Credit Card Transaction Fraud Detection

University of California, San Diego

Vivek Rayalu

Table of Contents

1. Executive Summary

This project involved building a machine learning model to detect fraudulent credit card transactions. The focus was on optimizing the model to achieve a low fraud detection rate (FDR) at the top 3% of predicted probabilities, specifically for out-of-time (OOT) validation.

The final model resulted in a good FDR score of 0.6 for the top 3% of predicted probabilities for out-of-time validation. Using our model, we anticipate to save approximately $50 million per year.

2. Description of Data

The dataset used to train our model includes **97,852 Credit Card Transaction records**, each with detailed information across **10** different fields. It has transaction amounts, merchant details, and indicators of fraudulent activity to facilitate fraud detection analysis. With complete data entries for the majority of fields, it is valuable for understanding patterns in credit card usage and enhancing fraud detection algorithms. The dataset is structured to provide insights into daily transactional behaviors and is essential for developing predictive models that can identify potentially fraudulent transactions effectively.

Numeric Fields Table

- Removed ‘cardnum’ because it is a categorical variable.

Field Name	# Records Have Values	% Populated	# Zeros	Min	Max	Mean	Std. Dev.	Most Common
Merch zip	93149	0.952	0	1	99999	44,684.186	28371.57	38118
Amount	97852	1	0	0.01	3102045.53	425.466	9949.8	3.62

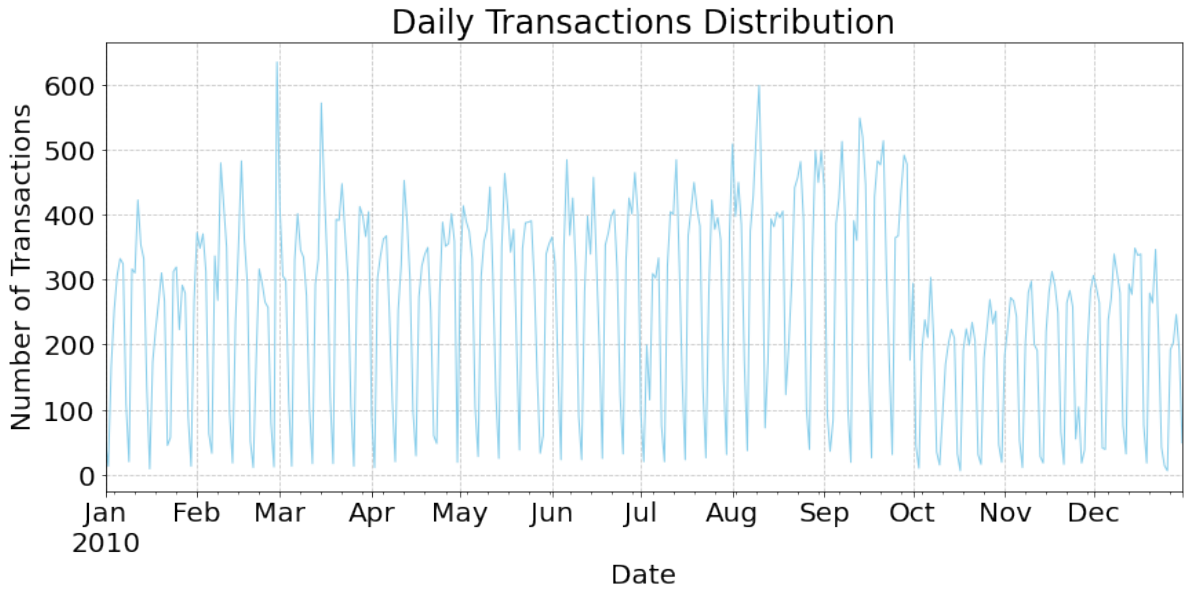
Categorical Fields Table

Field Name	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
Date	97852	100.0%	0	365	2/28/10
Merchnum	94455	96.5%	0	13091	930090121224
Merch description	97852	100.0%	0	13126	GSA-FSS-ADV
Merch state	96649	98.8%	0	227	TN
Transtype	97852	100.0%	0	1	P
Recnum	97852	100.0%	0	97852	1
Fraud	97852	100.0%	95805	2	0

Field Distribution Visualization:

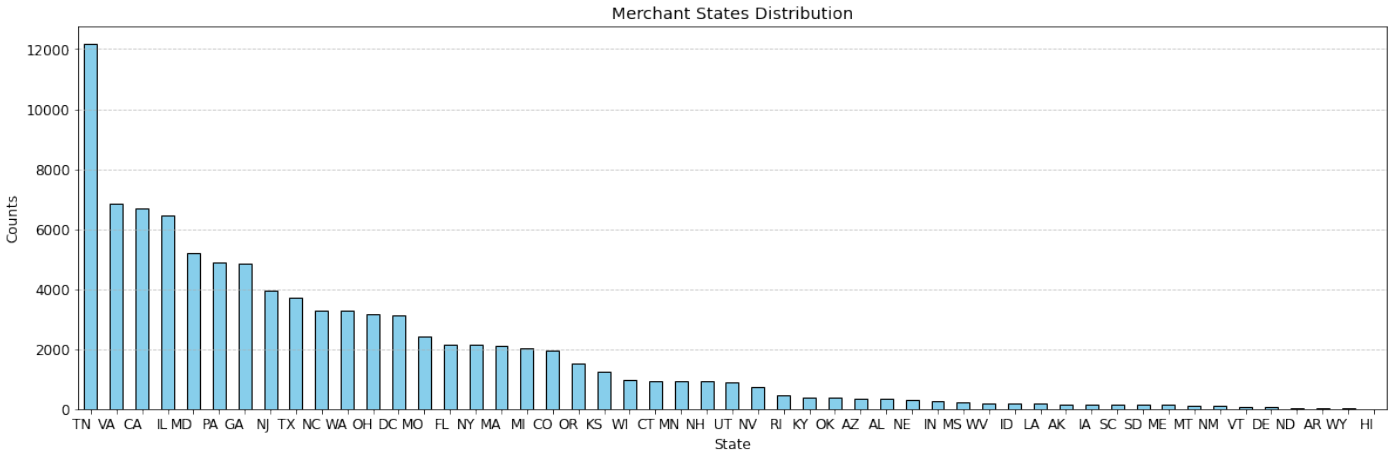
- **Field Name: Date**

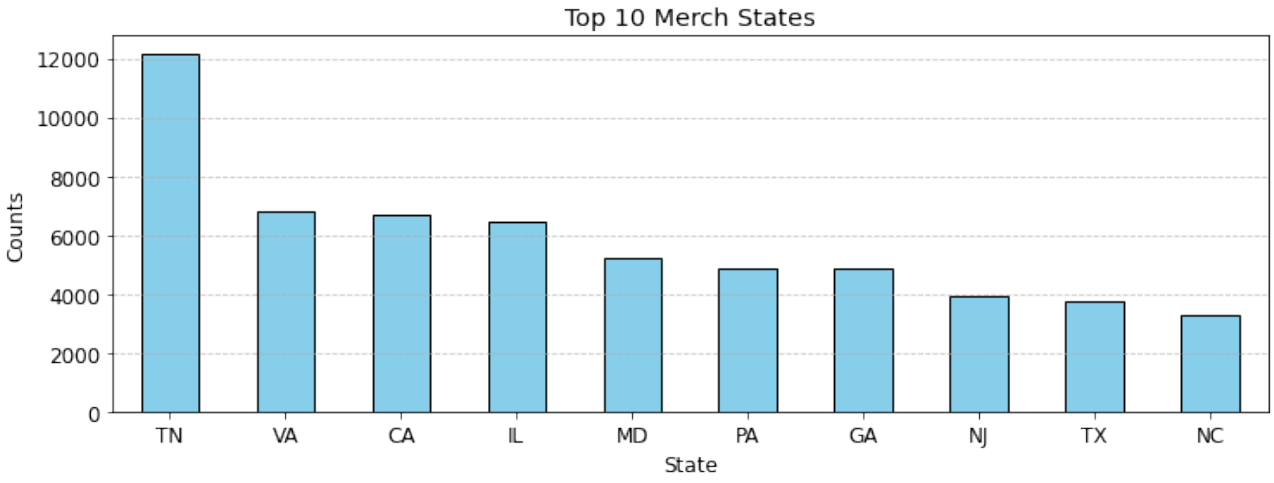
Description: Transaction date when the credit card activity occurred. This field helps in analyzing transaction volume over time and can be used to create daily or weekly distribution plots to observe temporal trends.



- **Field Name: Merch state**

Description: The state in which the merchant is located, a categorical field that allows for regional analysis of transactions across the United States.





3. Data Cleaning

The data cleaning process for the credit card transaction fraud detection project involved several key steps to ensure the dataset's accuracy and relevance for subsequent analysis. Here is a summary of the exclusions, outliers, and methods for imputation used during the data cleaning phase:

Date Conversion:

- Converted the transaction date field from a string format to datetime64[ns] type to facilitate date-based operations and analyses.

