Fraud Analytics in Credit Card Transactions

1. Executive Summary

After analyzing 96,753 records of **Credit Card Transactions with fraud labels**, we built a **Random Forest Model with 10 Variables** to detect credit card transaction frauds. Our final model successfully achieved **55.31% FDR** at **3% population** for **OOT** Data, meaning that our final model can eliminate about **55.31%** of frauds by declining only about **3%** of the transactions without any overfitting or underfitting. We anticipate an overall savings of **$20,592,000** per year by using our final model.

2. Data Observation

- **Overview of Data**
  The data is a collection of **real Credit Card Transactions for business purposes from a US government organization**. The data including **1,059** fraud labels invented is to build models that can detect credit card transactions fraud. The data covers the time of **year 2010** with total **96,753 records** and **10 fields**.

- **Statistics Tables of Data**
  The followings are summary of statistics for numeric and categorical fields. We can observe that there are **null values** in **Merchnum, Merch State, Merch Zip** fields and an **outlier** with large transaction amount of $3,102,045.53. These values were fixed in data cleaning process.

  - **Numeric Table**

Field Name	% Populated	Min	Max	Mean	Stdev	% 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

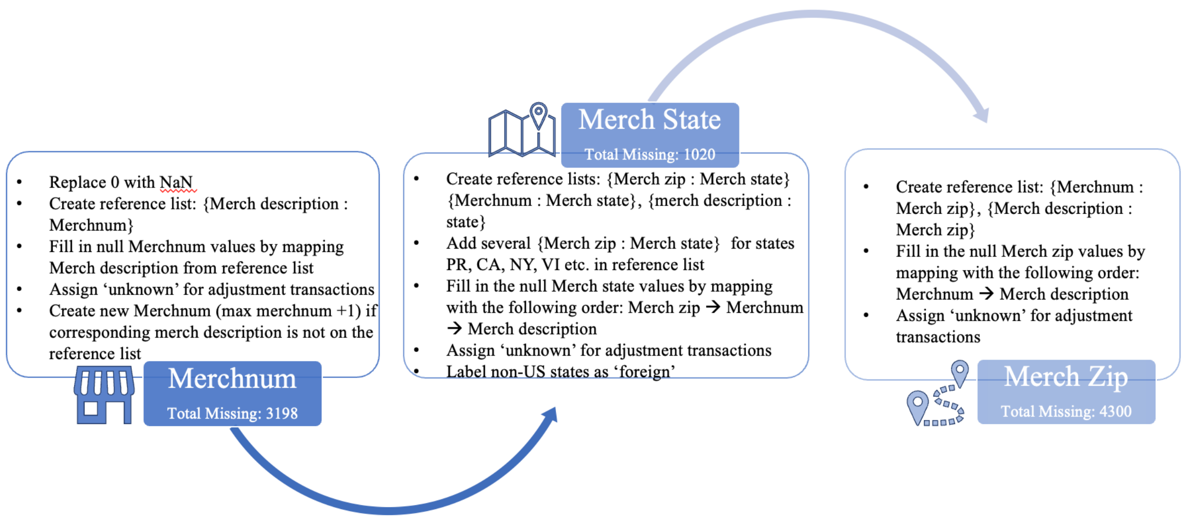
  - **Categorical Table**

Field Name	% Populated	# Unique Values	Most Common Field 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 Cleaning

- **Data Exclusions**
  Removed an **outlier** with large transaction amount of $3,102,045.53 and only kept transactions with **"P" transtype**.

- **Missing Field Values Imputation**
  Filled in missing values in **Merchnum, Merch State, and Merch Zip** fields. The imputation methods are described as below:



4. Variable Creation

- **Identity Fraud Modes/Motivation of Variables**
  Individual fraudsters steal others' credit cards or credit card information at specific location such as gas stations or online. Then they will use credit cards or credit card information to make purchases with high frequency in a short period of time, usually with a spending pattern from small to large amounts for each account at a specific merchant.

- **Variables**
  Motivated by the transactional fraud mode, we created 10 linking entities by combining original fields and four kinds of variables to check frequency, amount, and uniqueness of transactions: Days since, Amount, Velocity/Relative Velocity, and Counts by entities.

  - **Target Encoding**: In addition, we also had one target encoded variable that was converted from categorical date fields into numeric.

  - **Entities**: We created **10 entities** by linking original fields.
    ['Cardnum','Merchnum','card_merch','card_zip','card_state','merch_zip','card_zip3','Card_Merchedesc','Merchnum_desc','Card_Merchnum_desc']

- **Summary of Independent Variables**
  The following table shows a summary of variables. (total **1,424 Independent Variables**)

Family of Variables	Description of Variables	# Variables
Target Encoded Variable for day of week: 'Dow_Risk'	Average of the dependent variable 'Fraud' for all transactions in each day of week.	1
Days Since Variables	# days since the most recent transaction was seen with that specific entity.	10
Velocity Variables	# transactions with the same entity over the past {0,1,3,7,14,30,60} days.	70
Category - Amount Bins Variable	Category assigned by transaction amount based on the percentile 1st-5th	1
Amount Variables	Average, max, median, total, actual/average, actual/max, actual/median, actual/total, difference variance of amounts at the specific entity over the past {0, 1, 3, 7, 14, 30, 60} days.	899
Relative Velocity Variables	# transactions with that entity seen in the recent past {0,1} days over # transactions with that same entity seen in the past {7,14,30,60} days.	160
Counts by Entities Variables	# unique transactions with one entity that is linked to other entities over the past {1,3,7,14,30,60} days.	277
New Variables - for Online Transactions: "online_frequency"	For each Cardnum, the ratio of total number of online purchases in 30 days (current period) over average online purchases in 2010 (annual online purchases/12).	1
New Variables - For Gas Station Transactions: "gas_station_frequency"	For each Cardnum, the ratio of total number of gas station purchases in 30 days (current period) over average gas station purchases in 2010 (annual purchases/12).	1
New Variables - Amount Difference STD Ratio	For each Cardnum, the ratio of amount difference std in short period {1,7} over amount difference std in long period {30,60}	4
	Total Independent Variables	**1,424**
	Original Fields: Recnum and Fraud	2
	Total Variables (Including Recnum and Dependent Variable)	**1,426**

5. Feature Selection

- **Motivation**

After deduplication, we have **1,424** independent variables. Since dimensionality is high, data becomes sparse quickly and all points become outliers, causing a **curse of dimensionality**. Therefore, we implemented feature selection to reduce the number of independent variables.

- **Feature Selection Methods**

There were two steps for this feature selection: filter, and wrapper. In project 2, because OOT data, **data for the last two months in 2010**, shows strong impact of seasonality, we **included OOT data** in feature selection to take seasonality into consideration.

  - **Filter**: we used KS as the univariate measure to calculate correlations between each independent variable and Fraud. We sorted all independent variables by the KS-filter score in a descending order and chose the first **300**.

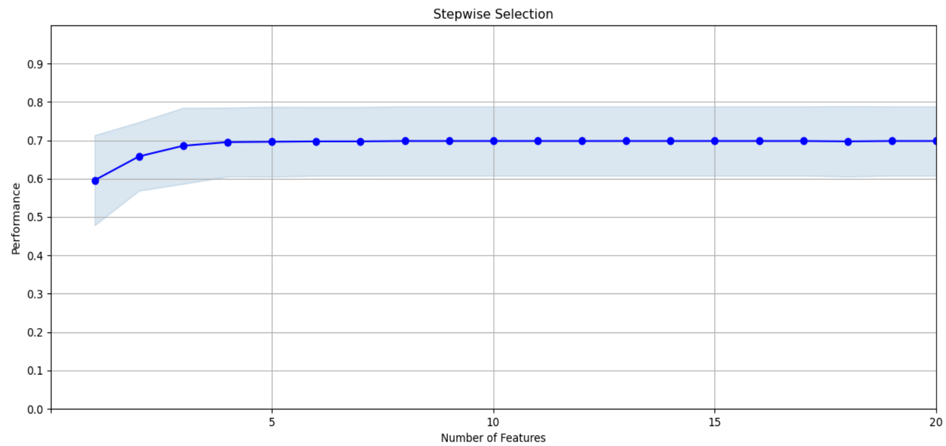
  - **Wrapper**: we used **Forward Selection** to build **LGBM** models (n_estimators=20, num_leaves=4, cv=4) by adding a variable until there was no significant improvement in the detection rate. We reduced the number of independent variables into **20**.

  - **List of Final Variables**

    The following is a list of final variables with **num_filter = 300 and num_wrapper = 20**

Wrapper Order	Variable	Filter Score
1	card_merch_total_14	0.630048056
2	card_zip3_max_14	0.629514577
3	Card_Merchdesc_count_7	0.367250198
4	Cardnum_avg_14	0.487201443
5	card_zip_max_0	0.543262985
6	card_merch_avg_0	0.512410575
7	Card_Merchnum_desc_max_0	0.533277329
8	card_zip3_med_3	0.498349452
9	Card_Merchnum_desc_avg_0	0.509146364
10	Card_Merchdesc_avg_0	0.50912471
11	card_merch_med_3	0.503946371
12	Card_Merchnum_desc_med_3	0.499138341
13	Card_Merchnum_desc_med_1	0.498892516
14	Card_Merchnum_desc_avg_1	0.511187128
15	Card_Merchdesc_med_3	0.498150751
16	Card_Merchdesc_med_0	0.49087379
17	Card_Merchnum_desc_med_0	0.490852136
18	Card_Merchdesc_avg_3	0.518653917
19	Card_Merchdesc_avg_1	0.514152093
20	card_merch_avg_3	0.52550209

- **Plot (300,20) LGBM Forward Selection**
  From the plot, we can see that the **saturation point is at 5** number of features with performance around **0.70**. To be conservative, we will **keep 10 variables** for modeling.



