Chaiyapuk Titinanapun https://www.linkedin.com/in/kk-titinanapun

Document Project 02

Executive summary

We utilized credit card transaction data from 2010 to create a credit card transaction fraud detection model that incorporated data from both credit card companies and merchants. Through out-of-time validation, we determined that the optimal threshold for fraud detection is a score cutoff of 4% of the total population. By adopting this threshold, we expect to achieve an estimated overall savings of $20,148,000, which considers the costs and benefits of fraud detection. Specifically, we anticipate a gain of $400 for every instance of fraud successfully detected, as well as a loss of $20 for each false positive.

Section 1: Description of data

This is transaction data containing credit card transactions in 2010. There are 10 fields with 96,753 transactions.

Summary Tables

1) Numerical Table

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

2) Categorical Table

Field Name	% Populated	# Unique Value	Most Common Value
Recnum	100.00%	96,753	N/A
Cardnum	100.00%	1,645	5142148452
Merchnum	96.51%	13,091	930090121224
Merch description	100.00%	13,126	GSA-FSS-ADV
Merch state	98.76%	227	TN
Merch zip	95.19%	4,567	38118
Transtype	100.00%	4	P
Fraud	100.00%	2	0

Section 2: Data cleaning

The data had been filtered and kept only Transtype == 'P' before missing value imputation.

There are three variables with missing values which are Merchnum, Merch state, and Merch zip.

- For Merchnum and Merch zip, replacing by the value from the first record with the mode of 'Merch_description'.
- For Merch state with zipcode, using the state of that zipcode if it is in the US or “Unknown” if it is outside of the US. With no zipcode, replacing by the value from the first record with the mode of 'Merch_description'.

Section 3: Variable creation

Credit card transaction fraud is the case when a credit card is used by someone else other than the owner. The supervised algorithm will look for characteristics of fraud cases based on variables that are created from account and merchant information. The algorithm will investigate if any card has been used in abnormal number of times or amounts.

The variable will be created with transaction information as follows:

Description of variables	# Variables created	# Variables after dedup
Day of the Week Risk Variable: Fraud case percentage for each day of the week (with Statistical Smoothing to overall average percentage)	1	1
Benford's Law Variables: The ratio between amounts which first digit is 3-9 and amounts which first digit is 1-2, normalize by 1.096 according to Benford's Law (plus Statistical Smoothing to 1).	2	2
Days Since Variables: # days since an application with that entity has been seen. Entities list is attached below	29	29
Frequency Variables: # records with the same entity over the last {0,1,3,7,14,30} days	174	164
Amount Variables: Aggregation of "transaction amounts" with the same entity over the last {0,1,3,7,14,30} days, including average, max, total, standard deviation, actual/avg, actual/max, and actual/total	1,218	1,154
Amount Difference Variables: Aggregation of "differences between recent transaction amount and previous amounts" with the same entity over the last {0,1,3,7,14,30} days, including average, min, max, total, and standard deviation	870	825
Relative Velocity Variables: The ratio between {# records, total amounts} with the same entity over the past {0,1} days and the average of {# records, total amounts} with the same entity over the past {3,7,14,30} days	464	448
Relative Velocity per Day Since Variables: The ratio between each Relative Velocity variables and Days Since variables with the same entity	232	232
Counts by Entities: For each of the same entity over the past {0,1,3,7,14,30,60} days, # unique values of other entities	5,684	937
Total # of variables	**8,674**	**3,792**

Entities list: ['Cardnum', 'Merchnum', 'Merch_description', 'Merch_num_des', 'Merch_num_state', 'Merch_num_zip', 'Merch_des_state', 'Merch_des_zip', 'Merch_state_zip', 'Merch_num_des_state', 'Merch_num_des_zip', 'Merch_num_state_zip', 'Merch_des_state_zip', 'Merch_all_info', 'card_Merchnum', 'card_Merch_description', 'card_Merch_state', 'card_Merch_zip', 'card_Merch_num_des', 'card_Merch_num_state', 'card_Merch_num_zip', 'card_Merch_des_state', 'card_Merch_des_zip', 'card_Merch_state_zip', 'card_Merch_num_des_state', 'card_Merch_num_des_zip', 'card_Merch_num_state_zip', 'card_Merch_des_state_zip', 'card_Merch_all_info']

Section 4: Feature selection

With roughly 4000 variables, the computational power and time necessary to explore models will be enormous. Feature selection will allow much faster nonlinear model runs to optimize model architecture and hyperparameters. The process can be done in two steps.

The first step is filtering variables independently based on their univariate model performance measure. Kolmogorov-Smirnov test for goodness of fit is used for this step. The result from the first step has 600 variables.

The second step is using a wrapper model with forward selection. The variables from the first step will be included into the model one by one. This will remove correlation between variables as highly correlated variables will not get selected. LightGBM with n_estimators = 30 and num_leaves = 4 is used for this step. The result from the second step has 25 variables.