Exclusions:

- **Transaction Type**: Removed rows where the transaction type was not 'P' (representing 'Purchase') to focus solely on purchase transactions.
- **Transaction Amount**: Excluded transactions above $3,000,000 to remove outliers that could skew the analysis. This ensured the dataset only included transactions representative of typical customer behavior.
- **Resulting Dataset**: The initial dataset of 97,852 records was reduced to 97,496 records after these cleaning steps.

Merchnum Field Cleaning:

- **Imputation**: Addressed 3,279 records with missing Merchnum values by mapping non-null Merch descriptions to their corresponding Merchnum, setting specific values to "unknown," and creating unique Merchnum for remaining missing entries. This ensured every transaction had a complete and consistent set of identifying information, resulting in no null values in the Merchnum field.

State Field Cleaning:

- **Imputation**: Handled 1,028 records with missing Merch state information by mapping known Merch zip codes to their corresponding states, setting placeholders for specific transaction types, and converting non-standard state entries to a consistent label 'foreign'. This process reduced missing state entries to zero and maintained valuable geographic information.

Zip Field Cleaning:

- **Imputation**: Started with 4,347 records with missing Merch zip values. Used dictionaries to fill missing zip codes based on Merchnum and Merch descriptions, applied state-based imputation strategies, and finally set remaining missing values to 'unknown'. This process resulted in no missing values in the Merch zip field.

Outlier Treatment for Amount:

- **Investigation**: Identified and flagged a noticeable spike in transactions amounting to $3.62 by plotting the top 15 most frequent transaction amounts. Transactions between $3.62 and $3.80, particularly those related to 'FEDEX,' were flagged for further review due to their abnormal frequency.
- **Findings**: The histogram showed a significant spike at $3.62, which was marked as an outlier. Post-flagging, the distribution of transaction amounts appeared more typical.

4. Variable Creation

The creation of these variables was driven by the need to capture various patterns and anomalies in transaction data that could indicate fraudulent activities. The variables were designed to provide a comprehensive view of transaction behavior over time and across different entities, enabling the model to detect subtle and complex fraud patterns.

Temporal Patterns: By spanning multiple timeframes, these variables capture both immediate and extended transaction behaviors.
Statistical Metrics: Employing a range of metrics such as count, average, maximum, median, total, and ratios facilitates in-depth statistical analysis.
Variability and Ratios: Analysis of transaction variability and ratios assists in pinpointing unusual patterns potentially indicative of fraud.
Entity Combinations: Utilizing unique count metrics for different entity pairings aids in monitoring transaction diversity and detecting anomalies.

Outlier Detection: Squared transaction counts amplify the model's ability to recognize anomalous patterns that may indicate fraud.

Geographical Analysis: The introduction of Cardnum_geo_distance measures the geographical distance between sequential transactions, which can be crucial for identifying fraudulent operations occurring over disparate locations within short intervals.

Variable Table

Description	Metrics	# of variables created
Average fraud percentage by day of the week	Avg.	1
Amount: amount spent	Avg.	1
Fraud: binary variable: Fraud (1) or Non Fraudulent (0)		1
Each category includes variables calculated over multiple timeframes (0, 1, 3, 7, 14, 30, 60 days), featuring metrics such as count, average, maximum, median, total, and ratios of actual values to statistical norms, with 64 variables generated per category.	Count, Avg, Max, Med, Total, Actual/ Stat ratios	1408
Metrics calculated over various timeframes (7, 14, 30, 60 days) for each field, encompassing counts and total amounts	Count, total	352
Ratio of the actual value to the derived value, computed over specific periods (7, 14, 30, 60 days) and for different states of a given entity, illustrating normalized metrics that highlight deviations or consistencies over time.	Ratio	176
Variability of transactions by a specific field over various time windows (1, 3, 7, 14, 30 days)	Avg, max, med.	396
Unique count metrics for various entity combinations and time frames, used to monitor transaction diversity and detect anomalies indicative of potential fraud.	Count	648
Squared counts of transaction occurrences for various entity combinations and time frames, enhancing the sensitivity of models to detect anomalies and potential fraud patterns.	Count	168
Amount Cat: Amount field binned into categories		1
NEW VARIABLE: Distance Between Transactions: Cardnum_geo_distance		1
Total Variables:	–	3153

5. Feature Selection

Feature Selection Methods Used:

- Forward Stepwise Selection:

Approach: Forward stepwise selection was run multiple times with varying numbers of filters (100, 200, 400) and a fixed number of wrappers (20).

Results:

- With 100 filters, the model reached peak performance quickly and maintained a performance level of 0.7.
- Increasing the filters to 200 resulted in a slightly higher performance of just over 0.7.
- With 400 filters, the best performance was achieved at nearly 0.75, albeit with a significant increase in computation time.

Conclusion: Despite the longer runtime, the highest performance was obtained with 400 filters, making it the preferred setting for feature selection.

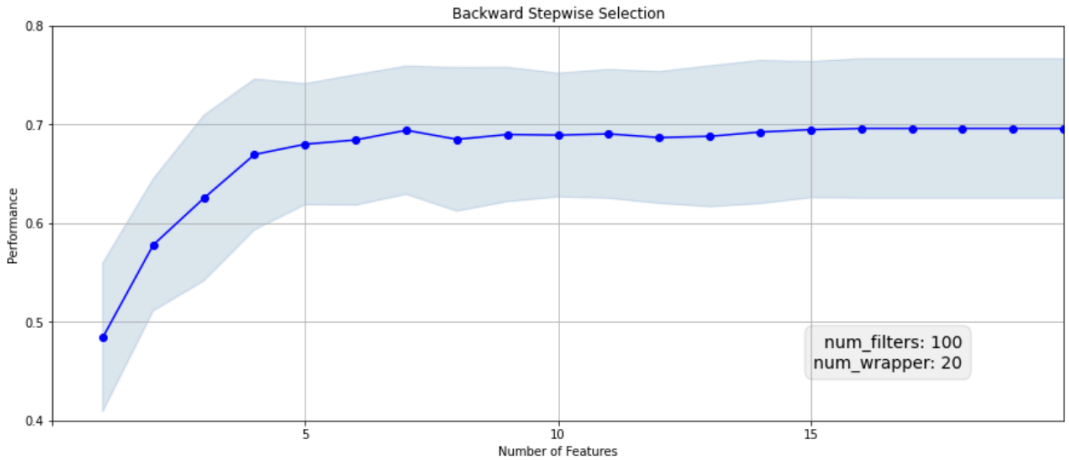
- Backward Stepwise Selection:

Approach: Backward stepwise selection was also tested with the same filter and wrapper values as forward selection (100, 200, 400).

Results:

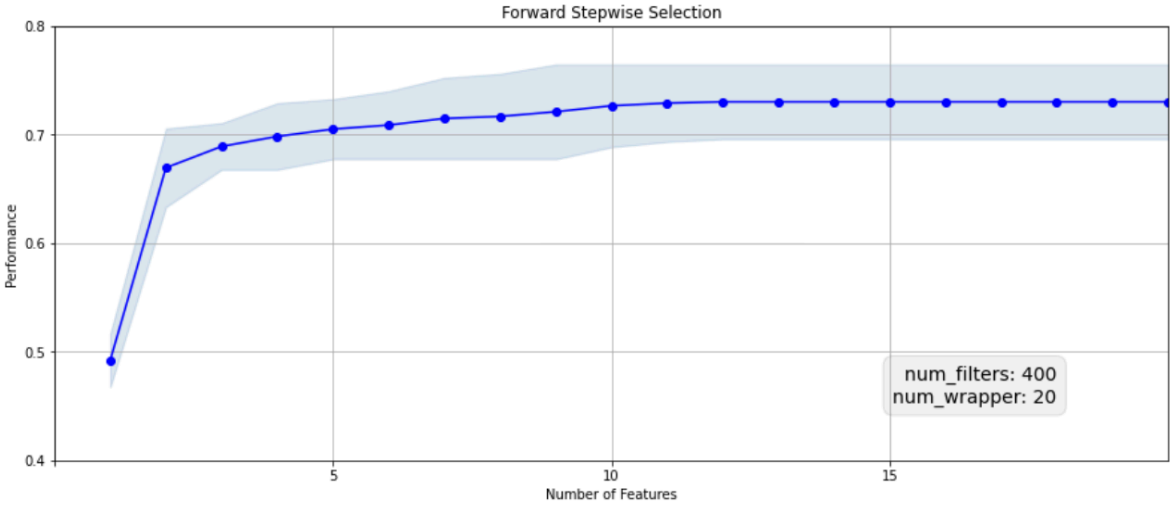
- At 100 filters, performance steadily increased and plateaued at 0.7.
- The process became too computationally intensive for 200 and 400 filters, with the kernel crashing during execution.

Conclusion: Due to the practical limitations encountered with higher filter settings, backward stepwise selection was not as viable as forward stepwise selection for this project.



Final Choice for Feature Selection Method:

Forward stepwise selection with 400 filters was chosen due to its superior performance (0.75) despite the increased computation time. This one-time computational cost is justified by the resulting improved accuracy in fraud detection.