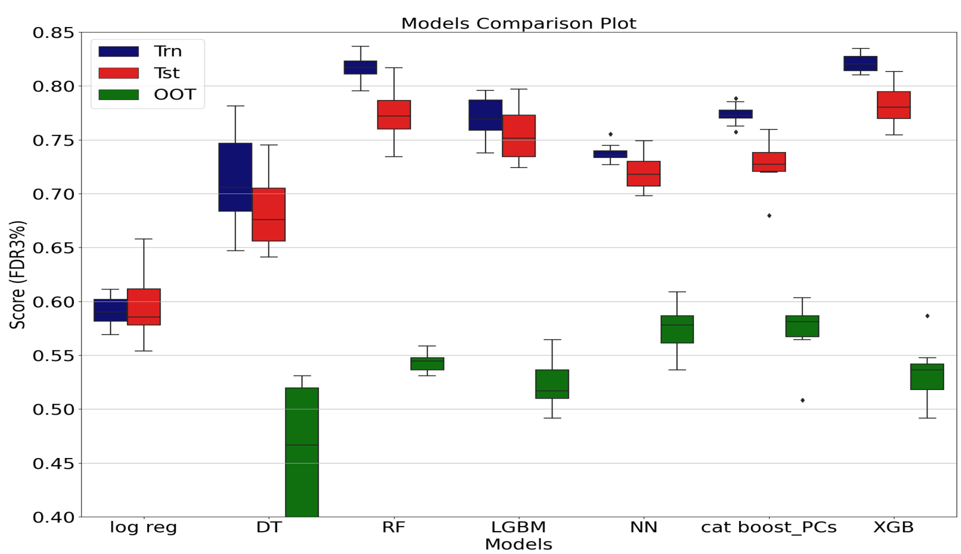
6. Preliminary Model Exploration

- **Hyperparameters Selection and Model Analysis**

We started from **a linear model** - logistic regression and tried **6 nonlinear models** with the **number of variables 10**. We firstly used the default hyperparameters and then tuned hyperparameters making the model overfitting. After overfitting, we lowered the complexity of the model (smaller depth or hidden layers) to find the best hyperparameters for each model.

- **Models Selection**

From the plot below, we can see that all nonlinear models perform better than logistic regression model. Within nonlinear models, we will choose **Random Forest Model** with higher mean of training and test, higher mean of oot, smaller variation of oot, and smaller diff between training and test.



○ Model Exploration Table

Models Exploration Table														
Models	Hyperparameters									Average FDR at 3%			Models Analysis	
Logistic Regression	Number of Variables	max_iter		solver		penalty		C		Train	Test	OOT	DIFF (trn-tst)	Performance
1	10	20		lbfgs		l2		1		0.598	0.589	0.378	0.009	
2	10	20		lbfgs		none		0.25		0.598	0.588	0.380	0.011	Best Model
3	10	20		lbfgs		l2		0.1		0.596	0.591	0.378	0.005	
Decision Tree	Number of Variables	splitter	max_depth	min_samples_split		min_samples_leaf		max_features		Train	Test	OOT	DIFF (trn-tst)	Performance
1	10	best	5	50		30		5		0.679	0.667	0.466	0.012	
2	10	best	10	10		10		10		0.926	0.743	0.322	0.184	Overfitting
3	10	best	20	25		20		10		0.905	0.750	0.333	0.155	Overfitting
4	10	best	30	50		50		10		0.834	0.761	0.379	0.073	
5	10	random	5	50		30		5		0.547	0.542	0.355	0.004	Underfitting
6	10	random	10	Formula Bar		10		10		0.723	0.680	0.453	0.043	Best Model
Random Forest	Number of Variables	n_estimators	max_depth	min_samples_split		min_samples_leaf		max_features		Train	Test	OOT	DIFF (trn-tst)	Performance
1	10	5	5	100		100		10		0.722	0.696	0.475	0.026	Underfitting
2	10	100	5	50		30		10		0.716	0.707	0.550	0.009	
3	10	15	30	10		10		5		0.984	0.790	0.459	0.194	Overfitting
4	10	15	20	20		15		10		0.942	0.788	0.398	0.154	Overfitting
5	10	10	15	30		30		10		0.855	0.791	0.463	0.064	
6	10	5	15	50		50		10		0.800	0.776	0.545	0.024	Best Model
Lightgbm	Number of Variables	n_estimators	max_depth	num_leaves	min_split_gain	reg_lambda	reg_alpha	learning_rate	subsample	Train	Test	OOT	DIFF (trn-tst)	Performance
1	10	100	2	256	0	0	0	0.1	1	0.775	0.745	0.513	0.030	
2	10	20	2	2	0	0	0	0.1	1	0.671	0.663	0.475	0.008	
3	10	100	30	256	0	0.3	0	0.25	1	0.998	0.754	0.328	0.244	Overfitting
4	10	100	5	10	0	0.3	0.5	0.25	1	0.901	0.764	0.374	0.137	Overfitting
5	10	100	2	10	0.5	0	0.5	0.25	1	0.689	0.637	0.409	0.051	Underfitting
6	10	100	2	100	0.5	0	0	0.1	1	0.775	0.772	0.529	0.003	Best Model
Catboost	Number of Variables	iterations	depth	random_strength		l2_leaf_reg		learning_rate		Train	Test	OOT	DIFF (trn-tst)	Performance
1	10	1000	2	0.5		1		0.1		0.825	0.791	0.462	0.034	
2	10	1000	8	0.5		1		0.1		0.991	0.780	0.351	0.211	Overfitting
3	10	5	10	0.5		5		0.1		0.668	0.670	0.387	-0.002	Underfitting
4	10	1000	2	1		1		0.03		0.711	0.699	0.550	0.011	
5	10	1000	2	1		5		0.1		0.776	0.711	0.574	0.066	Best Model
6	10	1000	2	1		5		0.03		0.711	0.687	0.559	0.024	
Xgboost	Number of Variables	max_depth	min_child_weight	subsample	reg_lambda	reg_alpha	gamma	learning_rate		Train	Test	OOT	DIFF (trn-tst)	Performance
1	10	2	1	1	1	0.5	0	0.3		0.835	0.797	0.498	0.038	
2	10	6	1	1	1	0.5	0	0.3		0.985	0.816	0.383	0.169	Overfitting
3	10	8	1	1	1	0.5	0	0.3		0.999	0.796	0.369	0.203	Overfitting
4	10	2	0.5	1	1	0	0	0.25		0.821	0.786	0.545	0.035	Best Model
5	10	2	0.5	1	1	0	0.25	0.3		0.836	0.793	0.506	0.043	
6	10	2	0.5	1	1	0	0	0.1		0.764	0.747	0.530	0.017	
7	10	2	1	1	1	0	0	0.1		0.765	0.746	0.532	0.019	
Neural Network	Number of Variables	lden_layer_si	activation	solver	learning_rate	alpha		learning_rate_init		Train	Test	OOT	DIFF (trn-tst)	Performance
1	10	(100,)	relu	lbfgs	constant	0.0001		0.001		0.746	0.704	0.573	0.042	
2	10	(5,)	logistic	lbfgs	constant	0.0001		0.001		0.709	0.697	0.517	0.012	
3	10	(10,10)	logistic	lbfgs	constant	0.1		0.001		0.724	0.713	0.434	0.011	
4	10	(20,20,20)	logistic	lbfgs	constant	0.1		0.001		0.519	0.500	0.301	0.019	Underfitting
5	10	(100,)	relu	lbfgs	adaptive	0.0001		0.001		0.741	0.718	0.585	0.023	Best Model
6	10	(10,10)	relu	lbfgs	adaptive	0.0001		0.001		0.738	0.699	0.546	0.039	
7	10	(20,20,20)	relu	lbfgs	adaptive	0.0001		0.001		0.771	0.730	0.454	0.042	A little overfitting

7. Final Model Performance

- **Final Model**

The followings are the details of our final model.

  - **Model Architecture**: **Random Forest**

  - **Model Hyperparameters**

N_Estimators	5
Max_Depth	15
Min_Samples_Split	50
Min_Samples_Leaf	50
Max_Features	10

  - **Final Variables:** Here is the list of **10 final independent variables**.

    ['card_merch_total_14','card_zip3_max_14','Card_Merchdesc_count_7','Cardnum_avg_14','card_zip_max_0','card_merch_avg_0','Card_Merchnum_desc_max_0','card_zip3_med_3','Card_Merchnum_desc_avg_0','Card_Merchdesc_avg_0']

- **Summary Columns**

  We train final model on training and evaluate the performance on testing and OOT population. The followings are explanation for each column in summary table.

Population Bin %	Percentage of population # records
# Records	Every 1% of population # records
# Goods	The increase in # goods with an increase of 1% of population records
# Bads	The increase in # bads with an increase of 1% of population records
% Goods	# Goods / # Records
% Bads	# Bads / # Records
Total # Records	Population bin % of population # records
Cumulative Goods	Total # goods in bin % of population
Cumulative Bads	Total # bads in bin % of population
% Cumulative Goods	Total # goods in bin % of population / Total # goods in 100 % of population
FDR (% Cumulative Bads)	Total # bads in bin % of population / Total # bads in 100 % of population
KS	% Cumulative Goods - % Cumulative Bads. It measures how well the goods and bads are separated.
FPR	Cumulative Goods / Cumulative Bads. It measures probability that we predict one record as bad that is actually good.