The selected variables after the second step are as follows:

wrapper order	variable	filter score
1	card_Merch_num_state_total_14	0.676008511
2	card_Merchnum_max_30	0.648741321
3	card_Merch_des_zip_total_1	0.634249492
4	Merch_description_total_0	0.581051024
5	card_Merch_state_zip_total_14	0.670751528
6	card_Merch_des_state_max_30	0.654714819
7	card_Merch_des_zip_max_30	0.651810225
8	card_Merch_des_state_zip_max_30	0.651810225
9	card_Merch_num_des_zip_total_0	0.602714009
10	card_Merch_des_state_zip_total_30	0.652168078
11	card_Merch_num_state_zip_total_0	0.607052787
12	card_Merch_num_state_zip_total_30	0.655072672
13	card_Merch_des_zip_total_30	0.652155541
14	card_Merch_num_zip_total_0	0.60704025
15	card_Merch_des_zip_total_0	0.605781583
16	card_Merch_state_zip_total_0	0.61025694
17	card_Merch_num_state_total_30	0.659830082
18	Merch_state_zip_variability_avg_14	0.40467819
19	card_Merch_num_des_total_0	0.60717186
20	card_Merch_des_state_total_30	0.657050858
21	card_Merch_des_state_total_0	0.610189287
22	card_Merch_description_total_0	0.610164213
23	card_Merch_zip_total_0	0.609993663
24	card_Merch_num_des_zip_total_14	0.65931699
25	card_Merch_description_total_30	0.656837729

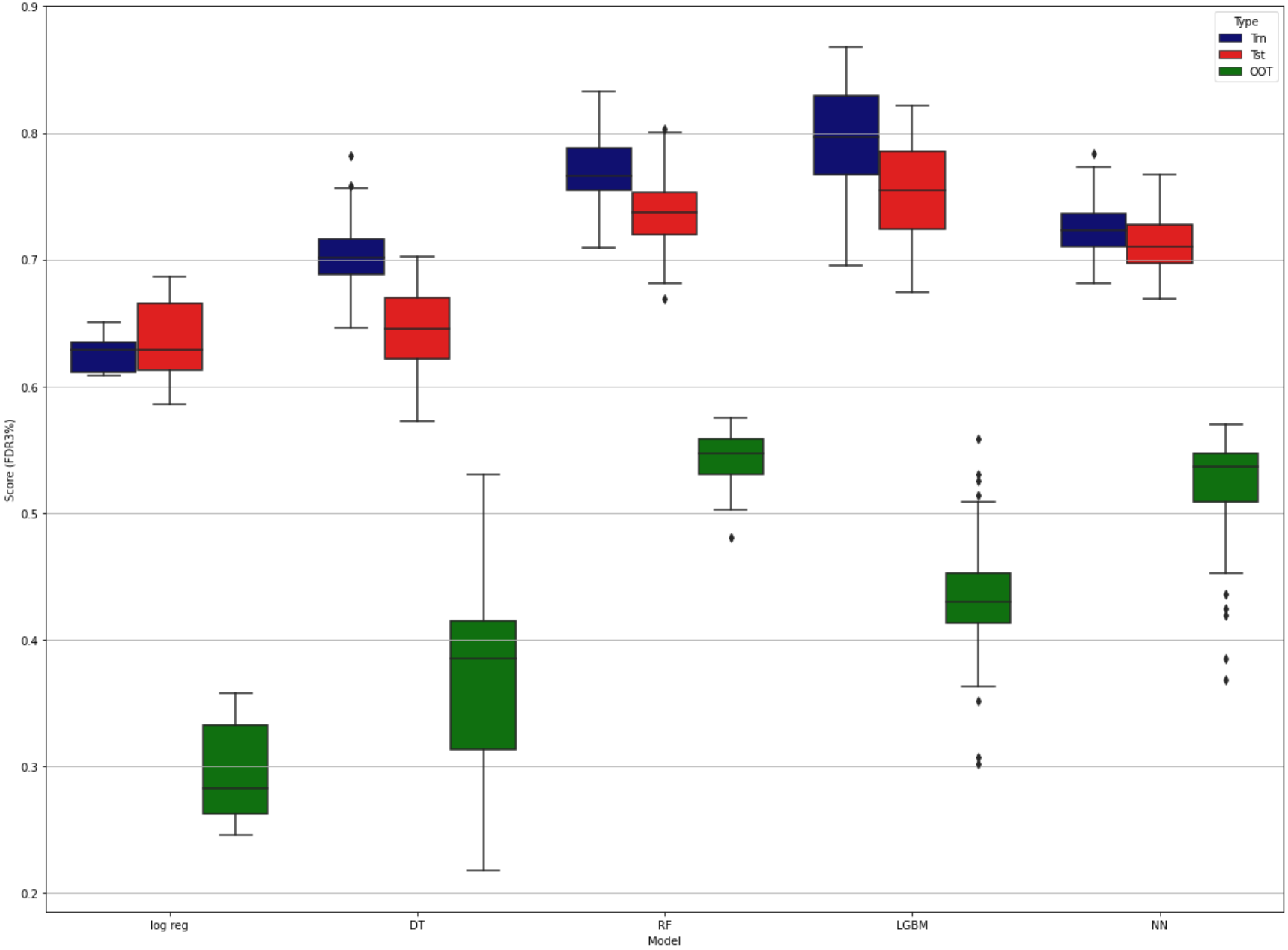
Section 5: Preliminary models exploration

Variables from feature selection process will be used to create several supervised models. Logistic regression models are used as baseline for comparison. Non-linear models are expected to perform better than logistic regression models. Each model will be run 5 times and will be selected based on their average fraud detection rate at 3% population with test dataset.