6. Preliminary Model Exploration

In the model exploration phase of the project, various machine learning algorithms were explored to develop a robust fraud detection model. The algorithms tested include Logistic Regression, Decision Tree Classifier, Random Forest Classifier, LightGBM, and Neural Network (MLPClassifier). Each of these algorithms was selected for its unique strengths in handling complex datasets and its potential to accurately identify fraudulent transactions. Through a systematic evaluation of their performance, the most effective model was identified and refined for further analysis.

Model Description:

1. **Logistic regression** is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. It is used for binary classification problems. It predicts the probability that a given input point belongs to a certain class.

2. **Decision trees** are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.
3. **Random Forest** is an ensemble method that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.
4. **LightGBM** is a gradient boosting framework that uses tree-based learning algorithms. It is designed to be distributed and efficient with the following advantages: faster training speed and higher efficiency, lower memory usage, and better accuracy.
5. **Neural Network - Multi-layer Perceptron (MLP)** is a class of feedforward artificial neural network (ANN). The term MLP is used ambiguously, sometimes loosely to mean any feedforward ANN, sometimes strictly to refer specifically to networks composed of multiple layers of perceptrons.

Model Exploration Tables:

- Note: Best models in each class are highlighted in green to show best parameters

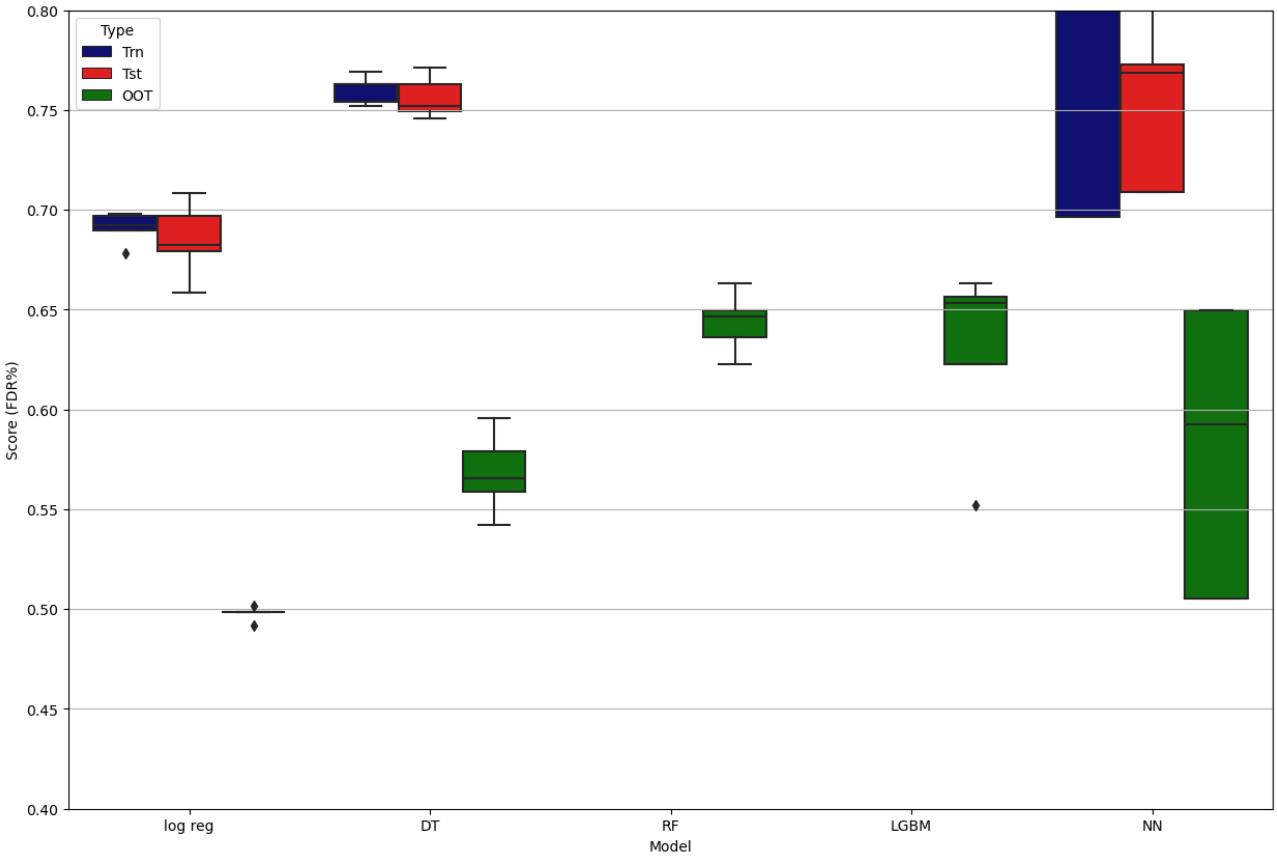
Model	Parameter					Average FDR at 3%		
Logistic Regression	**Number of Variables**	**Penalty**	**C**	**solver**	**L1 ratio**	**Train**	**Test**	**OOT**
1	5	L2	1	lbfgs	None	0.690	0.686	0.497
2	5	elasticnet	1	saga	0.5	0.693	0.684	0.497
3	10	L2	0.5	lbfgs	None	0.692	0.683	0.495
4	10	elasticnet	1	saga	0.5	0.686	0.686	0.495
5	**15**	**L2**	**0.5**	**lbfgs**	**None**	**0.689**	**0.692**	**0.497**

Decision Tree	**Number of Variables**	**Criterion**	**Splitter**	**Max Depth**	**Min Samples Split**	**Min Samples Leaf**	**Train**	**Test**	**OOT**
1	10	Gini	Best	None	2	1	1.000	0.621	0.394
2	10	Gini	Random	None	200	100	0.681	0.660	0.490
3	**15**	**Gini**	**Best**	**20**	**200**	**100**	**0.765**	**0.746**	**0.544**
4	15	Gini	Random	None	100	50	0.744	0.723	0.530
5	20	Gini	Best	40	200	100	0.763	0.738	0.564
6	20	Gini	Random	20	100	100	0.712	0.712	0.500
7	20	Gini	Best	50	100	50	0.805	0.763	0.563

Random Forest	Number of Variables	Num of estimators	Criterion	Max Depth	Min Samples Split	Min Samples Leaf	Train	Test	OOT
1	10	100	Gini	None	2	1	1.000	0.857	0.612
2	10	50	Entropy	None	100	30	0.821	0.793	0.593
3	10	150	Entropy	50	100	30	0.821	0.791	0.587
4	15	100	Gini	None	2	1	1.000	0.855	0.623
5	15	50	Entropy	None	100	30	0.816	0.803	0.581
6	15	150	Entropy	50	100	30	0.823	0.792	0.578
7	**20**	**100**	**Gini**	**None**	**2**	**1**	**1.000**	**0.865**	**0.647**
8	20	50	Entropy	None	100	30	0.828	0.796	0.608
9	20	150	Entropy	50	100	30	0.826	0.799	0.610

Boosted Tree (LGBM)	Number of Variables	Num of Leaves	Max Depth	Learning Rate	Num of estimators	Train	Test	OOT
1	10	30	-1	0.1	200	1.000	0.866	0.587
2	10	60	1	0.01	100	0.695	0.692	0.477
3	10	100	10	0.05	50	0.966	0.856	0.566
4	15	30	-1	0.1	200	1.000	0.875	0.570
5	15	60	1	0.01	100	0.696	0.683	0.475
6	15	100	10	0.05	50	0.977	0.856	0.589
7	20	30	-1	0.1	200	1.000	0.885	0.611
8	20	60	1	0.01	100	0.696	0.687	0.474
9	**20**	**100**	**10**	**0.05**	**50**	**0.962**	**0.861**	**0.631**

Neural Network	Number of Variables	Hidden Layer Sizes	Activation	Alpha	Learning Rate	Learning rate init	Max iterations	Train	Test	OOT
1	10	(100,)	Relu	0.0001	constant	0.001	500	0.685	0.642	0.490
2	10	(200,)	Relu	0.0001	Adaptive	0.01	250	0.602	0.576	0.444
3	10	(300,)	Relu	0.001	Constant	0.005	1000	0.519	0.490	0.397
4	**15**	**(100,)**	**Relu**	**0.0001**	**constant**	**0.001**	**500**	**0.743**	**0.730**	**0.566**
5	15	(200,)	Relu	0.0001	Adaptive	0.01	250	0.715	0.655	0.502
6	15	(300,)	Relu	0.0001	Adaptive	0.001	1000	0.741	0.639	0.455
7	20	(100,)	Relu	0.0001	constant	0.001	500	0.740	0.654	0.490
8	20	(200,)	Relu	0.0001	Adaptive	0.001	250	0.760	0.674	0.490
9	20	(300,)	Relu	0.001	Adaptive	0.0001	1000	0.630	0.614	0.482



Boxplot of Trn/Tst/OOT for all models explored

7. Final Model Performance

The final model chosen for the fraud detection model was a LightGBM (LGBM) classifier. This model was selected based on its superior performance on both the test and out-of-time (OOT) datasets compared to other models evaluated. We can see from the above tables that the LGBM classifier performs significantly better than the other models, although one of the Random Forest models does come close in performance, this could be due to randomness to an extent.

