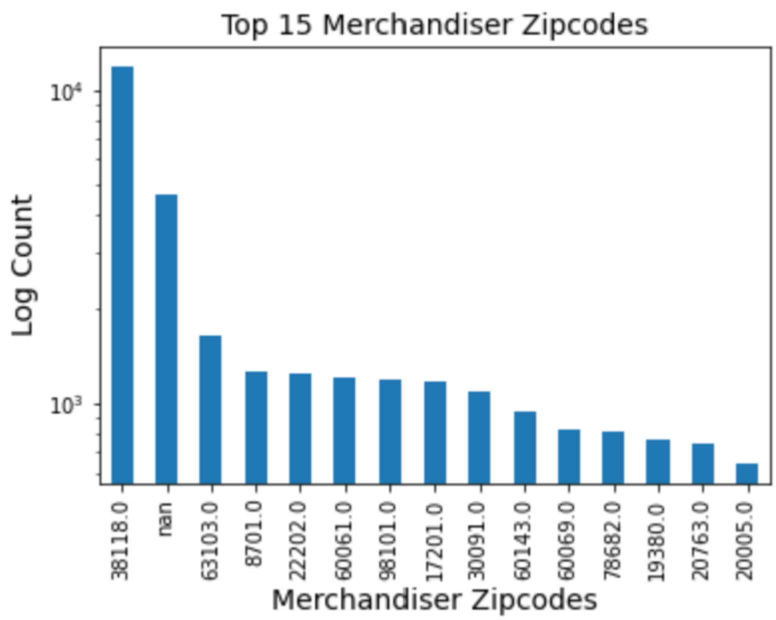
(7) Field Name: Merch State
<u>Description:</u> The state where each credit card transaction took place. The state with the most transactions is Tennessee with 12,305.



(8) Field Name: Merch Zip

Description: The zip code where each transaction took place

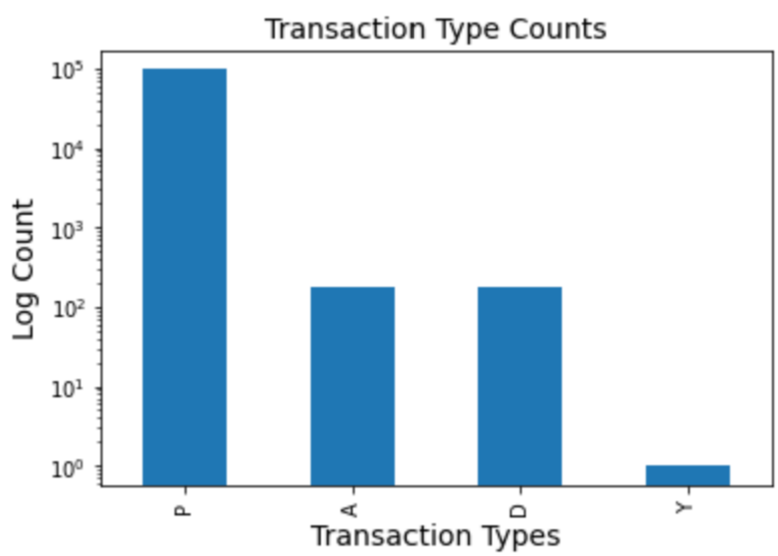
The most common zip code is 38118 (Tennessee) with a count of 11,868. This field also has a large number of N/A values at 4,656.



(9) Field Name: Transtype

Description: One of four transaction types: P, A, D or Y.

'P' transactions, which stand for Purchase, are the most common with a count of 96,398.



(10) Field Name: Fraud Label

Description: A binary classification, with Fraud=0 indicating no fraud detected, and Fraud=1 indicating fraud detection.
Fraud=0 (no fraud) is the most common with a count of 95,694.