The result has been provided as follows:

Model		Num of Variables	Parameters				Average FDR at 3%		
Logistic Regression	Iteration	Num of Variables	penalty				Train	Test	OOT
	1	10	N/A				0.626	0.636	0.295
	2	10							
Decision Tree	Iteration	Num of Variables	min_impurity_decrease		min_sample_split	min_samples_leaf	Train	Test	OOT
	1	10	0.00003		30	15	0.700	0.664	0.451
	2	10	0.00003		60	30	0.692	0.688	0.436
	3	10	0.00003		120	60	0.691	0.664	0.456
	4	10	0.00002		30	15	0.711	0.648	0.390
Random Forest	Iteration	Num of Variables	min_impurity_decrease	min_sample_split	min_samples_leaf	n_estimators	Train	Test	OOT
	2	10	0.0001	2	1	20	0.693	0.687	0.522
	3	10	0.00001	60	30	20	0.779	0.760	0.549
	4	10	0.00002	60	30	20	0.753	0.736	0.540
	5	10	0.000005	60	30	20	0.794	0.759	0.530
LightGBM	Iteration	Num of Variables	min_split_gain	max_depth	min_child_samples	num_leaves	Train	Test	OOT
	1	10	1	2	100	20	0.766	0.747	0.511
	2	10	0.01	2	100	20	0.768	0.751	0.541
	3	10	1	5	1000	20	0.803	0.763	0.447
	4	10	1	10	1000	20	0.816	0.784	0.437
Neural Network	Iteration	Num of Variables	hidden_layer_sizes		alpha	learning_rate_init	Train	Test	OOT
	1	10	(10, 10)		0.01	0.01	0.710	0.710	0.528
	2	10	(15, 15)		0.01	0.01	0.720	0.702	0.553
	3	10	(20, 20)		0.01	0.01	0.748	0.724	0.501
	4	10	(25, 25)		0.01	0.01	0.755	0.734	0.479
	5	10	(30, 30)		0.01	0.01	0.764	0.733	0.478

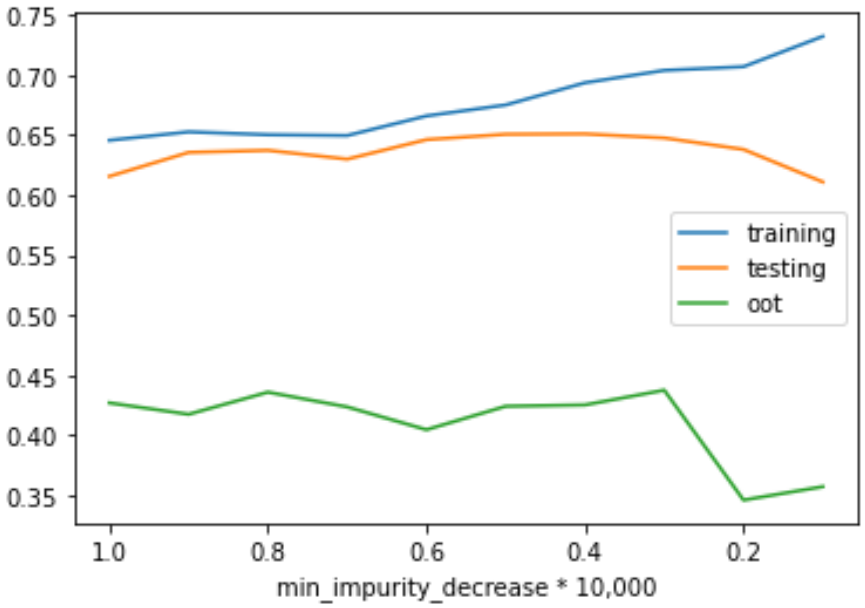
Chaiyapuk Titinanapun https://www.linkedin.com/in/kk-titinanapun



DecisionTreeClassifier

Complexity: min_impurity_decrease from 0.0001 to 0.00001

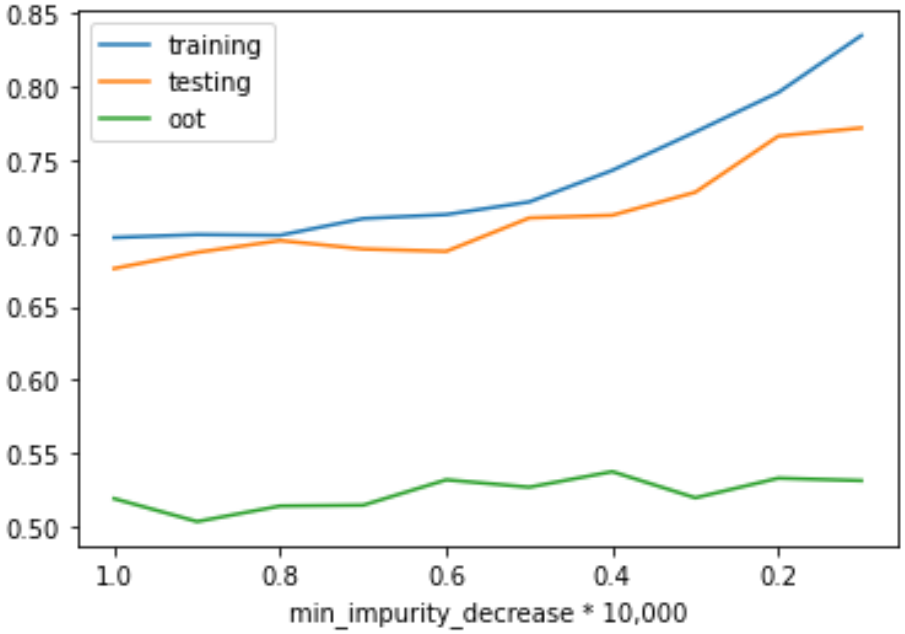
Hyperparameters: max_depth=None, min_samples_split=2, min_samples_leaf=1