Model Parameters for LightGBM:

Number of Variables: 20

- This refers to the number of features or predictors used by the model. The selection of 20 variables means that the model uses the top 20 most important features to make its predictions.

Number of Leaves: 100

- This sets the maximum number of leaves in one tree. More leaves can improve the model's ability to capture complex patterns but can also increase the risk of overfitting.

Max Depth: 10

- This sets the maximum depth of the tree. Limiting the depth helps to control overfitting by preventing the model from growing too complex.

Learning Rate: 0.05

- The learning rate controls how quickly the model adapts to the problem. A lower learning rate makes the model learn more slowly and steadily, which can help in finding a more generalized solution.

Number of Estimators: 50

- This refers to the number of boosting iterations or trees in the model. More estimators can improve accuracy but also increase computation time and risk of overfitting.

The model highlighted in green (LGBM) was chosen as the final model due to its well-balanced performance across the train, test, and OOT datasets. With a training accuracy of 0.962, it also maintained a high test performance of 0.861 and an OOT performance of 0.631. These metrics indicate that the model generalizes well and effectively identifies fraudulent transactions in both the test and out-of-time datasets, without overfitting to the training dataset.

Boosted Tree (LGBM)	Number of Variables	Num of Leaves	Max Depth	Learning Rate	Num of estimators	Train	Test	OOT
9	20	100	10	0.05	50	0.962	0.861	0.631

Best Model Parameters and Performance on Trn/Tst/OOT

Training Dataset:

Number of Records: 59684

Number of Non Fraudulent Transactions: 58463

Number of Fraudulent Transactions: 1221

Fraud Detection Rate: 0.0204

bin	#recs	#g	#b	%g	%b	tot	cg	cb	%cg	FDR	KS	FPR	Fraud Savings	FP Loss	Overall Savings
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	597.0	1.0	596.0	0.16750418760469013	99.83249581239531	597.0	1.0	596.0	0.0017106883810045163	48.534201954397396	48.53249126601639	0.0016778523489932886	596000.0	30.0	595970.0
2.0	597.0	73.0	524.0	12.22780569514238	87.77219430485762	1194.0	74.0	1120.0	0.1265909401943342	91.20521172638436	91.07862078619003	0.06607142857142857	1120000.0	2220.0	1117780.0
3.0	597.0	522.0	75.0	87.43718592964824	12.562814070351763	1791.0	596.0	1195.0	1.0195702750786917	97.31270358306189	96.2931333079832	0.49874476987447697	1195000.0	17880.0	1177120.0
4.0	596.0	592.0	4.0	99.32885906040268	0.671140939597322	2387.0	1188.0	1199.0	2.0322977966333653	97.63843648208469	95.60613868545133	0.9908256880733946	1199000.0	35640.0	1163360.0
5.0	597.0	587.0	10.0	98.3249581239531	1.6750418760469046	2984.0	1775.0	1209.0	3.036471876283016	98.45276872964169	95.41629685335867	1.468155500413565	1209000.0	53250.0	1155750.0
6.0	597.0	594.0	3.0	99.49748743718592	0.5025125628140756	3581.0	2369.0	1212.0	4.052620774599699	98.6970684039088	94.6444476293091	1.9546204620462047	1212000.0	71070.0	1140930.0
7.0	597.0	595.0	2.0	99.66499162479062	0.33500837520938376	4178.0	2964.0	1214.0	5.070480361297386	98.8599348534202	93.78945449212281	2.441515650741351	1214000.0	88920.0	1125080.0
8.0	597.0	595.0	2.0	99.66499162479062	0.33500837520938376	4775.0	3559.0	1216.0	6.088339947995073	99.0228013029316	92.93446135493653	2.926809210526316	1216000.0	106770.0	1109230.0
9.0	597.0	595.0	2.0	99.66499162479062	0.33500837520938376	5372.0	4154.0	1218.0	7.1061995346927604	99.185667752443	92.07946821775023	3.4105090311986865	1218000.0	124620.0	1093380.0
10.0	596.0	596.0	0.0	100.0	0.0	5968.0	4750.0	1218.0	8.125769809771452	99.185667752443	91.05989794267154	3.8998357963875203	1218000.0	142500.0	1075500.0
11.0	597.0	597.0	0.0	100.0	0.0	6565.0	5347.0	1218.0	9.147050773231149	99.185667752443	90.03861697921185	4.389983579638752	1218000.0	160410.0	1057590.0
12.0	597.0	596.0	1.0	99.83249581239531	0.16750418760469188	7162.0	5943.0	1219.0	10.16662104830984	99.2671009771987	89.10047992888886	4.875307629204266	1219000.0	178290.0	1040710.0
13.0	597.0	597.0	0.0	100.0	0.0	7759.0	6540.0	1219.0	11.187902011769536	99.2671009771987	88.07919896542916	5.365053322395406	1219000.0	196200.0	1022800.0
14.0	597.0	596.0	1.0	99.83249581239531	0.16750418760469188	8356.0	7136.0	1220.0	12.207472286848228	99.3485342019544	87.14106191510616	5.849180327868853	1220000.0	214080.0	1005920.0
15.0	597.0	596.0	1.0	99.83249581239531	0.16750418760469188	8953.0	7732.0	1221.0	13.22704256192692	99.42996742671009	86.20292486478317	6.332514332514332	1221000.0	231960.0	989040.0
16.0	596.0	596.0	0.0	100.0	0.0	9549.0	8328.0	1221.0	14.24661283700561	99.42996742671009	85.18335458970448	6.820638820638821	1221000.0	249840.0	971160.0
17.0	597.0	597.0	0.0	100.0	0.0	10146.0	8925.0	1221.0	15.267893800465307	99.42996742671009	84.16207362624479	7.30958230958231	1221000.0	267750.0	953250.0
18.0	597.0	596.0	1.0	99.83249581239531	0.16750418760469188	10743.0	9521.0	1222.0	16.287464075543998	99.51140065146579	83.2239365759218	7.791325695581015	1222000.0	285630.0	936370.0
19.0	597.0	597.0	0.0	100.0	0.0	11340.0	10118.0	1222.0	17.308745039003696	99.51140065146579	82.2026556124621	8.27986906710311	1222000.0	303540.0	918460.0
20.0	597.0	595.0	2.0	99.66499162479062	0.33500837520938376	11937.0	10713.0	1224.0	18.326604625701382	99.6742671009772	81.34766247527583	8.752450980392156	1224000.0	321390.0	902610.0

Test Dataset:

Number of Records: 25580

Number of Non Fraudulent Transactions: 25051

Number of Fraudulent Transactions: 529

Fraud Detection Rate: 0.0206

bin	#recs	#g	#b	%g	%b	tot	cg	cb	%cg	FDR	KS	FPR	Fraud Savings	FP Loss	Overall Savings
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	256.0	24.0	232.0	9.375	90.625	256.0	24.0	232.0	0.09577779551440657	44.44444444444444	44.348666648930035	0.10344827586206896	232000.0	720.0	231280.0
2.0	256.0	88.0	168.0	34.375	65.625	512.0	112.0	400.0	0.44696304573389734	76.62835249042146	76.18138944468757	0.28	400000.0	3360.0	396640.0
3.0	255.0	211.0	44.0	82.74509803921569	17.25490196078431	767.0	323.0	444.0	1.2890094979647218	85.05747126436782	83.76846176640309	0.7274774774774775	444000.0	9690.0	434310.0
4.0	256.0	240.0	16.0	93.75	6.25	1023.0	563.0	460.0	2.2467874531087877	88.12260536398468	85.87581791087588	1.2239130434782608	460000.0	16890.0	443110.0
5.0	256.0	246.0	10.0	96.09375	3.90625	1279.0	809.0	470.0	3.228509857131455	90.03831417624521	86.80980431911375	1.721276595744681	470000.0	24270.0	445730.0
6.0	256.0	248.0	8.0	96.875	3.125	1535.0	1057.0	478.0	4.2182137441136565	91.57088122605364	87.35266748193999	2.211297071129707	478000.0	31710.0	446290.0
7.0	256.0	253.0	3.0	98.828125	1.171875	1791.0	1310.0	481.0	5.227871338494692	92.1455938697318	86.9177225312371	2.7234927234927233	481000.0	39300.0	441700.0
8.0	255.0	252.0	3.0	98.82352941176471	1.17647058823529	2046.0	1562.0	484.0	6.2335381913959615	92.72030651340997	86.486768322014	3.227272727272727	484000.0	46860.0	437140.0
9.0	256.0	255.0	1.0	99.609375	0.390625	2302.0	1817.0	485.0	7.251177268736531	92.91187739463602	85.6607001258995	3.7463917525773196	485000.0	54510.0	430490.0
10.0	256.0	255.0	1.0	99.609375	0.390625	2558.0	2072.0	486.0	8.2688163460771	93.10344827586206	84.83463192978496	4.26337448559670 75	486000.0	62160.0	423840.0
11.0	256.0	254.0	2.0	99.21875	0.78125	2814.0	2326.0	488.0	9.282464681937904	93.48659003831418	84.20412535637627	4.766393442622951	488000.0	69780.0	418220.0
12.0	256.0	254.0	2.0	99.21875	0.78125	3070.0	2580.0	490.0	10.296113017798707	93.86973180076629	83.57361878296759	5.26530612244898	490000.0	77400.0	412600.0
13.0	255.0	253.0	2.0	99.2156862745098	0.7843137254901933	3325.0	2833.0	492.0	11.305770612179742	94.25287356321839	82.94710295103864	5.758130081300813	492000.0	84990.0	407010.0
14.0	256.0	255.0	1.0	99.609375	0.390625	3581.0	3088.0	493.0	12.323409689520313	94.44444444444444	82.12103475492412	6.26369168356998	493000.0	92640.0	400360.0
15.0	256.0	254.0	2.0	99.21875	0.78125	3837.0	3342.0	495.0	13.337058025381117	94.82758620689656	81.4905281815154 4	6.751515151515152	495000.0	100260.0	394740.0
16.0	256.0	254.0	2.0	99.21875	0.78125	4093.0	3596.0	497.0	14.350706361241919	95.21072796934865	80.86002160810673	7.2354124748490944	497000.0	107880.0	389120.0
17.0	256.0	256.0	0.0	100.0	0.0	4349.0	3852.0	497.0	15.372336180062256	95.21072796934865	79.83839178928639	7.750503018108652	497000.0	115560.0	381440.0
18.0	255.0	253.0	2.0	99.2156862745098	0.7843137254901933	4604.0	4105.0	499.0	16.38199377444329	95.59386973180077	79.21187595735748	8.226452905811623	499000.0	123150.0	375850.0
19.0	256.0	255.0	1.0	99.609375	0.390625	4860.0	4360.0	500.0	17.39963285178386	95.78544061302682	78.38580776124296	8.72	500000.0	130800.0	369200.0
20.0	256.0	254.0	2.0	99.21875	0.78125	5116.0	4614.0	502.0	18.413281187644664	96.16858237547893	77.7553011878 3428	9.191235059760956	502000.0	138420.0	363580.0

OOT Dataset:

Number of Records: 12232

Number of Non Fraudulent Transactions: 11935

Number of Fraudulent Transactions: 297

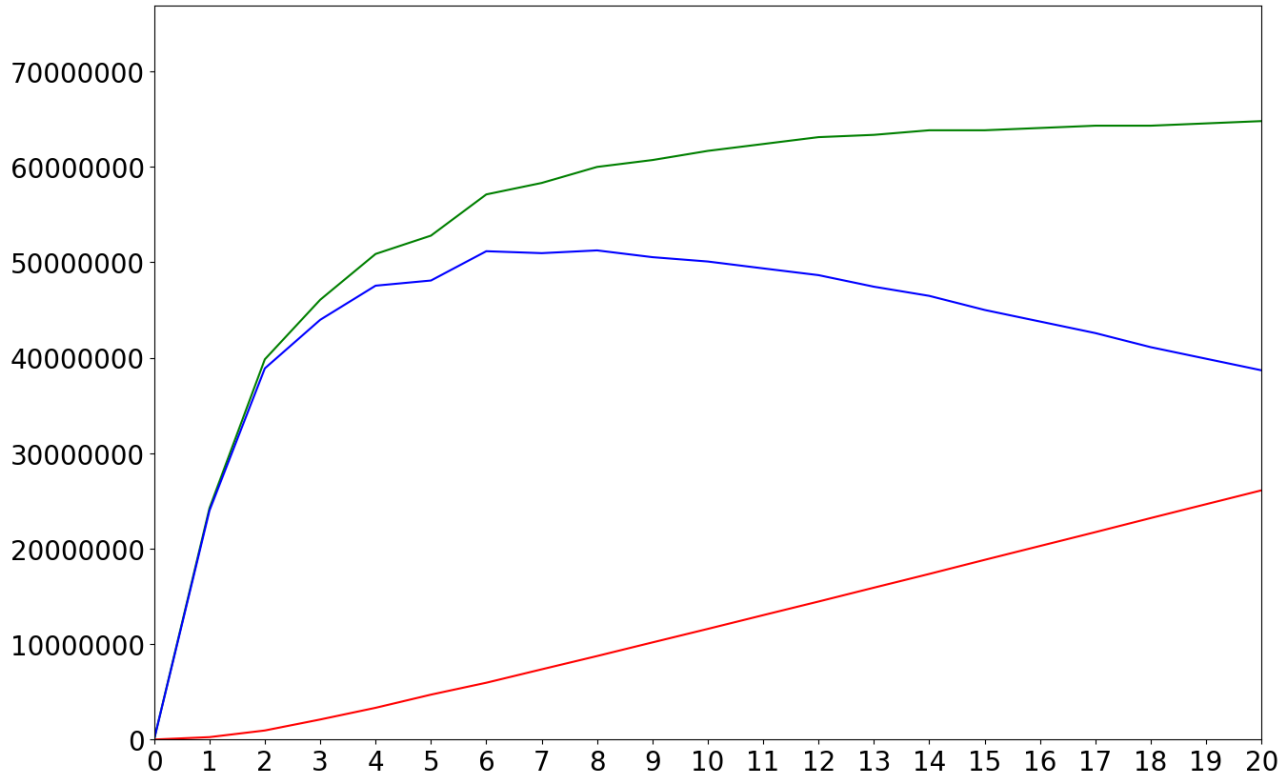
Fraud Detection Rate: 0.024

bin	#recs	#g	#b	%g	%b	tot	cg	cb	%cg	FDR	KS	FPR
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	122.0	19.0	103.0	15.573770491803279	84.42622950819673	122.0	19.0	103.0	0.1591956430666108	34.68013468013468	34.52093903706807	0.18446601941747573
2.0	123.0	61.0	62.0	49.59349593495935	50.40650406504065	245.0	80.0	165.0	0.6702974444909929	55.55555555555556	54.885258111064566	0.48484848484848486
3.0	122.0	90.0	32.0	73.77049180327869	26.229508196721312	367.0	170.0	197.0	1.4243820695433598	66.32996632996633	64.90558426042297	0.8629441624365483
4.0	122.0	109.0	13.0	89.34426229508196	10.655737704918039	489.0	279.0	210.0	2.3376623376623376	70.70707070707071	68.36940836940838	1.3285714285714285
5.0	123.0	111.0	12.0	90.2439024390244	9.756097560975604	612.0	390.0	222.0	3.2677000418935904	74.74747474747475	71.47977470558116	1.7567567567567568
6.0	122.0	117.0	5.0	95.90163934426229	4.098360655737707	734.0	507.0	227.0	4.248010054461667	76.43097643097643	72.18296637651476	2.2334801762114536
7.0	122.0	119.0	3.0	97.54098360655738	2.4590163934426243	856.0	626.0	230.0	5.245077503142019	77.44107744107744	72.19599993793543	2.7217391304347824
8.0	123.0	113.0	10.0	91.869918699187	8.130081300813004	979.0	739.0	240.0	6.191872643485547	80.8080808080808	74.61620816459525	3.0791666666666666
9.0	122.0	116.0	6.0	95.08196721311475	4.918032786885249	1101.0	855.0	246.0	7.1638039379974865	82.82828282828282	75.66447889028534	3.475609756097561
10.0	122.0	119.0	3.0	97.54098360655738	2.4590163934426243	1223.0	974.0	249.0	8.160871386677838	83.83838383838383	75.677512451706	3.9116465863453813
11.0	123.0	120.0	3.0	97.5609756097561	2.439024390243901	1346.0	1094.0	252.0	9.166317553414327	84.84848484848484	75.68216729507051	4.341269841269841
12.0	122.0	116.0	6.0	95.08196721311475	4.918032786885249	1468.0	1210.0	258.0	10.138248847926267	86.86868686868686	76.7304380207606	4.689922480620155
13.0	122.0	119.0	3.0	97.54098360655738	2.4590163934426243	1590.0	1329.0	261.0	11.13531629660662	87.87878787878788	76.74347158218126	5.091954022988506
14.0	122.0	120.0	2.0	98.36065573770492	1.639344262295083	1712.0	1449.0	263.0	12.140762463343108	88.55218855218855	76.41142608884545	5.509505703422053
15.0	123.0	122.0	1.0	99.1869918699187	0.8130081300813004	1835.0	1571.0	264.0	13.162966066191872	88.88888888888889	75.72592282269702	5.950757575757576
16.0	122.0	121.0	1.0	99.18032786885246	0.8196721311475414	1957.0	1692.0	265.0	14.176790950984499	89.22558922558923	75.04879827460474	6.384905660377359
17.0	122.0	121.0	1.0	99.18032786885246	0.8196721311475414	2079.0	1813.0	266.0	15.190615835777127	89.56228956228956	74.37167372651244	6.815789473684211
18.0	123.0	119.0	4.0	96.7479674796748	3.2520325203252014	2202.0	1932.0	270.0	16.187683284457478	90.9090909090909	74.72140762463343	7.155555555555556
19.0	122.0	121.0	1.0	99.18032786885246	0.8196721311475414	2324.0	2053.0	271.0	17.201508169250104	91.24579124579125	74.04428307654115	7.575645756457565
20.0	122.0	122.0	0.0	100.0	0.0	2446.0	2175.0	271.0	18.22371177209887	91.24579124579125	73.02207947369239	8.025830258302584

8. Financial Curves and Recommended Cutoff

In this section, we analyze the financial impact of the fraud detection model using three financial curves: Fraud Savings, FP Loss, and Overall Savings. The plot shows these financial metrics over different cutoff thresholds.