- ## Summary Tables

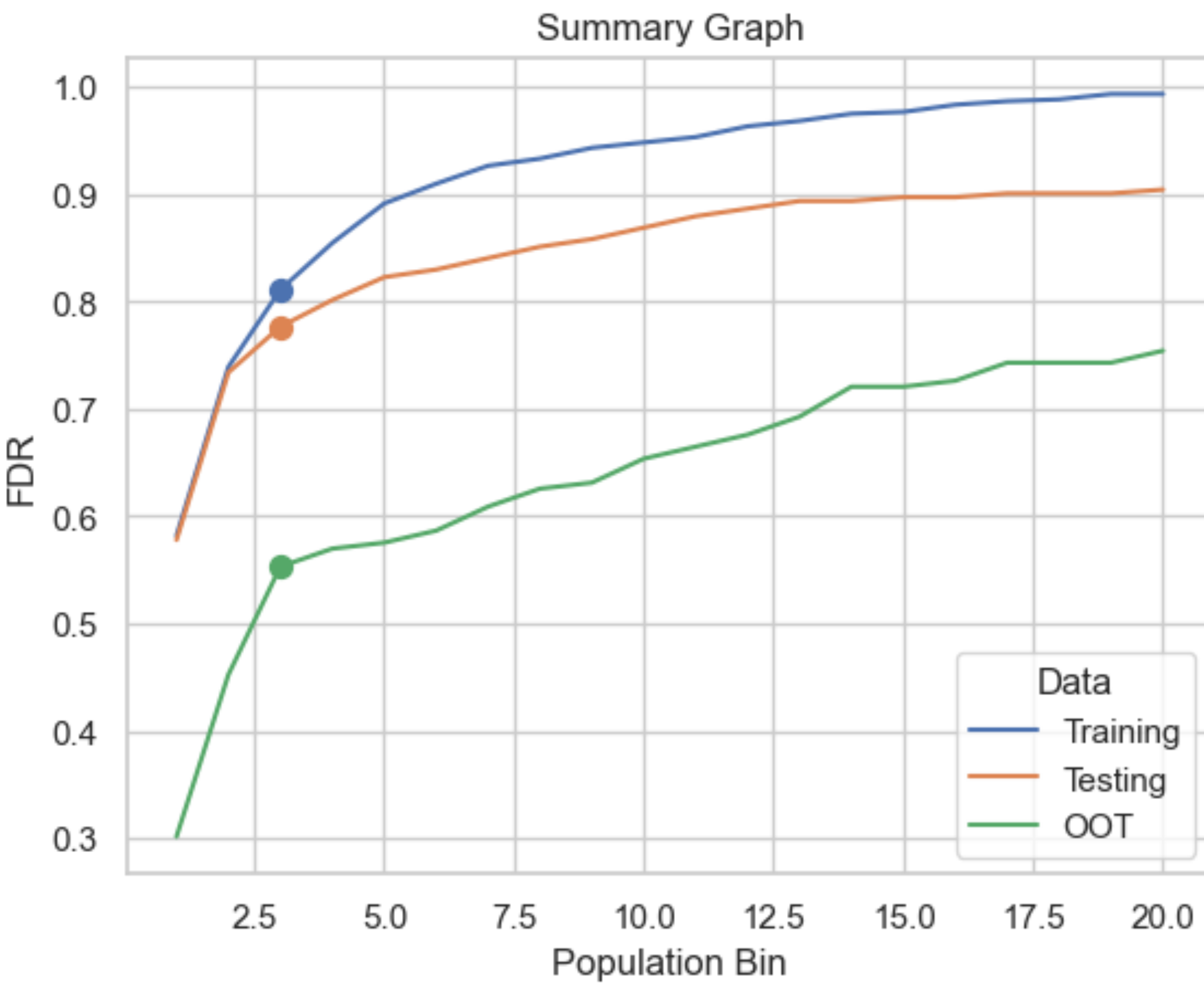
Training	Population Total # Records		Population Total # Goods		Population Total # Bads		Actual Fraud Rate					
	59,010		58,412		598		0.010237622					
Bin Statistics						Cumulative Statistics				Model Performance		
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Cumulative Goods	% Cumulative Bads (FDR)	KS	FPR
1	590	242	348	41.02%	58.98%	590	242	348	0.41%	58.19%	57.78%	0.70
2	590	496	94	84.07%	15.93%	1180	738	442	1.26%	73.91%	72.65%	1.67
3	590	547	43	92.71%	7.29%	1770	1285	485	2.20%	81.10%	78.90%	2.65
4	590	564	26	95.59%	4.41%	2360	1849	511	3.17%	85.45%	82.29%	3.62
5	590	568	22	96.27%	3.73%	2950	2417	533	4.14%	89.13%	84.99%	4.53
6	591	580	11	98.14%	1.86%	3541	2997	544	5.13%	90.97%	85.84%	5.51
7	590	580	10	98.31%	1.69%	4131	3577	554	6.12%	92.64%	86.52%	6.46
8	590	586	4	99.32%	0.68%	4721	4163	558	7.13%	93.31%	86.18%	7.46
9	590	584	6	98.98%	1.02%	5311	4747	564	8.13%	94.31%	86.19%	8.42
10	590	587	3	99.49%	0.51%	5901	5334	567	9.13%	94.82%	85.68%	9.41
11	590	587	3	99.49%	0.51%	6491	5921	570	10.14%	95.32%	85.18%	10.39
12	590	584	6	98.98%	1.02%	7081	6505	576	11.14%	96.32%	85.18%	11.29
13	590	587	3	99.49%	0.51%	7671	7092	579	12.14%	96.82%	84.68%	12.25
14	590	586	4	99.32%	0.68%	8261	7678	583	13.14%	97.49%	84.35%	13.17
15	591	590	1	99.83%	0.17%	8852	8268	584	14.15%	97.66%	83.50%	14.16
16	590	586	4	99.32%	0.68%	9442	8854	588	15.16%	98.33%	83.17%	15.06
17	590	588	2	99.66%	0.34%	10032	9442	590	16.16%	98.66%	82.50%	16.00
18	590	589	1	99.83%	0.17%	10622	10031	591	17.17%	98.83%	81.66%	16.97
19	590	587	3	99.49%	0.51%	11212	10618	594	18.18%	99.33%	81.15%	17.88
20	590	590	0	100.00%	0.00%	11802	11208	594	19.19%	99.33%	80.14%	18.87

Testing	Population Total # Records		Population Total # Goods		Population Total # Bads		Actual Fraud Rate					
	25,290		25,008		282		0.011276392					
Bin Statistics						Cumulative Statistics				Model Performance		
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Cumulative Goods	% Cumulative Bads (FDR)	KS	FPR
1	253	90	163	35.57%	64.43%	253	90	163	0.36%	57.80%	57.44%	0.55
2	253	209	44	82.61%	17.39%	506	299	207	1.20%	73.40%	72.21%	1.44
3	253	241	12	95.26%	4.74%	759	540	219	2.16%	77.66%	75.50%	2.47
4	253	246	7	97.23%	2.77%	1012	786	226	3.14%	80.14%	77.00%	3.48
5	252	246	6	97.62%	2.38%	1264	1032	232	4.13%	82.27%	78.14%	4.45
6	253	251	2	99.21%	0.79%	1517	1283	234	5.13%	82.98%	77.85%	5.48
7	253	250	3	98.81%	1.19%	1770	1533	237	6.13%	84.04%	77.91%	6.47
8	253	250	3	98.81%	1.19%	2023	1783	240	7.13%	85.11%	77.98%	7.43
9	253	251	2	99.21%	0.79%	2276	2034	242	8.13%	85.82%	77.68%	8.40
10	253	250	3	98.81%	1.19%	2529	2284	245	9.13%	86.88%	77.75%	9.32
11	253	250	3	98.81%	1.19%	2782	2534	248	10.13%	87.94%	77.81%	10.22
12	253	251	2	99.21%	0.79%	3035	2785	250	11.14%	88.65%	77.52%	11.14
13	253	251	2	99.21%	0.79%	3288	3036	252	12.14%	89.36%	77.22%	12.05
14	253	253	0	100.00%	0.00%	3541	3289	252	13.15%	89.36%	76.21%	13.05
15	253	252	1	99.60%	0.40%	3794	3541	253	14.16%	89.72%	75.56%	14.00
16	252	252	0	100.00%	0.00%	4046	3793	253	15.17%	89.72%	74.55%	14.99
17	253	252	1	99.60%	0.40%	4299	4045	254	16.17%	90.07%	73.90%	15.93
18	253	253	0	100.00%	0.00%	4552	4298	254	17.19%	90.07%	72.88%	16.92
19	253	253	0	100.00%	0.00%	4805	4551	254	18.20%	90.07%	71.87%	17.92
20	253	252	1	99.60%	0.40%	5058	4803	255	19.21%	90.43%	71.22%	18.84

OOT	Population Total # Records		Population Total # Goods		Population Total # Bads		Actual Fraud Rate					
	12,097		11,918		179		0.015019299					
Bin Statistics						Cumulative Statistics				Model Performance		
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Cumulative Goods	% Cumulative Bads (FDR)	KS	FPR
1	121	67	54	55.37%	44.63%	121	67	54	0.56%	30.17%	29.61%	1.24
2	121	94	27	77.69%	22.31%	242	161	81	1.35%	45.25%	43.90%	1.99
3	121	103	18	85.12%	14.88%	363	264	99	2.22%	55.31%	53.09%	2.67
4	121	118	3	97.52%	2.48%	484	382	102	3.21%	56.98%	53.78%	3.75
5	121	120	1	99.17%	0.83%	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	117	4	96.69%	3.31%	847	738	109	6.19%	60.89%	54.70%	6.77
8	121	118	3	97.52%	2.48%	968	856	112	7.18%	62.57%	55.39%	7.64
9	121	120	1	99.17%	0.83%	1089	976	113	8.19%	63.13%	54.94%	8.64
10	121	117	4	96.69%	3.31%	1210	1093	117	9.17%	65.36%	56.19%	9.34
11	121	119	2	98.35%	1.65%	1331	1212	119	10.17%	66.48%	56.31%	10.18
12	121	119	2	98.35%	1.65%	1452	1331	121	11.17%	67.60%	56.43%	11.00
13	121	118	3	97.52%	2.48%	1573	1449	124	12.16%	69.27%	57.12%	11.69
14	121	116	5	95.87%	4.13%	1694	1565	129	13.13%	72.07%	58.94%	12.13
15	121	121	0	100.00%	0.00%	1815	1686	129	14.15%	72.07%	57.92%	13.07
16	121	120	1	99.17%	0.83%	1936	1806	130	15.15%	72.63%	57.47%	13.89
17	120	117	3	97.50%	2.50%	2056	1923	133	16.14%	74.30%	58.17%	14.46
18	121	121	0	100.00%	0.00%	2177	2044	133	17.15%	74.30%	57.15%	15.37
19	121	121	0	100.00%	0.00%	2298	2165	133	18.17%	74.30%	56.14%	16.28
20	121	119	2	98.35%	1.65%	2419	2284	135	19.16%	75.42%	56.25%	16.92

- **Summary Graph**

The following is the plot for training, testing, and OOT FDR as population bin increases.