RandomForestClassifier

Complexity: min_impurity_decrease from 0.0001 to 0.00001

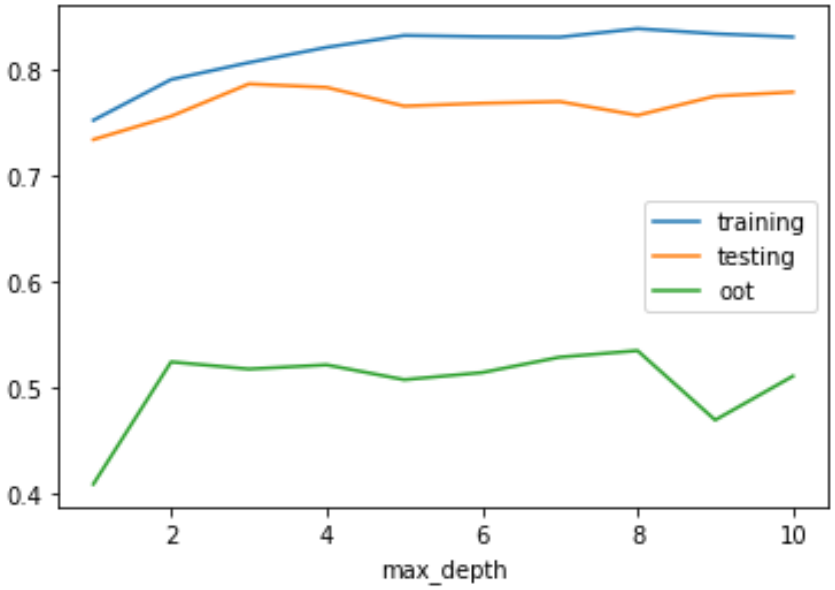
Hyperparameters: max_depth=None, min_samples_split=2, min_samples_leaf=1



LGBMClassifier

Complexity: max_depth from 1 to 10

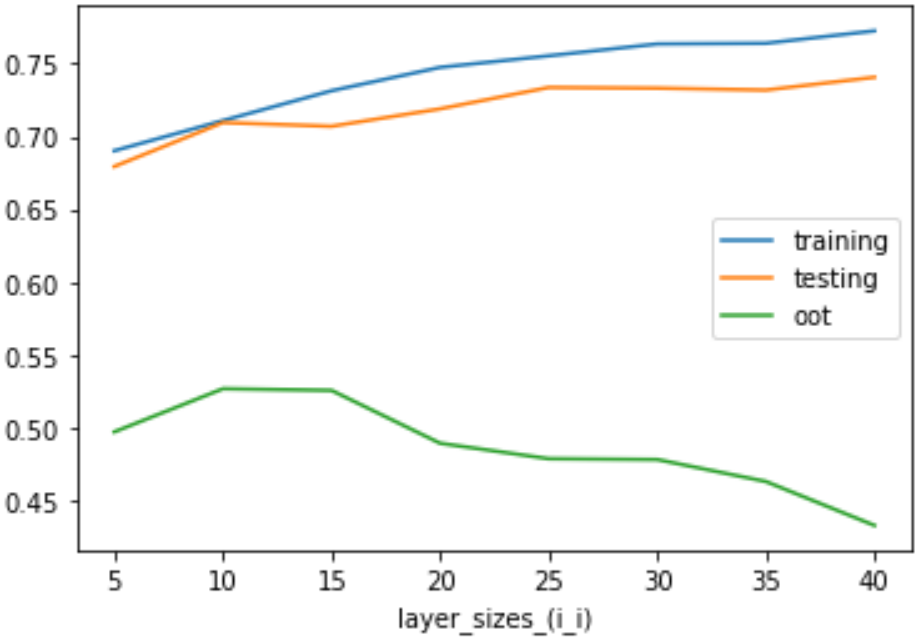
Hyperparameters: min_split_gain = 1, min_child_samples = 100, n_estimators = 1000, num_leaves = 10



Neural Network

Complexity: layer_sizes from (5, 5) to (40, 40)

Hyperparameters: alpha=0.0001, learning_rate_init=0.001, learning_rate='constant'



Section 6: Summary of results

RandomForestClassifier (n_estimators = 20, min_impurity_decrease = 0.00001, min_samples_leaf = 30, min_samples_split = 60) with 10 variables has been chosen as the final model. The result of the final model includes fraud detection rate for each population bin as the measure of model performance. The result is provided as follows:

Training	# Records	# Goods	# Bads	Fraud Rate
	59,009	58,397	612	1.037%

	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Cumulative Goods	FDR (% Cumulative Bad)	KS	FPR
1	590	244	346	41.36%	58.64%	590	244	346	0.42%	56.54%	56.12%	0.7052
2	590	488	102	82.71%	17.29%	1,180	732	448	1.25%	73.20%	71.95%	1.6339
3	590	563	27	95.42%	4.58%	1,770	1,295	475	2.22%	77.61%	75.40%	2.7263
4	590	583	7	98.81%	1.19%	2,360	1,878	482	3.22%	78.76%	75.54%	3.8963
5	590	577	13	97.80%	2.20%	2,950	2,455	495	4.20%	80.88%	76.68%	4.9596
6	591	588	3	99.49%	0.51%	3,541	3,043	498	5.21%	81.37%	76.16%	6.1104
7	590	587	3	99.49%	0.51%	4,131	3,630	501	6.22%	81.86%	75.65%	7.2455
8	590	583	7	98.81%	1.19%	4,721	4,213	508	7.21%	83.01%	75.79%	8.2933
9	590	585	5	99.15%	0.85%	5,311	4,798	513	8.22%	83.82%	75.61%	9.3528
10	590	587	3	99.49%	0.51%	5,901	5,385	516	9.22%	84.31%	75.09%	10.4360
11	590	583	7	98.81%	1.19%	6,491	5,968	523	10.22%	85.46%	75.24%	11.4111
12	590	587	3	99.49%	0.51%	7,081	6,555	526	11.22%	85.95%	74.72%	12.4620
13	590	585	5	99.15%	0.85%	7,671	7,140	531	12.23%	86.76%	74.54%	13.4463
14	590	585	5	99.15%	0.85%	8,261	7,725	536	13.23%	87.58%	74.35%	14.4123
15	590	585	5	99.15%	0.85%	8,851	8,310	541	14.23%	88.40%	74.17%	15.3604
16	590	588	2	99.66%	0.34%	9,441	8,898	543	15.24%	88.73%	73.49%	16.3867
17	591	589	2	99.66%	0.34%	10,032	9,487	545	16.25%	89.05%	72.81%	17.4073
18	590	589	1	99.83%	0.17%	10,622	10,076	546	17.25%	89.22%	71.96%	18.4542
19	590	587	3	99.49%	0.51%	11,212	10,663	549	18.26%	89.71%	71.45%	19.4226
20	590	589	1	99.83%	0.17%	11,802	11,252	550	19.27%	89.87%	70.60%	20.4582

Testing	# Records	# Goods	# Bads	Fraud Rate
	25,290	25,022	268	1.060%

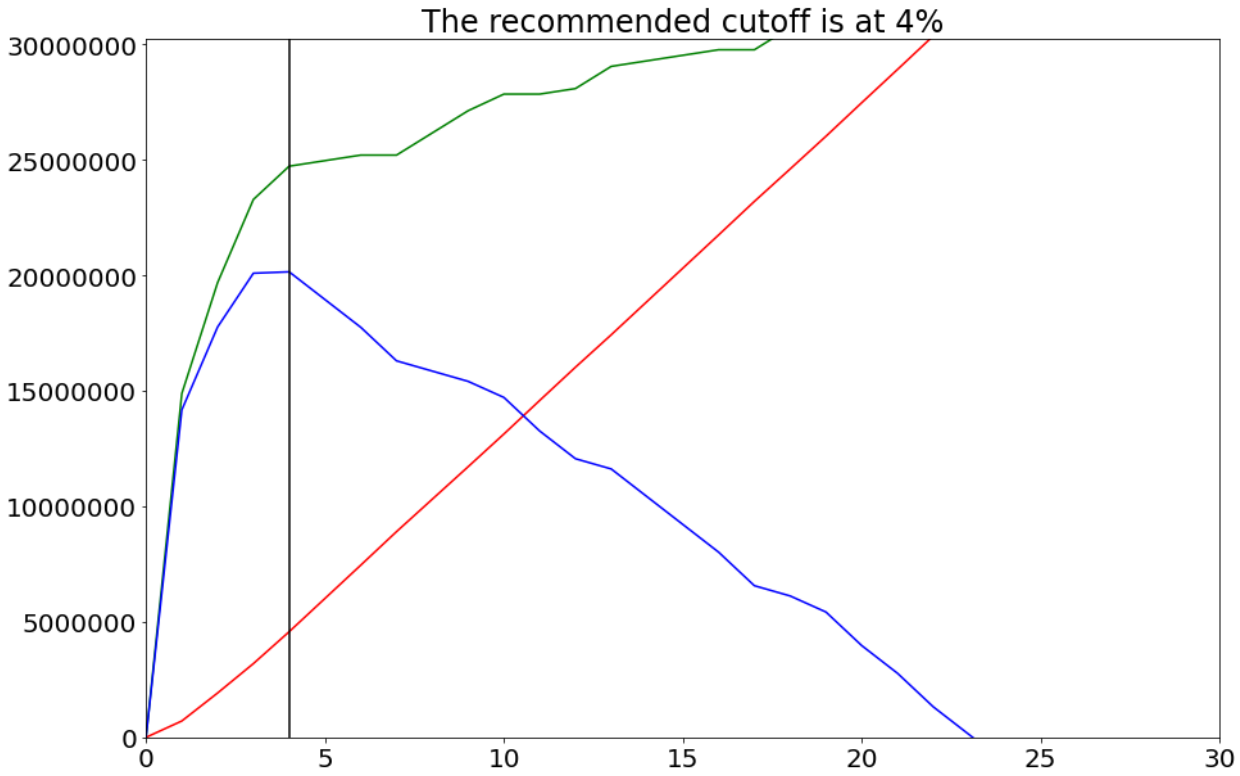
	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Cumulative Goods	FDR (% Cumulative Bad)	KS	FPR
1	253	99	154	39.13%	60.87%	253	99	154	0.40%	57.46%	57.07%	0.6429
2	253	217	36	85.77%	14.23%	506	316	190	1.26%	70.90%	69.63%	1.6632
3	253	237	16	93.68%	6.32%	759	553	206	2.21%	76.87%	74.66%	2.6845
4	253	249	4	98.42%	1.58%	1,012	802	210	3.21%	78.36%	75.15%	3.8190
5	252	244	8	96.83%	3.17%	1,264	1,046	218	4.18%	81.34%	77.16%	4.7982
6	253	251	2	99.21%	0.79%	1,517	1,297	220	5.18%	82.09%	76.91%	5.8955
7	253	251	2	99.21%	0.79%	1,770	1,548	222	6.19%	82.84%	76.65%	6.9730
8	253	253	-	100.00%	0.00%	2,023	1,801	222	7.20%	82.84%	75.64%	8.1126
9	253	253	-	100.00%	0.00%	2,276	2,054	222	8.21%	82.84%	74.63%	9.2523
10	253	252	1	99.60%	0.40%	2,529	2,306	223	9.22%	83.21%	73.99%	10.3408
11	253	251	2	99.21%	0.79%	2,782	2,557	225	10.22%	83.96%	73.74%	11.3644
12	253	251	2	99.21%	0.79%	3,035	2,808	227	11.22%	84.70%	73.48%	12.3700
13	253	251	2	99.21%	0.79%	3,288	3,059	229	12.23%	85.45%	73.22%	13.3581
14	253	251	2	99.21%	0.79%	3,541	3,310	231	13.23%	86.19%	72.97%	14.3290
15	253	251	2	99.21%	0.79%	3,794	3,561	233	14.23%	86.94%	72.71%	15.2833
16	252	251	1	99.60%	0.40%	4,046	3,812	234	15.23%	87.31%	72.08%	16.2906
17	253	249	4	98.42%	1.58%	4,299	4,061	238	16.23%	88.81%	72.58%	17.0630
18	253	253	-	100.00%	0.00%	4,552	4,314	238	17.24%	88.81%	71.57%	18.1261
19	253	253	-	100.00%	0.00%	4,805	4,567	238	18.25%	88.81%	70.55%	19.1891
20	253	253	-	100.00%	0.00%	5,058	4,820	238	19.26%	88.81%	69.54%	20.2521