Fraud Savings (green line): Represents the savings achieved by correctly identifying fraudulent transactions.
FP Loss (red line): Represents the cost incurred due to false positives, i.e., legitimate transactions wrongly flagged as fraudulent.
Overall Savings (blue line): Calculated as the difference between Fraud Savings and FP Loss, showing the net financial benefit.



Savings VS Score %

The recommended cutoff to maximize savings and also detect fraudulent transactions would be 4% according to the plot.

If we estimate a loss of $20 for each falsely flagged fraudulent transaction and a gain of $400 for each correctly identified fraudulent transaction, we can calculate the potential savings of our model when extended to a large-scale operation. For a realistic scenario involving millions of records, typical for any large credit card company, we apply a multiplier to approximate these savings. This multiplier is calculated as (12/2)×(10,000,000/100,000), where we consider that our out-of-time (OOT) dataset represents only 2 out of 12 months, and we scale up from 100,000 sample transactions to a full year of 10 million transactions.

The maximum possible amount that can be saved due to the model being implemented is around $51,252,000.

9. Summary

In this project, we developed an effective model for detecting fraudulent transactions in credit card data. We began by cleaning the dataset, converting the transaction date fields to appropriate formats, and excluding non-purchase transactions and extreme outliers to focus our analysis on relevant data. The initial dataset of 97,852 records was reduced to 97,496 records after these cleaning steps.

Next, we created new variables to capture various transaction patterns, enhancing the predictive power of our model. This included calculating metrics over multiple timeframes and aggregating transaction data in various ways. Our feature selection process used forward stepwise selection to identify the most impactful variables, ultimately selecting 20 key features for the final model. Other feature selection methods were tried but we found forward stepwise selection to be the most effective at selecting important variables which can be predictive of fraudulent transactions.

We explored several machine learning algorithms to identify the best approach for fraud detection, including Logistic Regression, Decision Tree, Random Forest, LightGBM, XGBoost, SVM, and Neural Networks. The LightGBM classifier was chosen for its balanced and strong performance across training, testing, and out-of-time datasets.

For the financial analysis, we estimated the potential savings from using the model on a large scale. Assuming a loss of $20 for each falsely flagged transaction and a gain of $400 for each correctly identified fraud, we applied a scaling factor to project the savings for a full year of data, accounting for millions of transactions. We calculated the Fraud Savings, FP Loss, and Overall Savings for each threshold, identifying the optimal cutoff point of 4% that maximized the financial benefits.

Our analysis showed that the model could achieve a Fraud Detection Rate (FDR@3%) of approximately 0.611 for the out-of-time dataset, highlighting its effectiveness in identifying fraudulent transactions. The maximum possible savings were found to be around $50 million, confirming the model's large potential financial impact if deployed.

Future improvements could include refining the feature selection process, experimenting with different model architectures, and continually updating the model with new data to maintain its accuracy. Additionally, incorporating real-time transaction data could further enhance the model's effectiveness in preventing fraud.

This project displayed a comprehensive approach to building and validating a fraud detection model, providing valuable insights and substantial financial benefits for large-scale credit card operations.

10. Appendix

Data Quality Report

1. Data Description

This dataset includes 97,852 **Credit Card Transaction records**, each with detailed information across 10 different fields. It has transaction amounts, merchant details, and indicators of fraudulent activity to facilitate fraud detection analysis. With complete data entries for the majority of fields, it is valuable for understanding patterns in credit card usage and enhancing fraud detection algorithms. The dataset is structured to provide insights into daily transactional behaviors and is essential for developing predictive models that can identify potentially fraudulent transactions effectively.

2. Summary Tables

Numeric Fields Table

- Removed 'cardnum' because it is a categorical variable.

Field Name	# Records Have Values	% Populated	# Zeros	Min	Max	Mean	Std. Dev.	Most Common
Merch zip	93149	0.952	0	1	99999	44,684.186	28371.57	38118
Amount	97852	1	0	0.01	3102045.53	425.466	9949.8	3.62

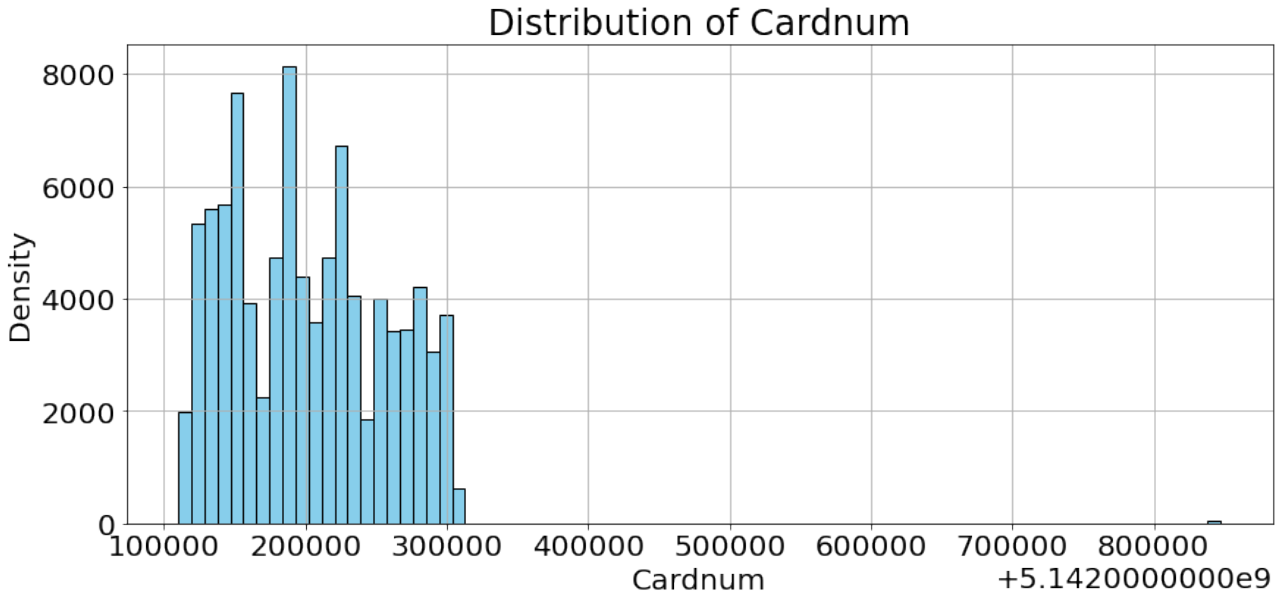
Categorical Fields Table

Field Name	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
Date	97852	100.0%	0	365	2/28/10
Merchnum	94455	96.5%	0	13091	930090121224
Merch description	97852	100.0%	0	13126	GSA-FSS-ADV
Merch state	96649	98.8%	0	227	TN
Transtype	97852	100.0%	0	1	P
Recnum	97852	100.0%	0	97852	1
Fraud	97852	100.0%	95805	2	0

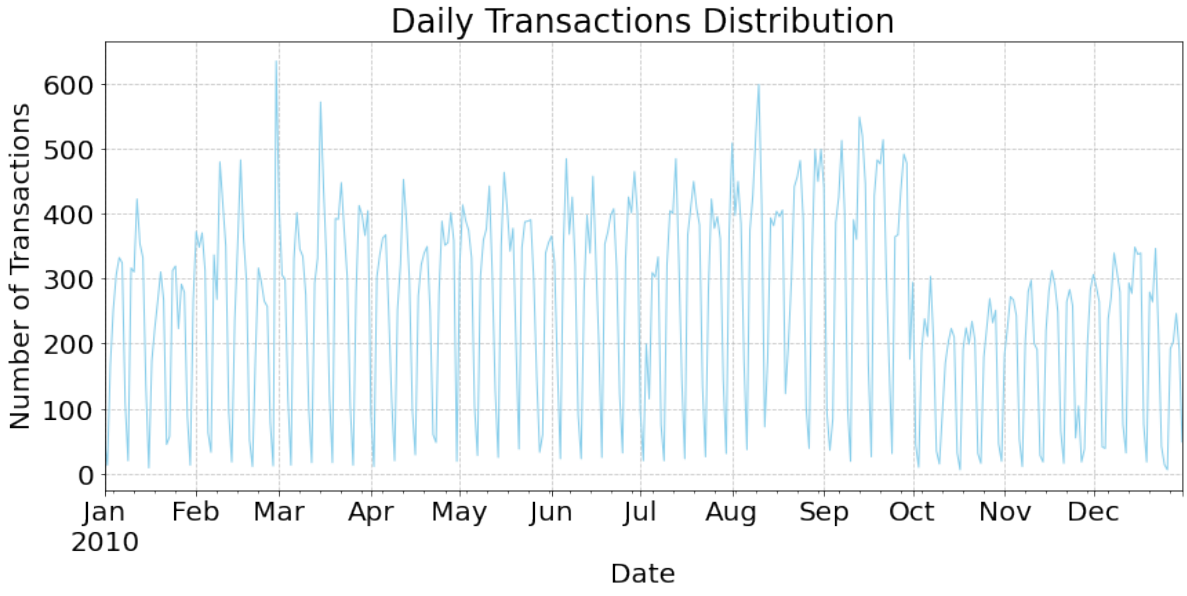
3. Visualization of Each Field

3. **Field Name: Recnum**
   Description: Ordinal unique positive integer for each credit card transaction record, ranging from 1 to 97,852.