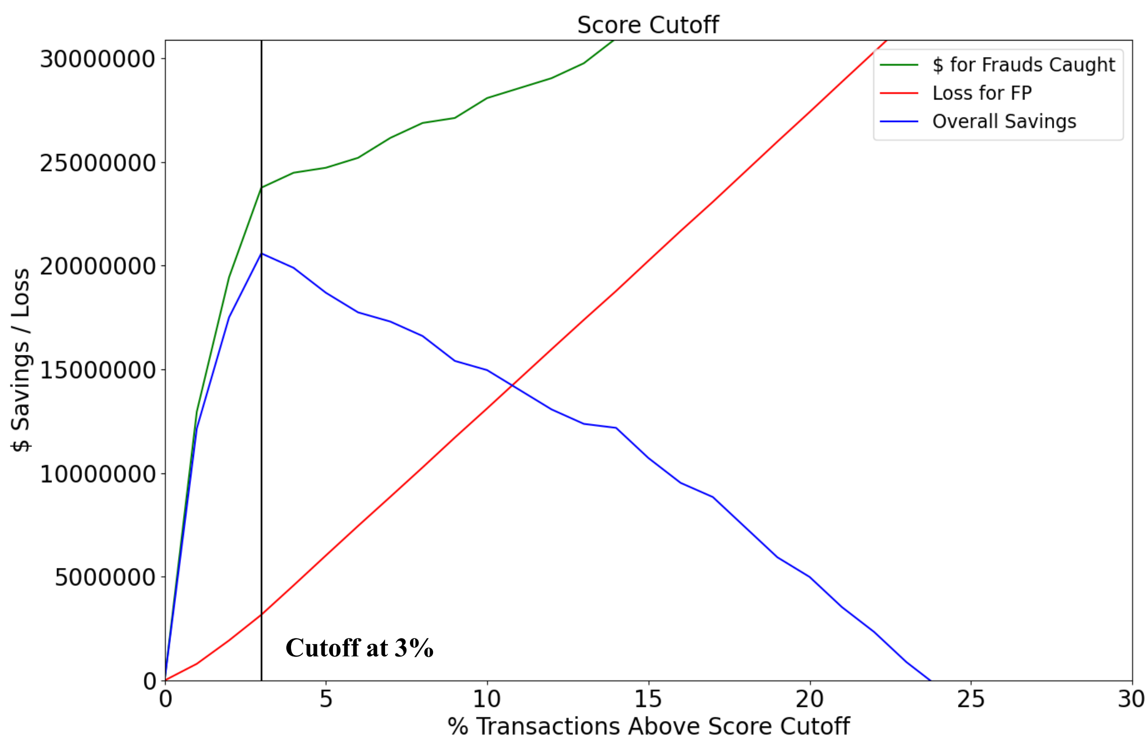
- **Conclusion**

Based on the table above, we can get **81.10%** FDR at 3% population for **training** data, **77.66%** FDR at 3% population for **testing** data, and **55.31%** FDR at 3% population for **OOT** data. In conclusion, OOT FDR shows that the final model can eliminate about **55.31%** of the fraud by declining only about **3%** of the transactions without any overfitting or underfitting.

8. Financial Curves and Score Cutoff

- **Financial Factors:** From business fraud manager, we know that **$400 saving** for every fraud caught and **$20 loss** for every false positive result. We assume that we took 100,000 samples out of 10 million population transactions per year.

- **Financial Curves:** Based on financial factors, we drew a plot showing savings for frauds caught, losses for false positive, and overall savings (difference between two values above) at all possible thresholds for score percentiles.



- **Score Cutoff**
  The financial curves show that overall savings reach to the optimal when cutoff is at 3%. In addition, we also want to deny as few transactions as possible. **Therefore, we recommend setting the score cutoff at 3%.**

- **Overall Savings**
  Using our final model, we anticipate **overall savings of $20,592,000/year** by multiplying overall savings of oot by **100** for sampling * **6** for 2-months oot data of 12 months.

9. Summary

In summary, we finished the whole pipeline to build a supervised fraud model including data observation, data cleaning, feature engineering, feature selection, model exploration, final model performance and recommended cutoff. We will describe the process in the following:

- **Data Observation and Data Cleaning**

After observing credit card transactions data that covers the time of **year 2010 with total 96,753 records and 10 fields**, we only kept **"P"** transactions and excluded one transaction **outlier** with high amount of $3,102,045.53. We also filled in missing values in **Merchnum, Merch State, and Merch Zip** fields.

- **Feature Creation and Selection**

In feature creation process, we created 10 linking entities and four kinds of variables to check frequency, amount, and uniqueness of transactions: Days since, Amount, Velocity/Relative Velocity, and Counts by entities. Then in the feature selection process, we first used **Filter** method to **keep only 300 variables** and used **Wrapper** method to keep only **20 variables** by applying **Forward Selection** to build **LGBM** models. We used all data including data for OOT in feature selection process to take seasonality into consideration. The overall performance at saturation point is around 0.70.

- **Model Exploration and Final Model**

We started from **a linear model** - logistic regression and tried **6 nonlinear models** with the **number of variables 10**. We built models on training data and evaluate performance of models by testing and OOT (data of the last two months). After comparing FDR for training, testing, and OOT, we finally chose **Random Forest Model** as our final model.

- **Final Model Performance and Financial Recommendation**

We built a random forest model with 10 variables. The **55.31%** OOT FDR and financial curves shows that our final model can eliminate about **55%** of frauds by declining only about **3%** of the transactions and save **$20,592,000 each year.**

Appendix

Data Quality Report

1. **Data Description**
   The data is a collection of **real Credit Card Transactions for business purposes from a US government organization**. The data including **1,059** fraud labels invented is to build models that can detect credit card transactions fraud. The data covers the time of **year 2010** with total **96,753 records** and **10 fields**.

2. **Summary Tables**
   The summary tables of numeric and categorical data are included in the report.

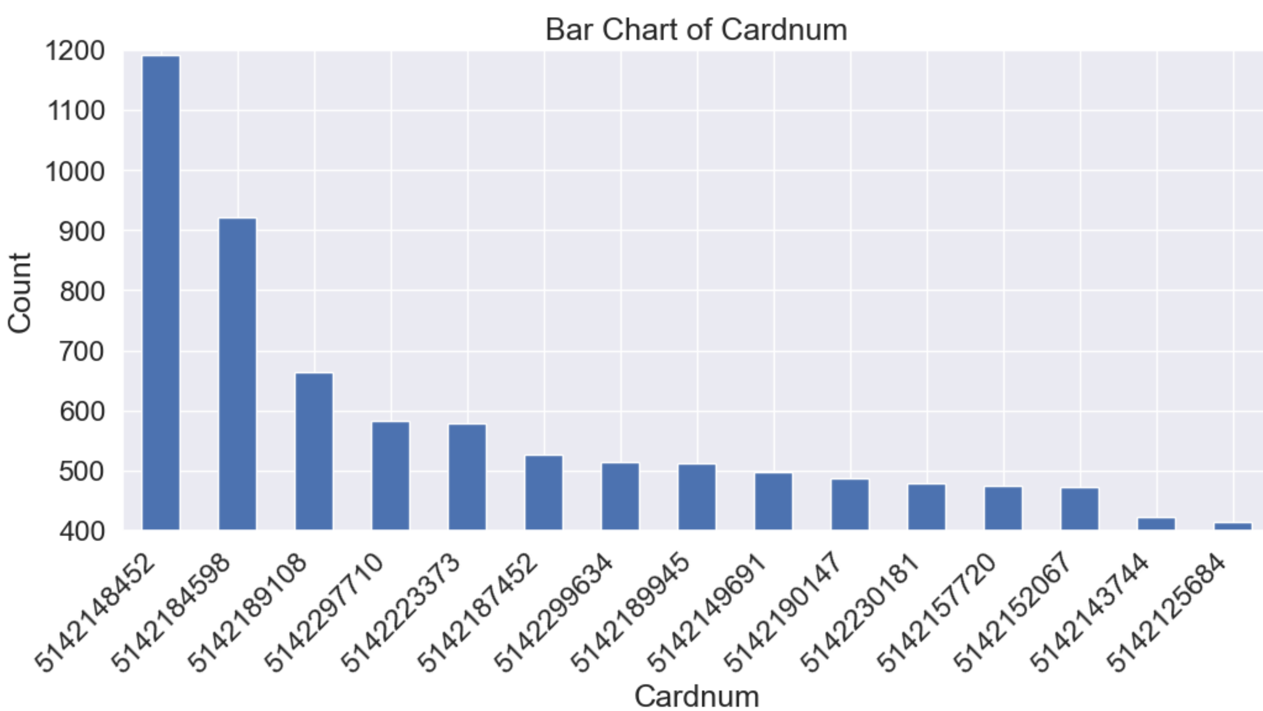
3. **Visualization of Each Field – Distribution**

(1) Field Name: Recnum

- **Visualization:** This record field has all unique values. Therefore, we don't need a histogram/distribution for this field.
- **Description:** this field is about record number of credit card transactions with ordinal unique positive integer from 1 to 96,753.

(2) Field Name: Cardnum

- **Visualization:** Bar Chart of Cardnum. The chart selects top **15** field values of Cardnum. For a better visualization of the shape, y axis starts from 400. (Data type of this field has been converted to string.)
- **Description:** Credit card number in each record/transaction. The most common Cardnum shown in transactions is 5142148452, with total amount of 1,192.