Out of Time	# Records	# Goods	# Bads	Fraud Rate
	12,098	11,919	179	1.480%

	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Cumulative Goods	FDR (% Cumulative Bad)	KS	FPR
1	121	59	62	48.76%	51.24%	121	59	62	0.50%	34.64%	34.14%	0.9516
2	121	101	20	83.47%	16.53%	242	160	82	1.34%	45.81%	44.47%	1.9512
3	121	106	15	87.60%	12.40%	363	266	97	2.23%	54.19%	51.96%	2.7423
4	121	115	6	95.04%	4.96%	484	381	103	3.20%	57.54%	54.35%	3.6990
5	121	120	1	99.17%	0.83%	605	501	104	4.20%	58.10%	53.90%	4.8173
6	121	120	1	99.17%	0.83%	726	621	105	5.21%	58.66%	53.45%	5.9143
7	121	121	-	100.00%	0.00%	847	742	105	6.23%	58.66%	52.43%	7.0667
8	121	117	4	96.69%	3.31%	968	859	109	7.21%	60.89%	53.69%	7.8807
9	121	117	4	96.69%	3.31%	1,089	976	113	8.19%	63.13%	54.94%	8.6372
10	121	118	3	97.52%	2.48%	1,210	1,094	116	9.18%	64.80%	55.63%	9.4310
11	121	121	-	100.00%	0.00%	1,331	1,215	116	10.19%	64.80%	54.61%	10.4741
12	121	120	1	99.17%	0.83%	1,452	1,335	117	11.20%	65.36%	54.16%	11.4103
13	121	117	4	96.69%	3.31%	1,573	1,452	121	12.18%	67.60%	55.42%	12.0000
14	121	120	1	99.17%	0.83%	1,694	1,572	122	13.19%	68.16%	54.97%	12.8852
15	121	120	1	99.17%	0.83%	1,815	1,692	123	14.20%	68.72%	54.52%	13.7561
16	121	120	1	99.17%	0.83%	1,936	1,812	124	15.20%	69.27%	54.07%	14.6129
17	121	121	-	100.00%	0.00%	2,057	1,933	124	16.22%	69.27%	53.06%	15.5887
18	121	117	4	96.69%	3.31%	2,178	2,050	128	17.20%	71.51%	54.31%	16.0156
19	121	118	3	97.52%	2.48%	2,299	2,168	131	18.19%	73.18%	54.99%	16.5496
20	121	121	-	100.00%	0.00%	2,420	2,289	131	19.20%	73.18%	53.98%	17.4733

Section 7: Recommended cutoff

Based on out-of-time validation, we recommend an optimal score cutoff of 4% for fraud detection, which yields an expected overall savings of $20,148,000. This estimation is generated by considering the costs and benefits of detecting fraud, assuming a gain of $400 for every fraud caught and a loss of $20 for every false positive.



Section 8: Summary

We utilized credit card transaction data from 2010 to create a credit card transaction fraud detection model that incorporated data from both credit card companies and merchants. Through out-of-time validation, we determined that the optimal threshold for fraud detection is a score cutoff of 4% of the total population. By adopting this threshold, we expect to achieve an estimated overall savings of $20,148,000, which considers the costs and benefits of fraud detection. Specifically, we anticipate a gain of $400 for every instance of fraud successfully detected, as well as a loss of $20 for each false positive.

Chaiyapuk Titinanapun https://www.linkedin.com/in/kk-titinanapun

Appendix: Data Quality Report

1. Data Description

This is application data containing credit card transactions in 2010. There are 10 fields with 96,753 transactions.