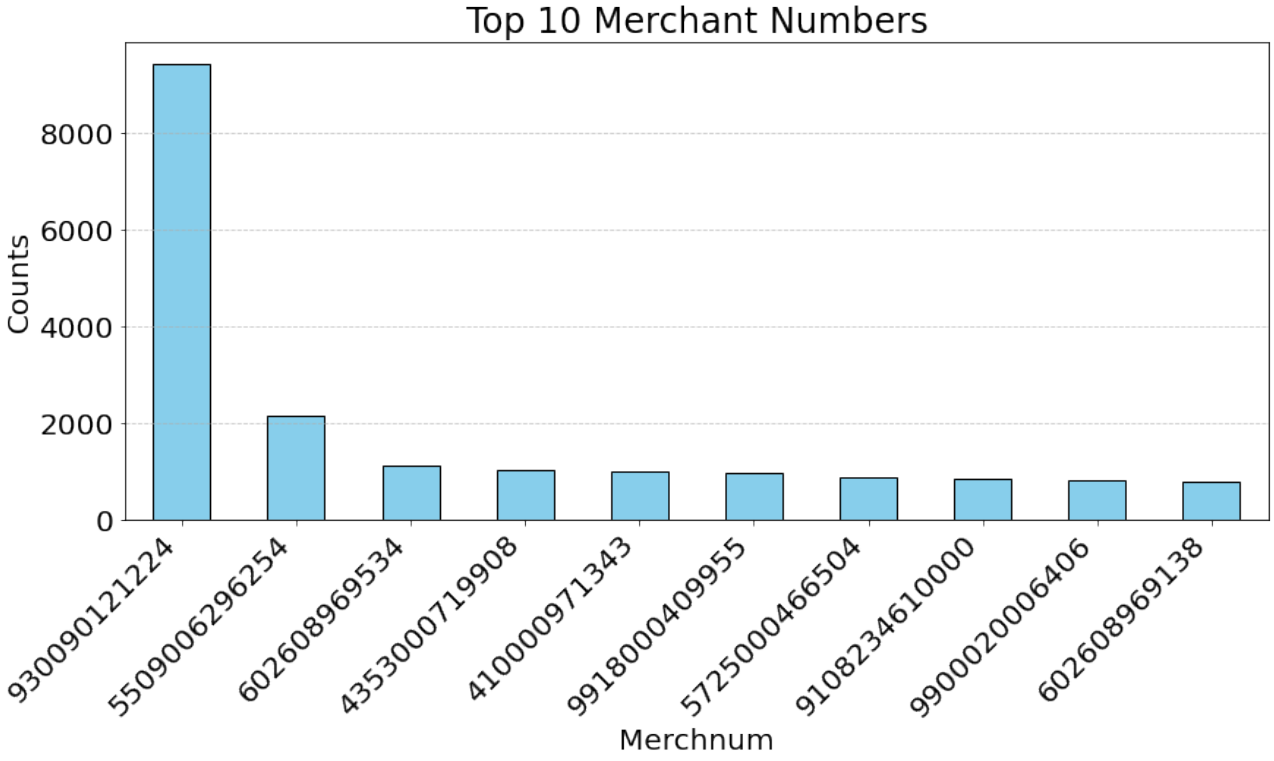
4. **Field Name: Cardnum**
   Description: Identifier for credit card numbers, a numerical field that contains unique values representing individual credit card accounts.



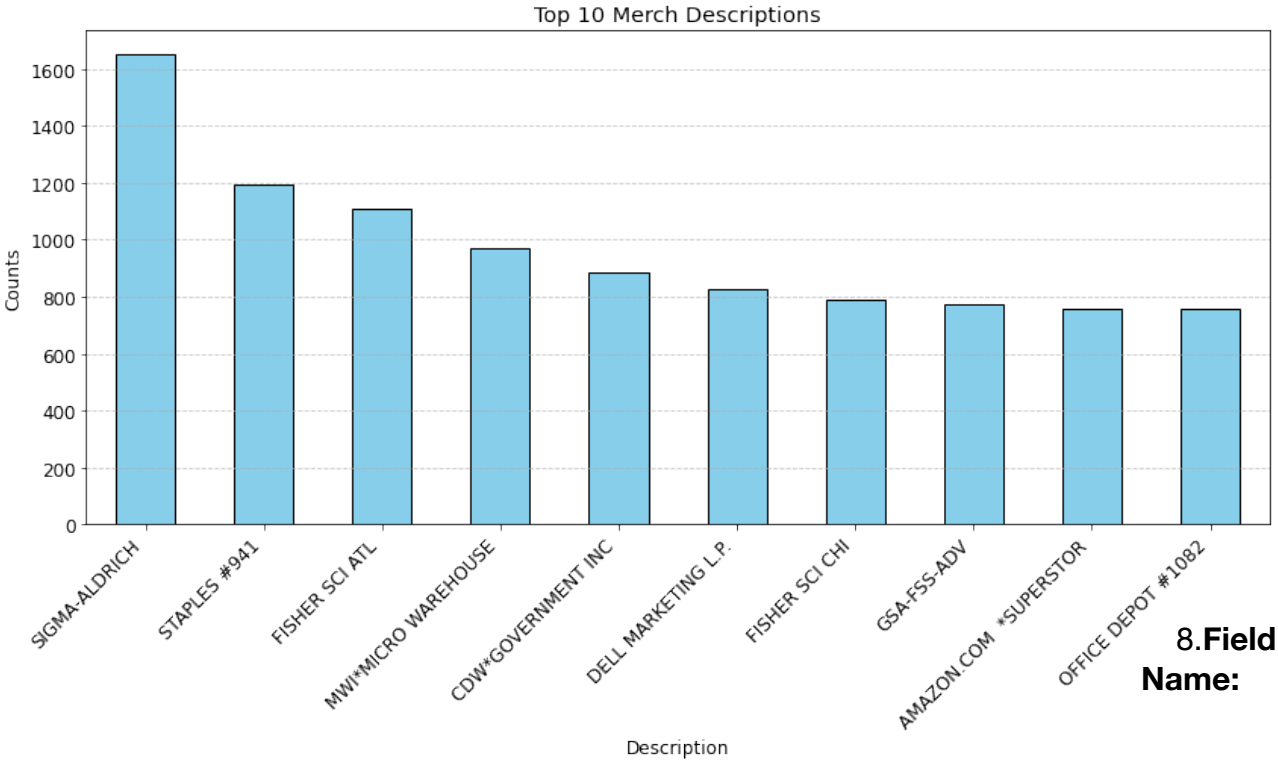
5. **Field Name: Date**
   Description: Transaction date when the credit card activity occurred. This field helps in analyzing transaction volume over time and can be used to create daily or weekly distribution plots to observe temporal trends.



6. **Field Name: Merchnum**
   Description: Merchant identifier number, an alphanumeric field with unique merchant codes. It may contain missing values and is used to identify where the transactions took place.