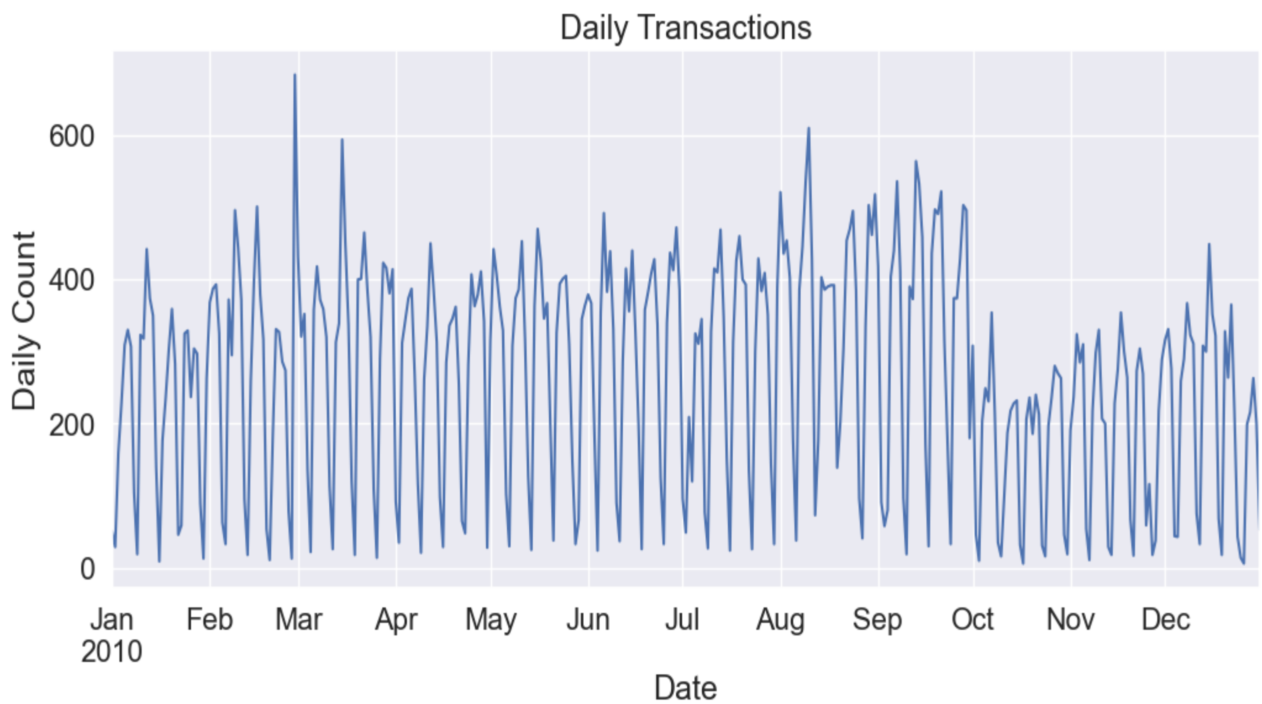


(3) Field Name: Date

There are three charts counting **total number of transactions** by **days, weeks, or months**: The Daily Transactions, The Proportion of Weekly Transactions for Both Fraudulent Transactions and Nonfraudulent Transactions, and The Monthly Transactions.

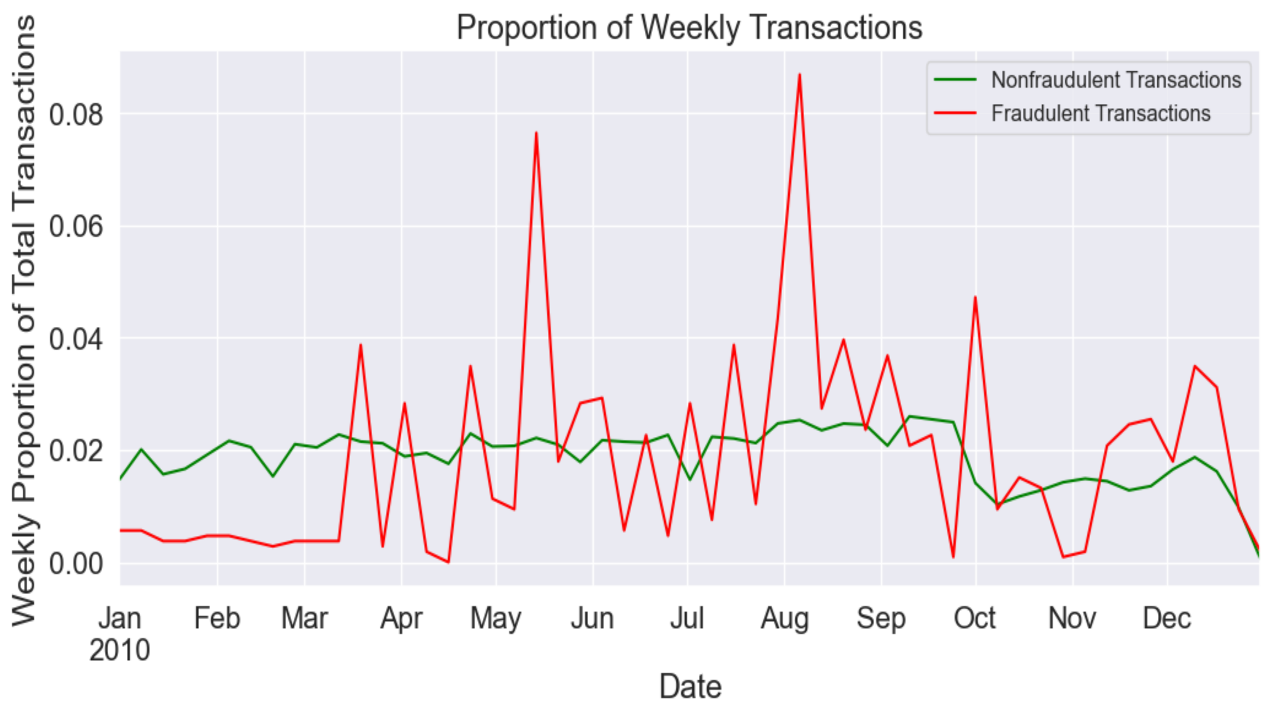
a. Visualization: Line Chart - **Daily** Transactions

- **Description:** A distribution of daily transactions amounts from 2010-01-01 to 2010-12-31. We can observe some recurring spikes showing regular increase in transactions each month.



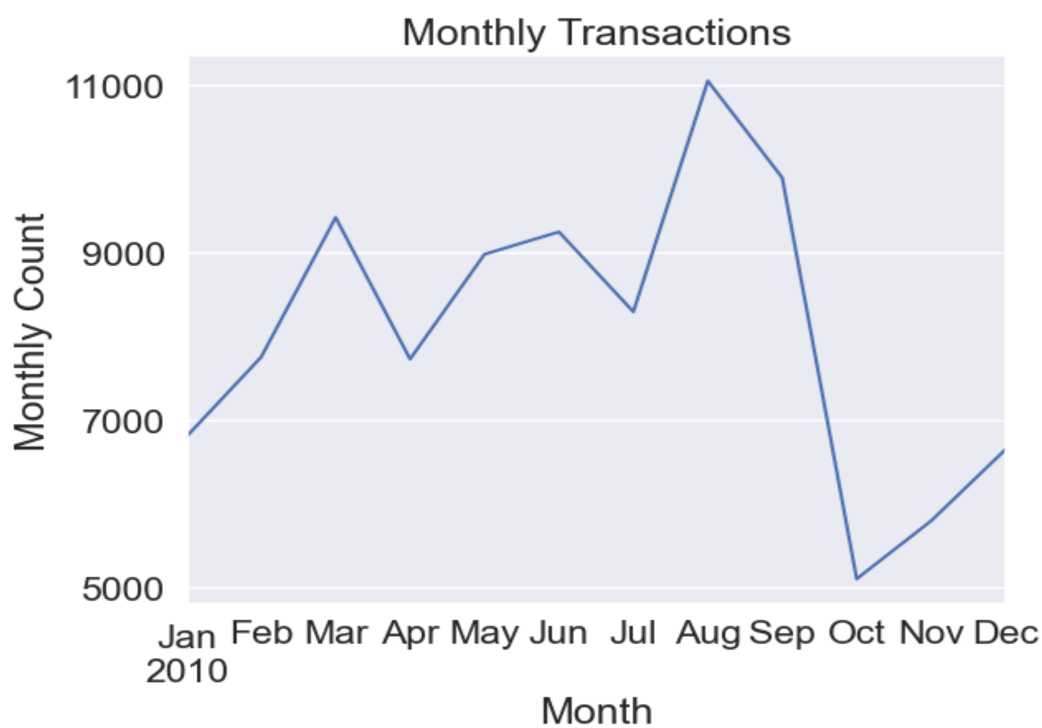
b. Visualization: Line Chart – Proportion of **Weekly** Transactions

- **Description:** A line chart representing both the proportion of weekly fraudulent transactions over total fraudulent transactions (**red** line) and the proportion of weekly nonfraudulent transactions over total nonfraudulent transactions (**green** line).



c. **Visualization:** Line Chart - **Monthly** Transactions

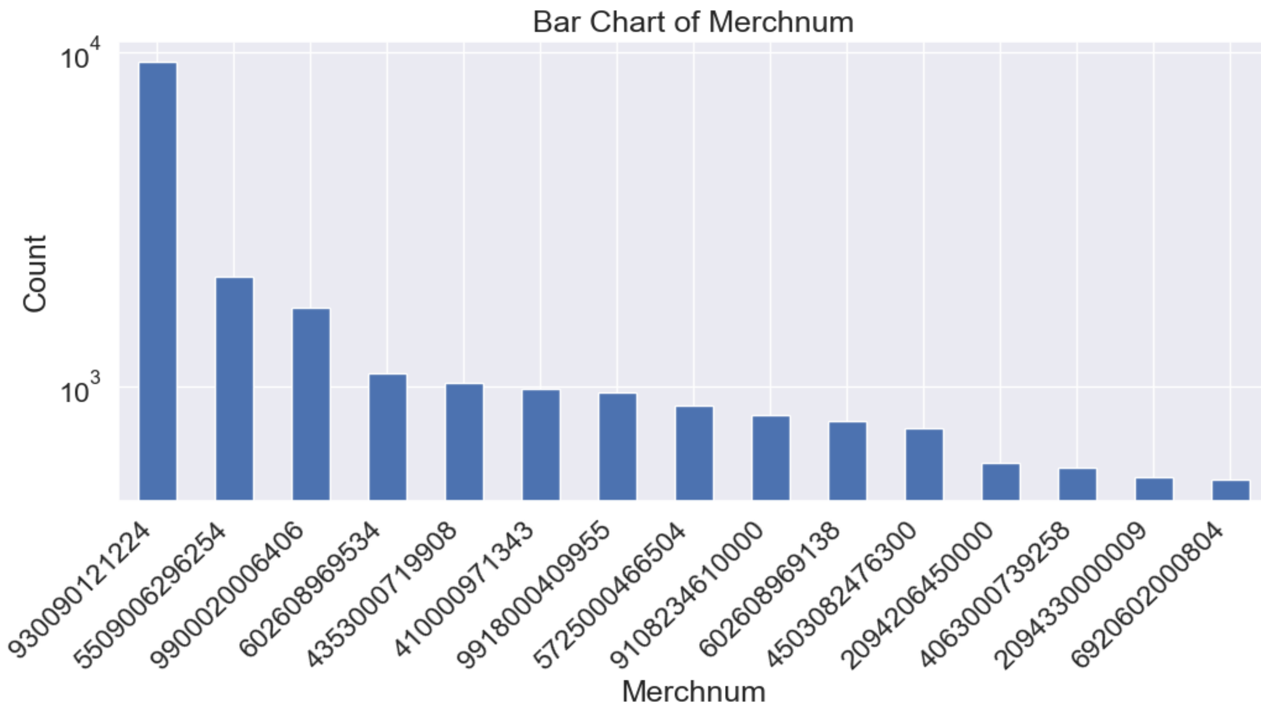
- **Description:** A distribution of monthly transactions amounts from 2010-01 to 2010-12. We can observe that monthly amount of credit card transactions dropped significantly from August to October.