2. Summary Tables

1) Numerical Table

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

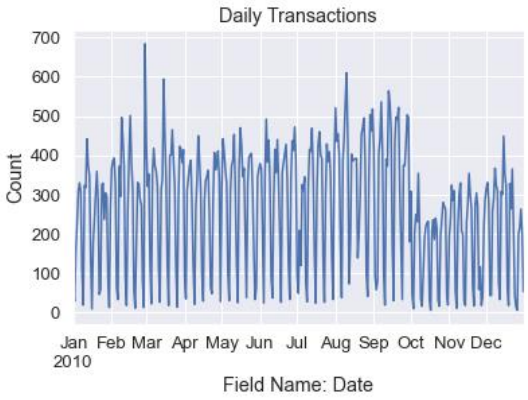
2) Categorical Table

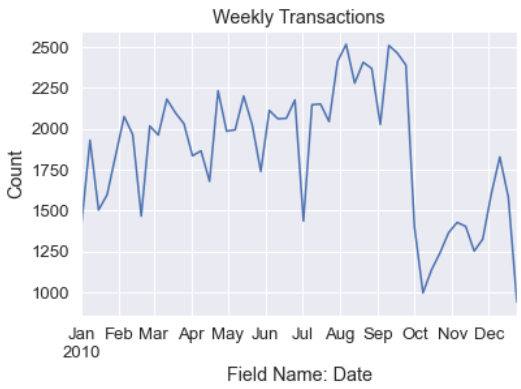
Field Name	% Populated	# Unique Value	Most Common Value
Recnum	100.00%	96,753	N/A
Cardnum	100.00%	1,645	5142148452
Merchnum	96.51%	13,091	930090121224
Merch description	100.00%	13,126	GSA-FSS-ADV
Merch state	98.76%	227	TN
Merch zip	95.19%	4,567	38118
Transtype	100.00%	4	P
Fraud	100.00%	2	0

3. Data Visualizations

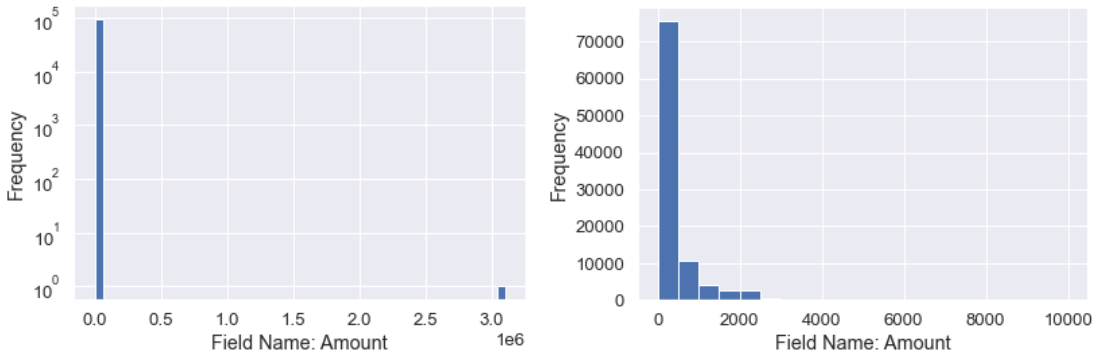
1) Field Name: Recnum
Description: Record Number. Ordinal unique positive integer for each record, from 1 to 96,753. There are no duplicate records.

2) Field Name: Date
Description: Date of transaction. Daily and weekly applications distribution across time are provided.

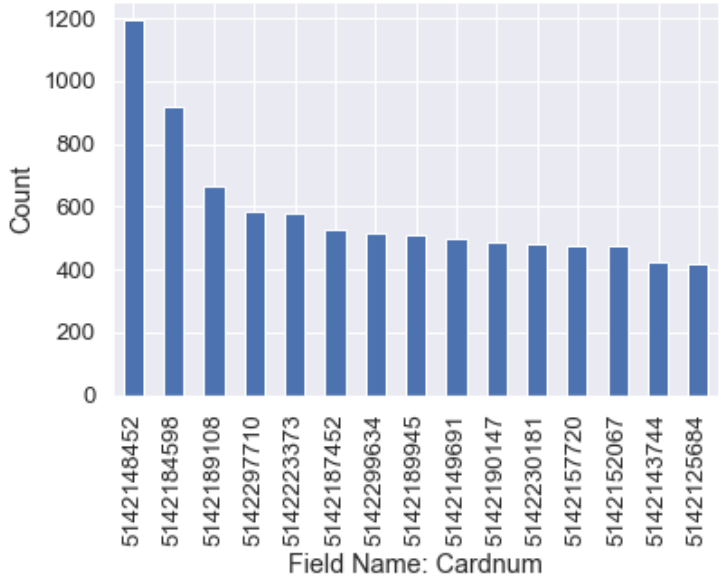




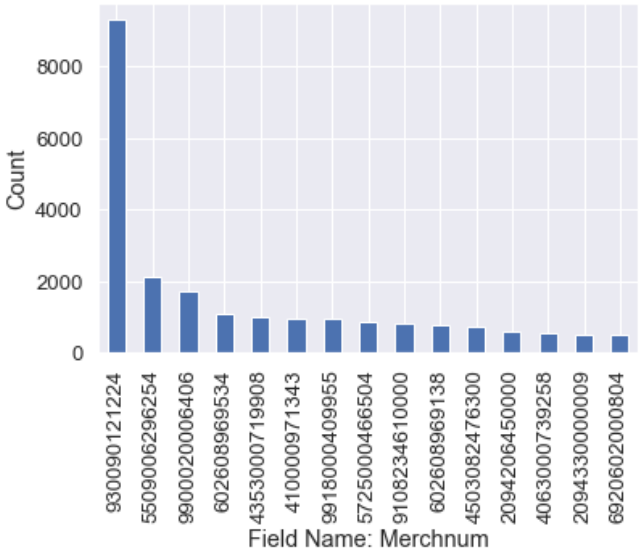
3) Field Name: Amount
Description: Transaction amount. As there are some outliers, the histogram for less than 10,000 is also provided.