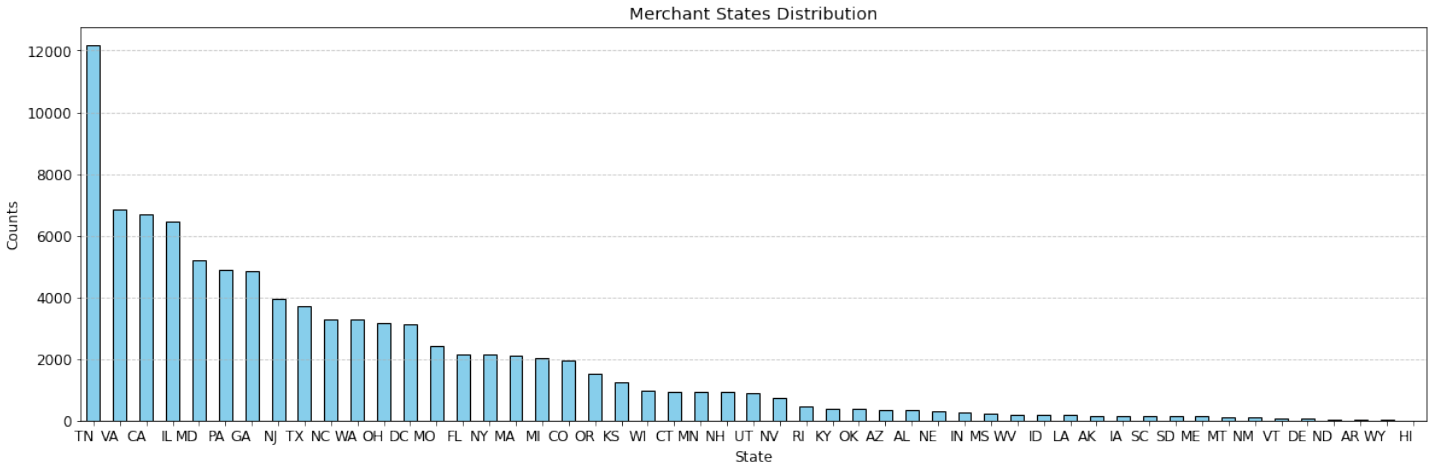


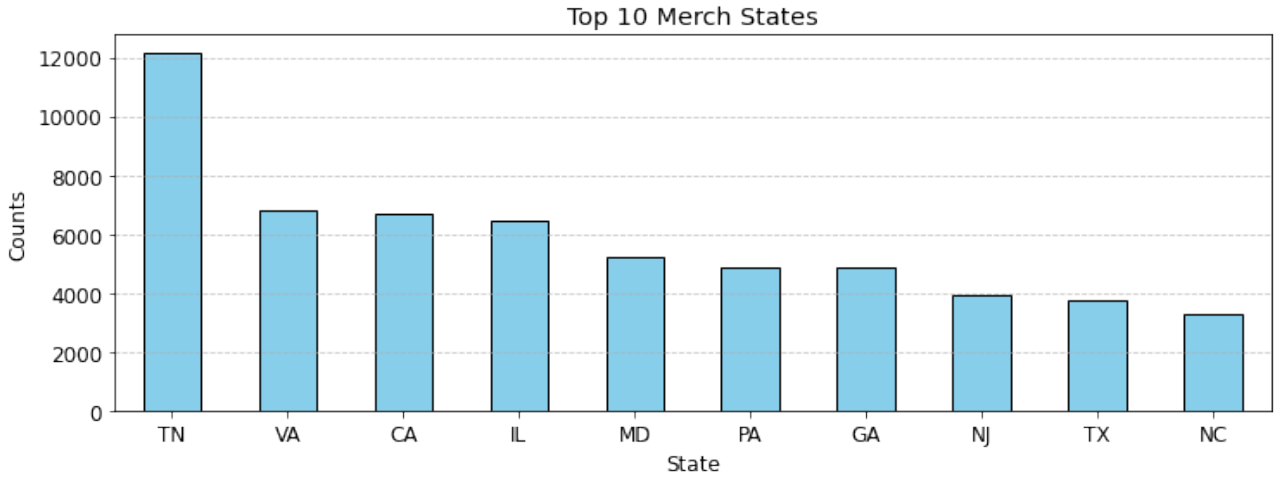
7. **Field Name: Merch description**
   Description: Description of the merchant where the credit card transaction occurred, providing textual information about the business or service provider.



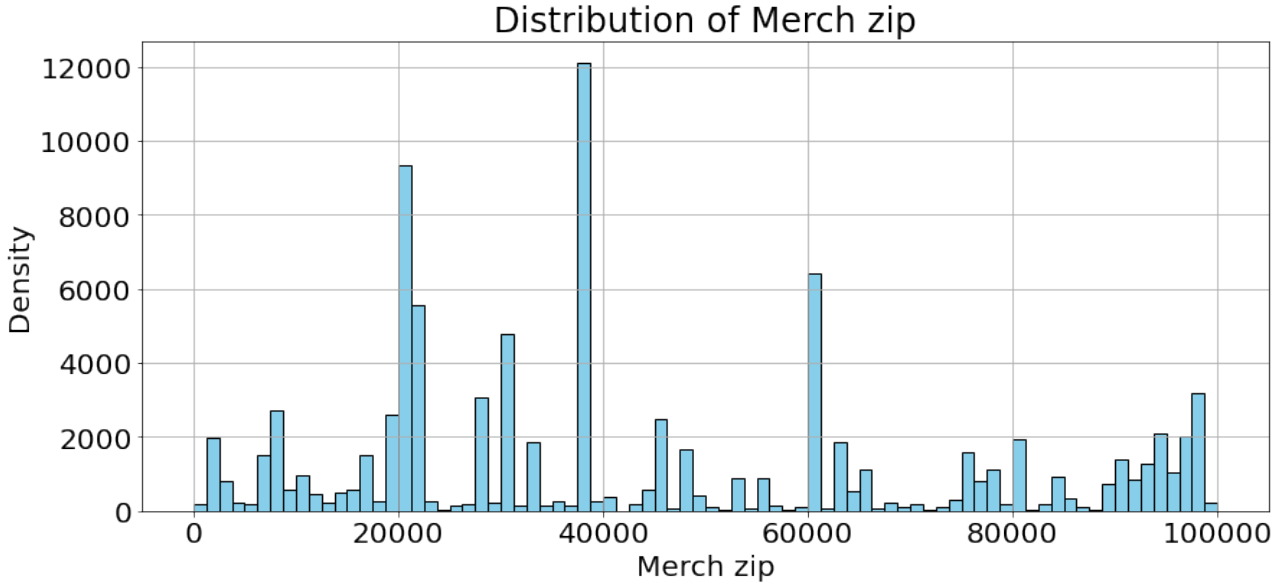
8.**Field Name:**

Merch state
Description: The state in which the merchant is located, a categorical field that allows for regional analysis of transactions across the United States.



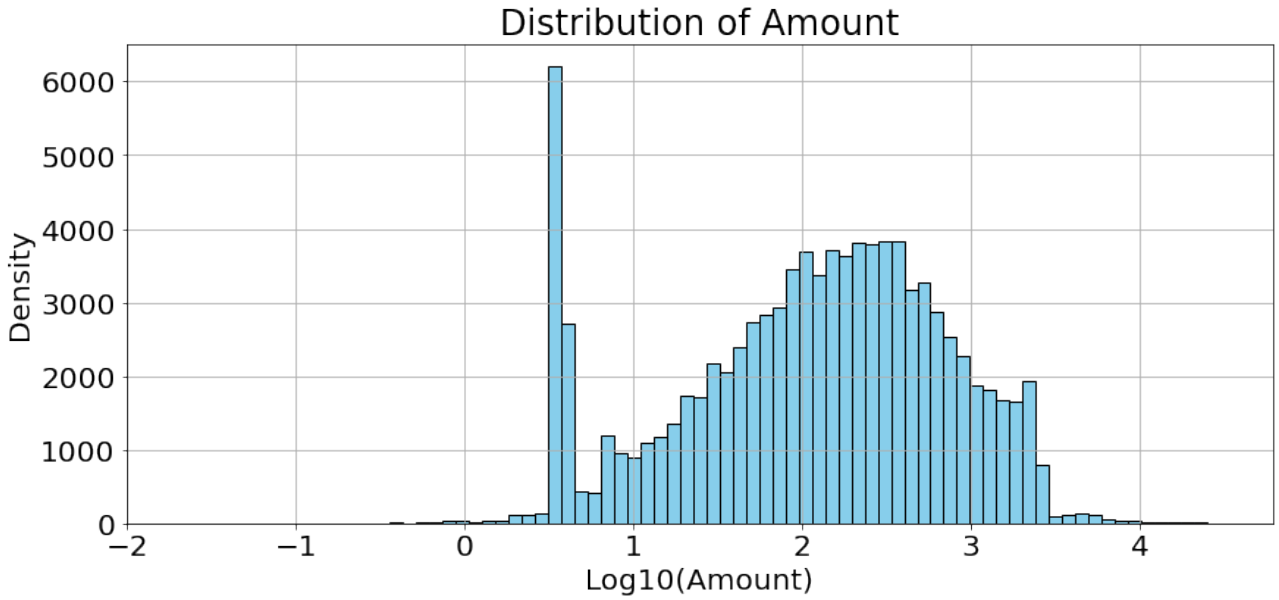


9. **Field Name: Merch zip**
Description: Postal ZIP code of the merchant's location, which is a numerical field that can be used for geospatial analysis of transactions.



10. **Field Name: Transtype**
    Description: Type of transaction performed, currently holding a singular value 'P', indicating a particular type of transaction across the dataset.

11. **Field Name: Amount**
    Description: The amount of the transaction in U.S. dollars, a numerical field critical for identifying and analyzing transaction sizes and patterns.



12. **Field Name: Fraud**
    Description: Binary indicator of whether a transaction was fraudulent ('1') or not ('0'), essential for building predictive models that can detect fraudulent activities. The plot shows the number of fraudulent and non fraudulent transactions.