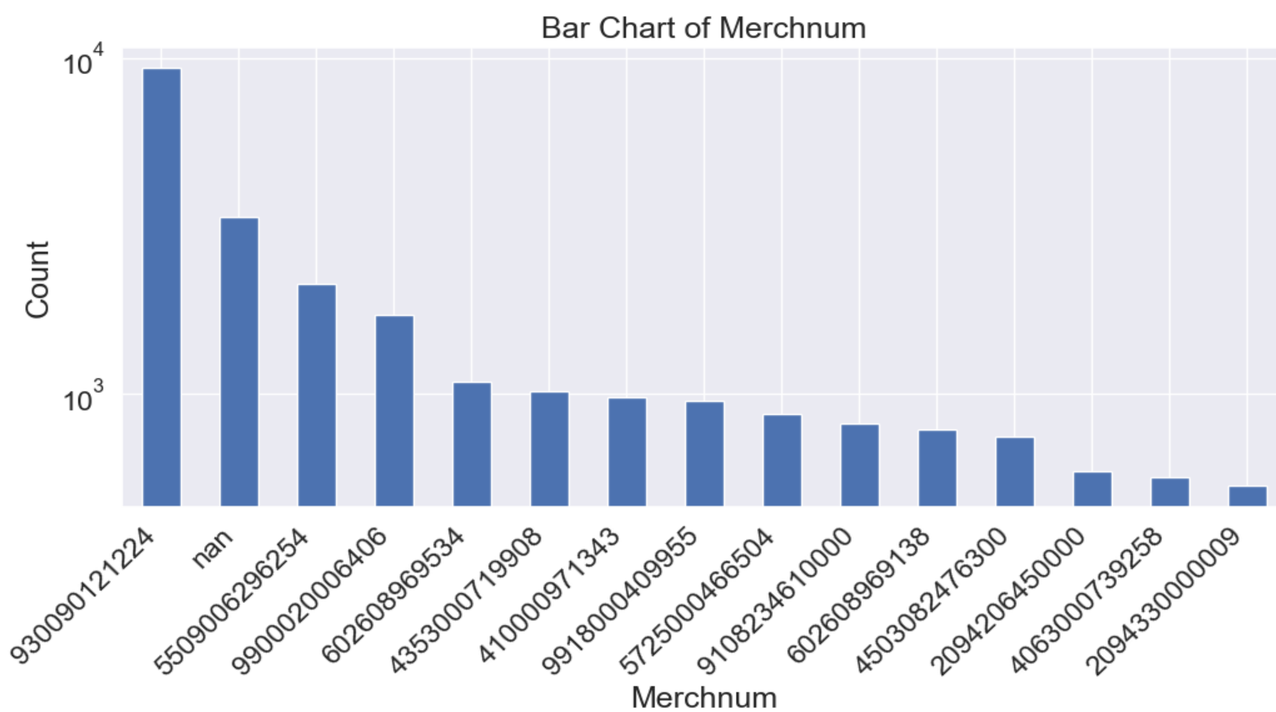
(4) Field Name: Merchnum

a. **Visualization:** Bar Chart of Merchnum (**null values excluded**). The chart selects top **15** field values of Merchnum.

- **Description:** Merchant number in each record/transaction. The most common Merchnum shown in transactions is 930090121224, with total amount of 9,310.

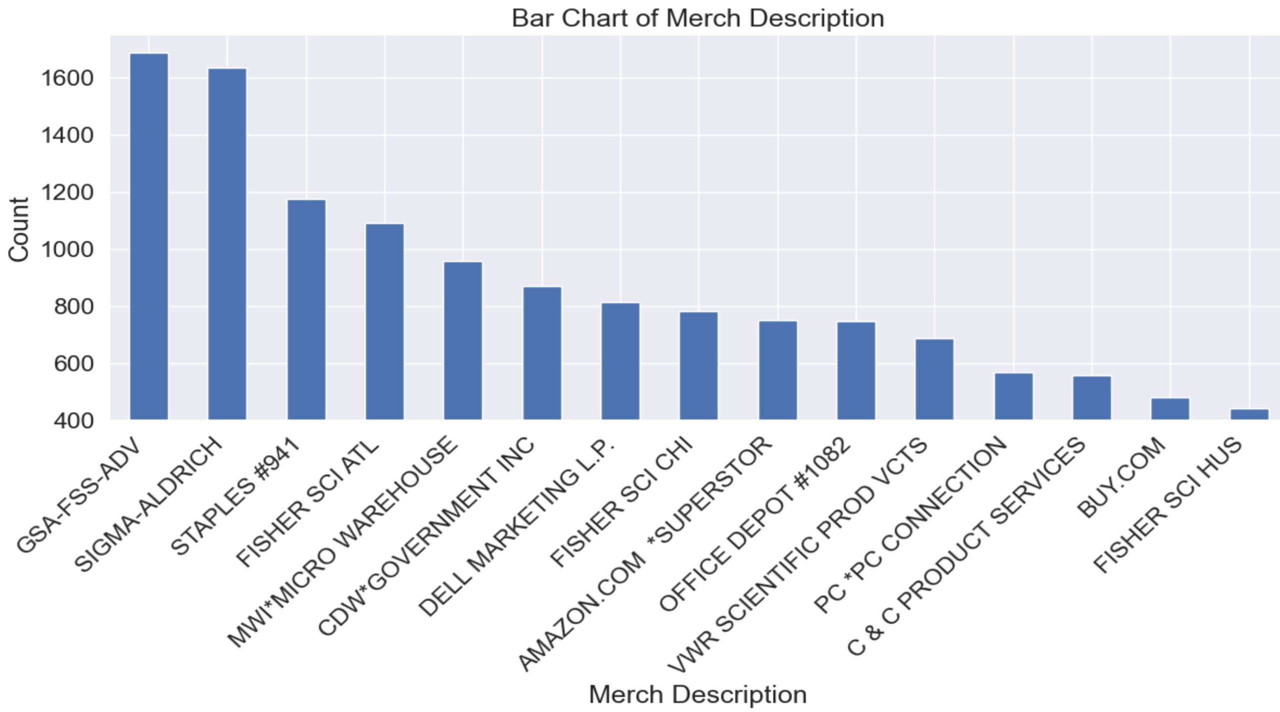


b. **Visualization:** Bar Chart of Merchnum (**null values included**). The chart selects top **15** field values of Merchnum.



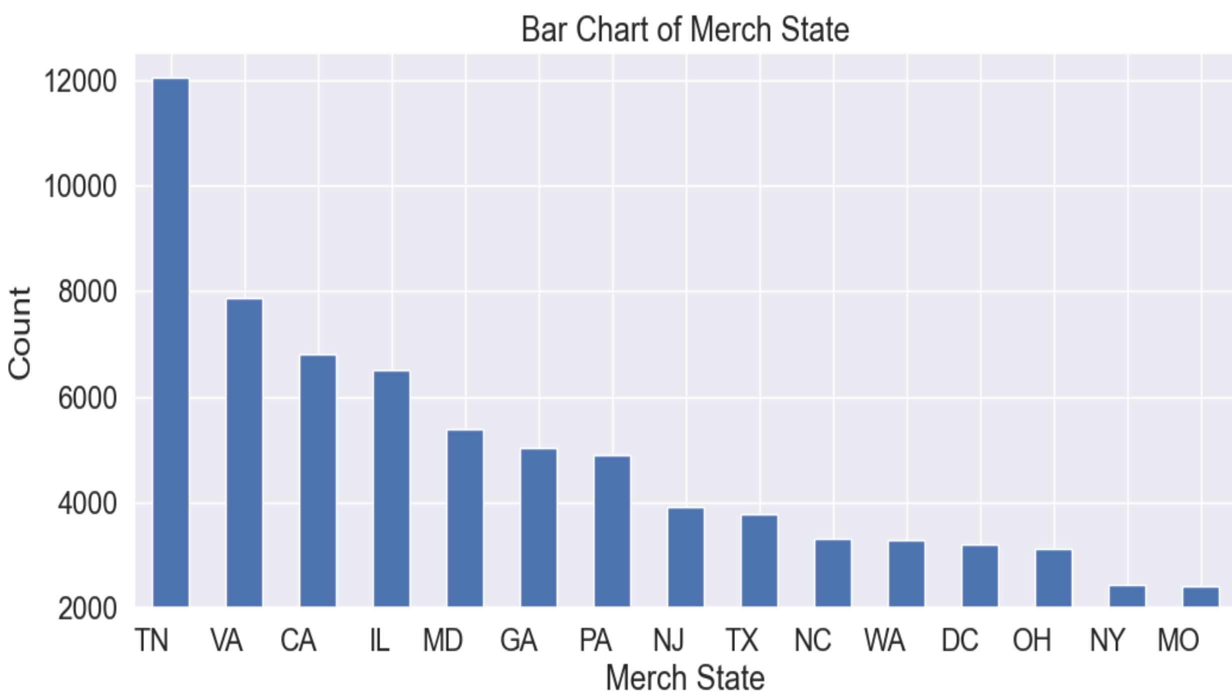
(5) Field Name: Merch description

- **Visualization:** Bar Chart of Merch Description. The chart selects top **15** field values of Merch description. For a better visualization of the shape, y axis starts from 400.
  Description: Merchant description in each record/transaction. The most common Merch description shown in transactions is GSA-FSS-ADV, with total amount of 1,688.