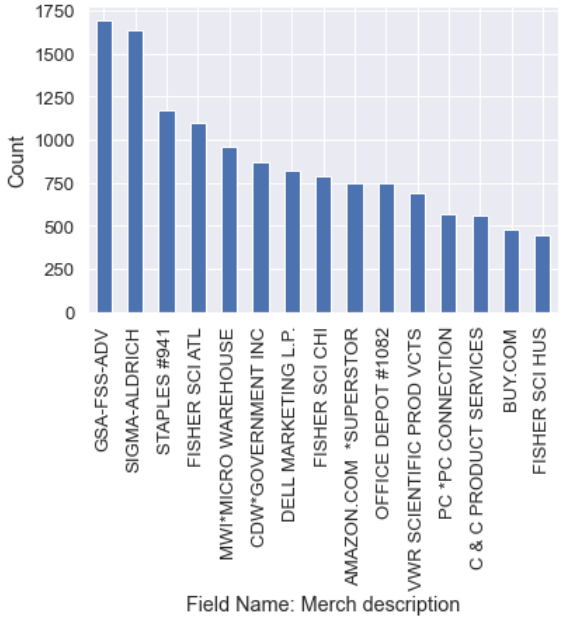
4) Field Name: Cardnum
Description: Credit Card Number. The distribution of the top 15 field values is provided. The most common value is 5142148452, accounting for 1.23% of all records.



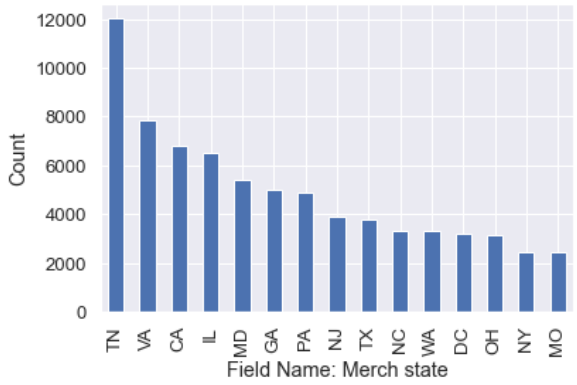
5) Field Name: Merchnum
Description: Merchant Number. The distribution of the top 15 field values is provided. The most common value is 930090121224, accounting for 9.62% of all records.



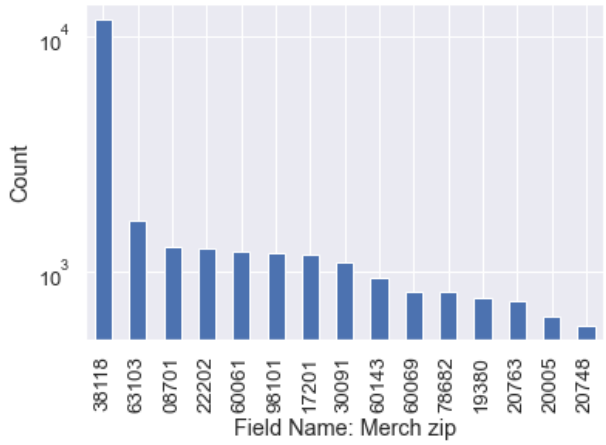
6) Field Name: Merch description
Description: Description of merchants. The distribution of the top 15 field values is provided. The most common value is GSA-FSS-ADV, accounting for 1.74% of all records.



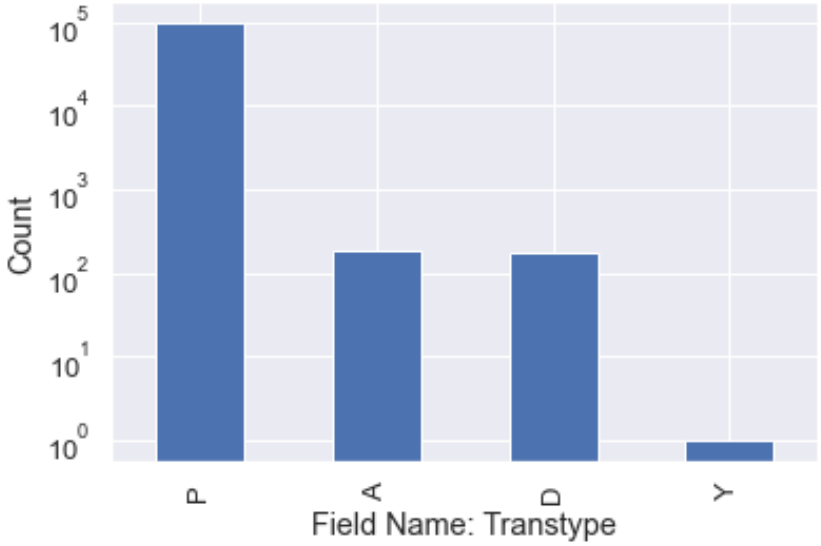
7) Field Name: Merch state
Description: Merchants' state. The distribution of the top 15 field values is provided. The most common value is TN, accounting for 12.44% of all records.



8) Field Name: Merch zip
Description: Merchants' zipcode. The distribution of the top 15 field values is provided. The most common value is 38118, accounting for 12.27% of all records.



9) Field Name: Transtype
Description: Transaction type. The distribution is provided. The most common value is P, accounting for 99.63% of all records.



10) Field Name: Fraud
Description: Binary field for Fraud Applications. 0 means not a fraud and 1 is a fraud. There are 1.09% of fraud accounts.