(6) Field Name: Merch state

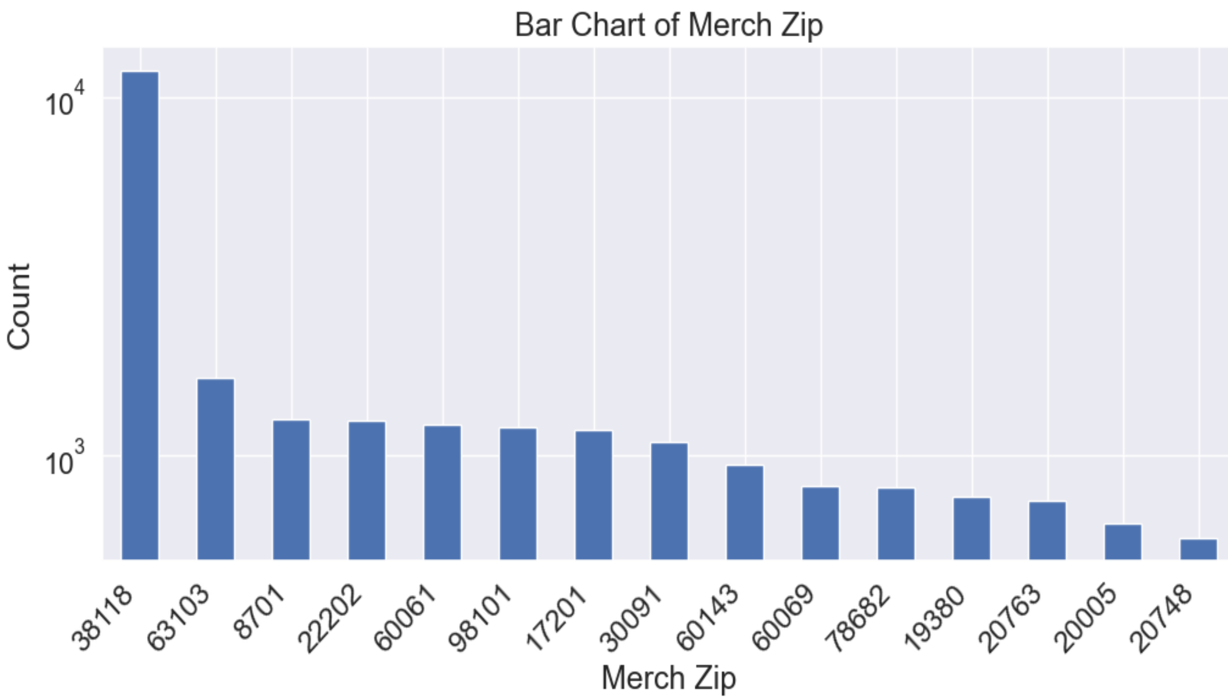
- **Visualization:** Bar Chart of Merch State. The chart selects top **15** field values of Merch state. For a better visualization of the shape, y axis starts from 2000.
- **Description:** Merchant state in each record/transaction. The most common Merch state shown in transactions is TN, with total amount of 12,035.



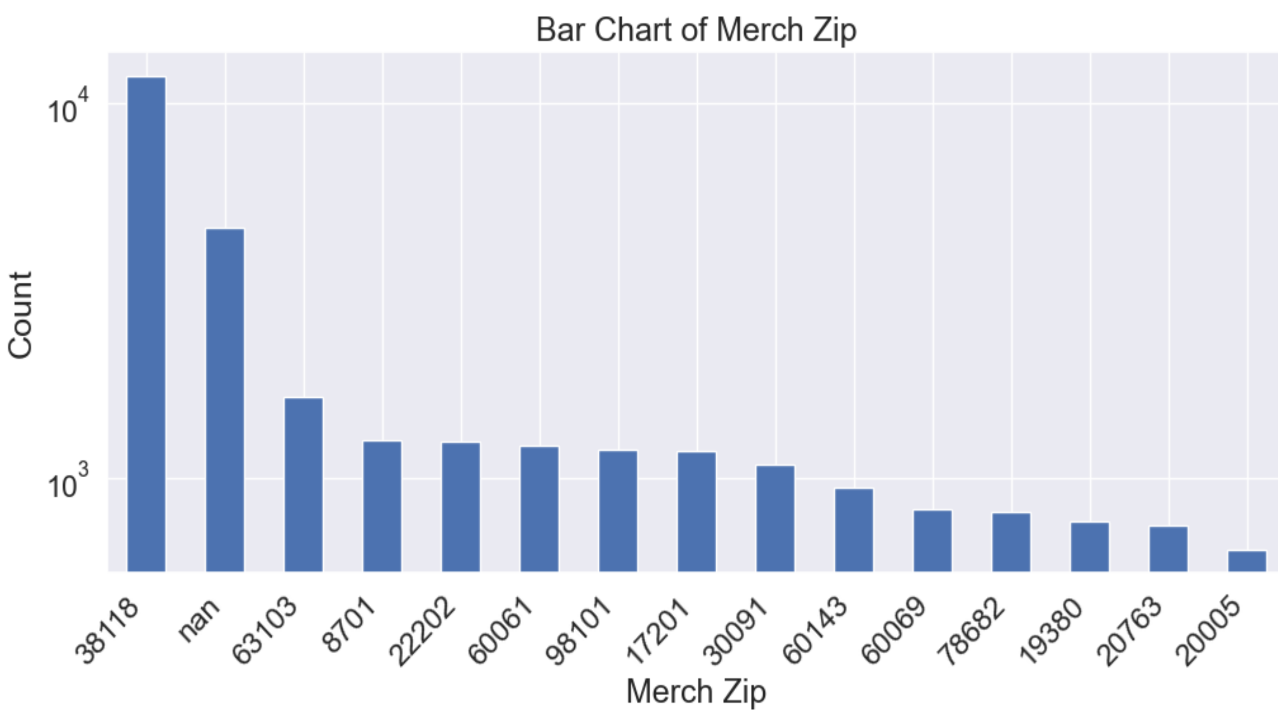
(7) Field Name: Merch zip

a. **Visualization:** Bar Chart of Merch Zip (**null values excluded**). The chart selects top **15** field values of Merch zip. (Date type of this field has been converted to string.)

- **Description:** Merchant zip code in each record/transaction. The most common Merch zip shown in applications is 38118, with total amount of 11,868.

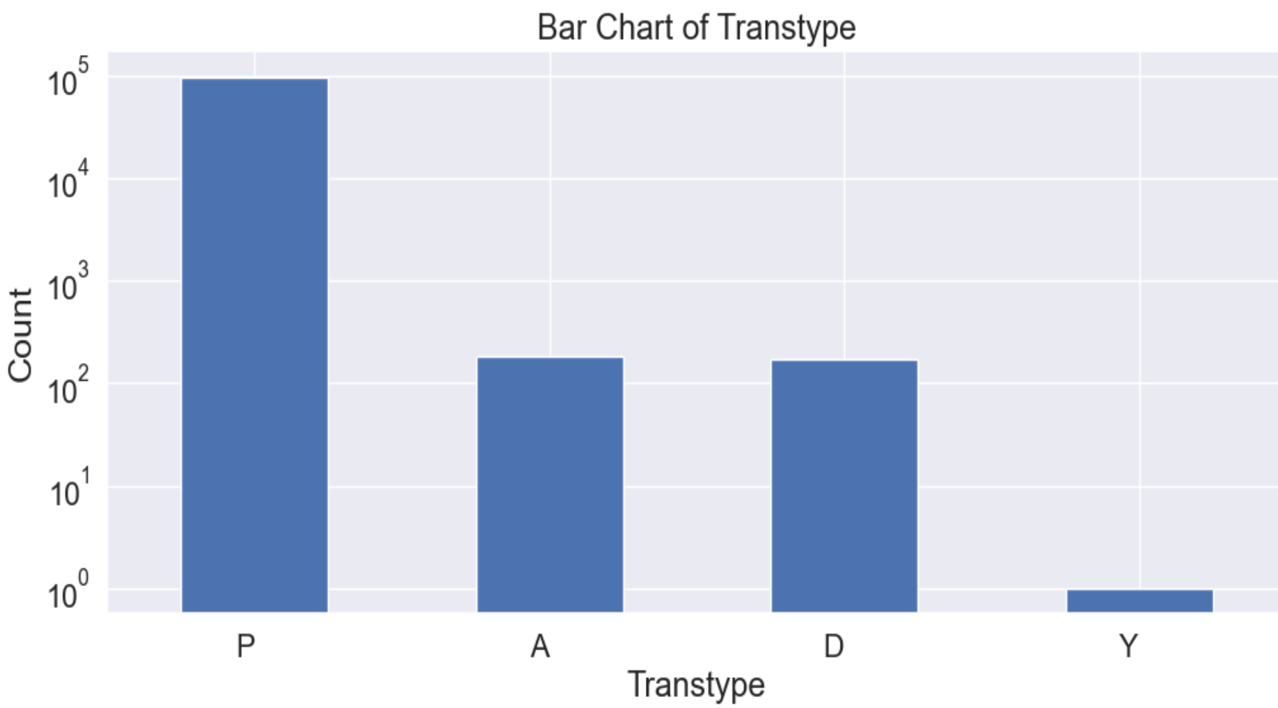


b. **Visualization:** Bar Chart of Merch Zip (**null values included**). The chart selects top **15** field values of Merch zip. (Date type of this field has been converted to string.)



(8) Field Name: Transtype

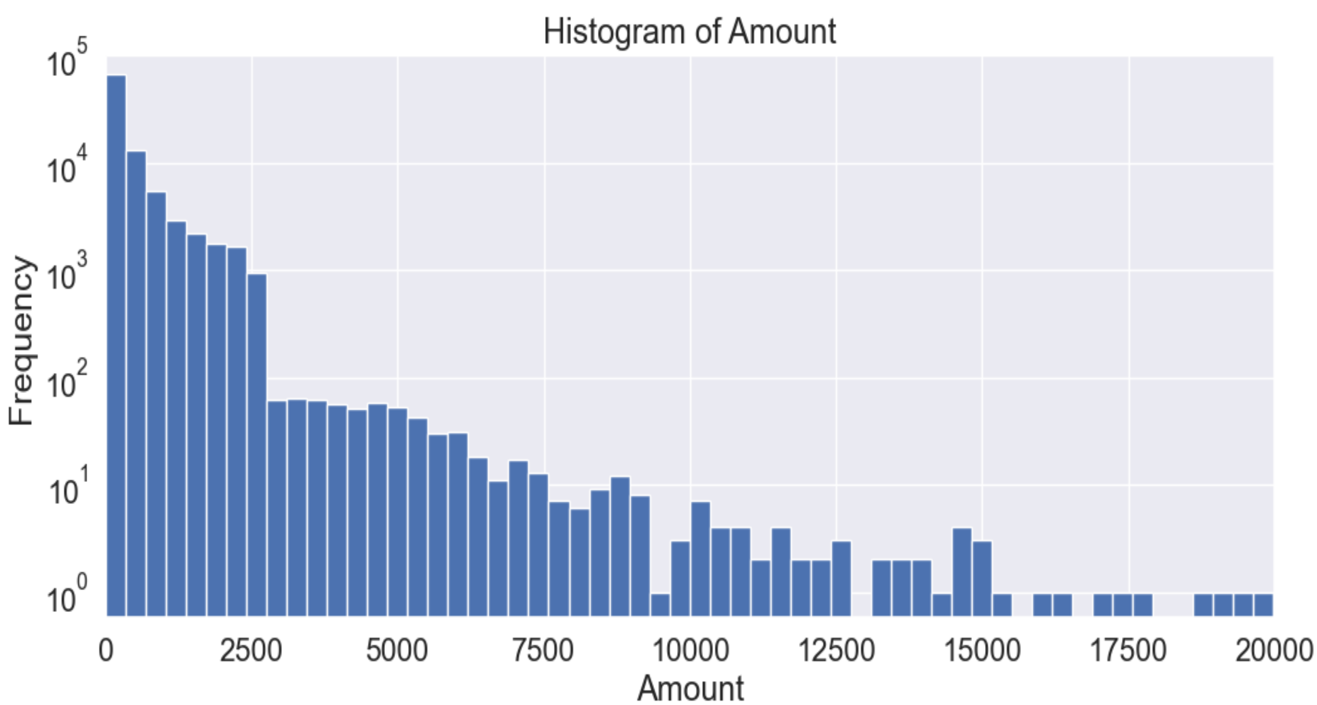
- **Visualization:** Bar Chart of Transtype. The chart shows all **4** types of transactions in this field.
- **Description:** Transaction type in each record. The most common transaction type shown is **P** meaning purchase, with total amount of 96,398.



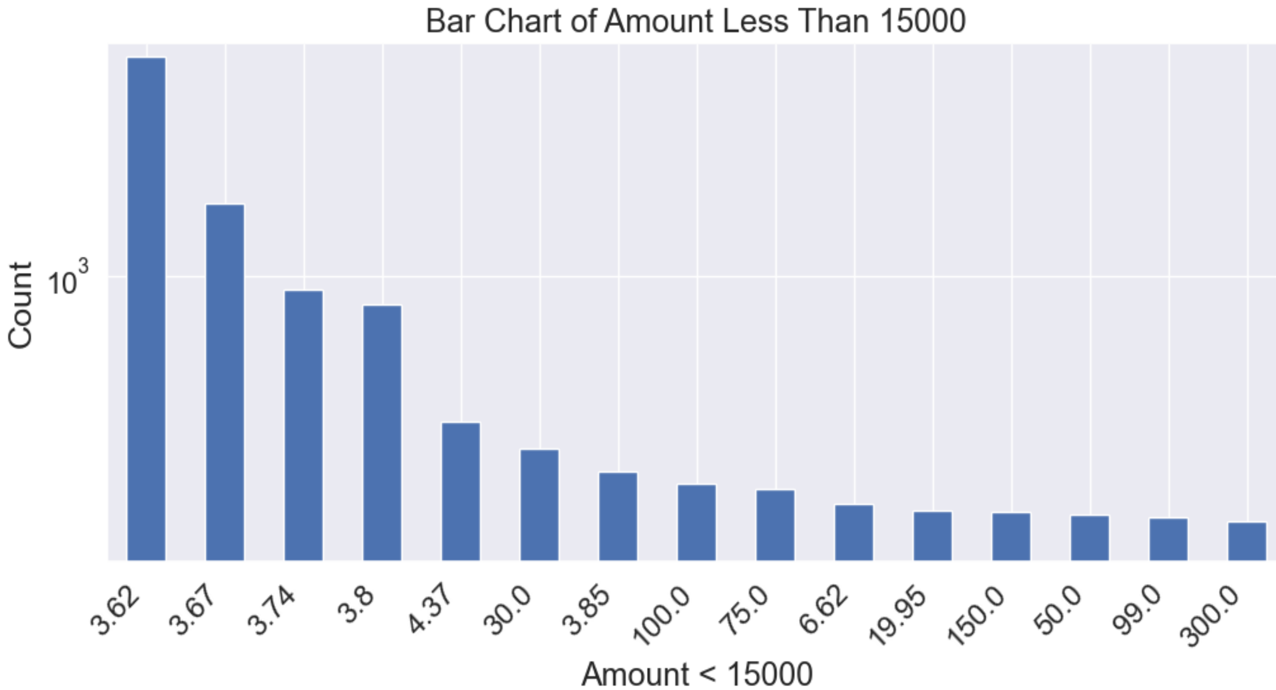
(9) Field Name: Amount

a. **Visualization:** Histogram – Histogram of Amount with a **range of x in [0, 20,000**], which covers most of the transaction amounts (around 99.97%).

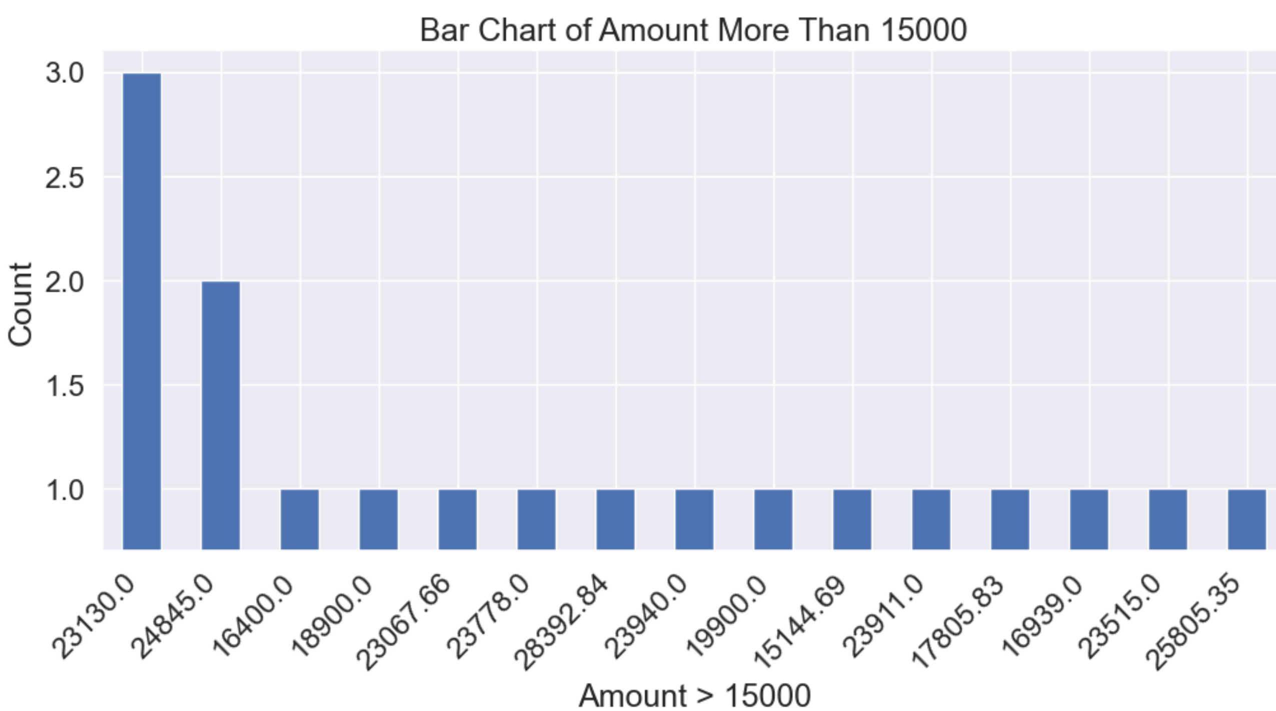
- **Description:** We can observe a big drop when amount goes over 2500, showing that most credit card transactions involve amounts of less than **$2500**. Moreover, when amount goes over **$15,000**, the count drops to close to 1, indicating there are several **outliers with large transaction amounts over $15,000**. In the below chart, we will discuss transaction amounts **below $15,000 and over $15,000**.



- **b. Visualization:** Bar Chart of Amount **Less Than 15000**. The chart selects top **15** field values of credit card transaction amounts that are less than $15,000.
- **Description:** When transaction amounts are under $15,000, the most common amount in transactions is $3.62, with a total count/frequency of 4,283.



- **c. Visualization:** Bar Chart of Amount **More Than 15000**. The chart selects top **15** field values of credit card transaction amounts that are more than $15,000.
- **Description:** When transaction amounts are over $15,000 (**outliers**), the most common amount in transactions is $23,130.00, with a total count/frequency of 3.



(10) Field Name: Fraud

- **Visualization:** The Bar Chart of Fraud Label (with blue bar = nonfraudulent records and red bar = fraudulent records).
- **Description:** The count of Fraud = 0 is 95,694. The count of Fraud = 1 is 1,059.