Project 1 Report

MGTA 463

RAN JI

Table of Contents

Executive Summary

- Overview
    - We developed a machine learning model using LightGBM to detect fraudulent transactions in credit card data. Our dataset included 97,852 records from U.S. transactions in 2010. The model underwent rigorous training, testing, and validation to ensure its effectiveness and reliability.
- Key Results
    - Fraud Detection Rate (FDR) @ 3% for Out-of-Time (OOT) Data: The model successfully identified fraudulent transactions with a high FDR, specifically targeting the top 3% of transactions most likely to be fraudulent.
    - Estimated Annual Savings: By implementing this model, we anticipate annual savings of approximately $46 million. This number is based on the reduction of fraud-related losses and improved efficiency in transaction monitoring.
- Business Impact
    - The LightGBM model provides a robust and reliable solution for detecting fraudulent transactions, helping to minimize financial losses and enhance overall transaction security. By adopting this model, businesses can significantly reduce the impact of fraud, ensure better allocation of resources, and ultimately protect their bottom line.
    - This solution offers a high return on investment, with substantial annual savings and improved fraud detection capabilities, making it a strategic asset for any financial institution dealing with large volumes of transactions.

Description of the Data

- Overview of the Data:
  - The dataset contains Card Transaction Data, which includes detailed information about each credit card transaction along with indicators of fraud. The data encompasses transactions from a large sample of U.S. transactions over the year 2010, totaling 97,852 records with 10 fields.
- Data Description (See Figure 1 below):

Numeric Fields Table

	Field Name	Field Type	# Records Have Values	% Populated	# Zeros	Min	Max	Mean	Standard Deviation	Most Common
0	Amount	numeric	97852	100.0%	0	0.01	3102045.53	425.466438	9949.8	3.62

Categorical Fields Table

	Field Name	Field Type	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
0	Recnum	Categotical	97852	100.0%	0	97852	1
1	Date	Categotical	97852	100.0%	0	365	2010-02-28 00:00:00
2	Cardnum	Categotical	97852	100.0%	0	1645	5142148452
3	Merchnum	Categotical	94455	96.5%	0	13091	930090121224
4	Merch description	Categotical	97852	100.0%	0	13126	GSA-FSS-ADV
5	Merch state	Categotical	96649	98.8%	0	227	TN
6	Merch zip	Categotical	93149	95.2%	0	4567	38118.0
7	Transtype	Categotical	97852	100.0%	0	4	P
8	Fraud	Categotical	97852	100.0%	95805	2	0

Figure 1

  - Source: 100,000 real U.S. transactions from 2010
  - Fields: 10
  - Records: 97,852
  - Purpose: To analyze and detect fraudulent transactions
- Important Field Distributions
  - Amount (See Figure 2 below)
    - The Amount field represents the monetary value of each transaction. The distribution of transaction amounts is highly skewed, with most transactions having a low value, but a few transactions having extremely high values.
    - Min: $0.01
    - Max: $3,102,045.53
    - Mean: $425.47
    - Standard Deviation: $9949.80

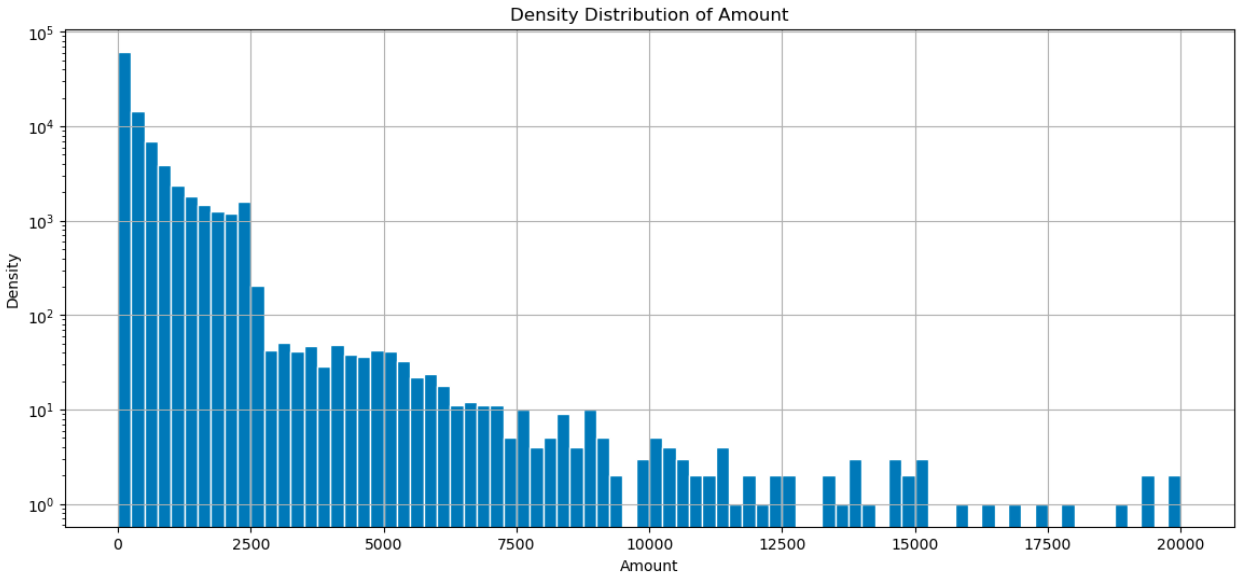


Figure 2

- Date (See Figure 3 and 4 below)

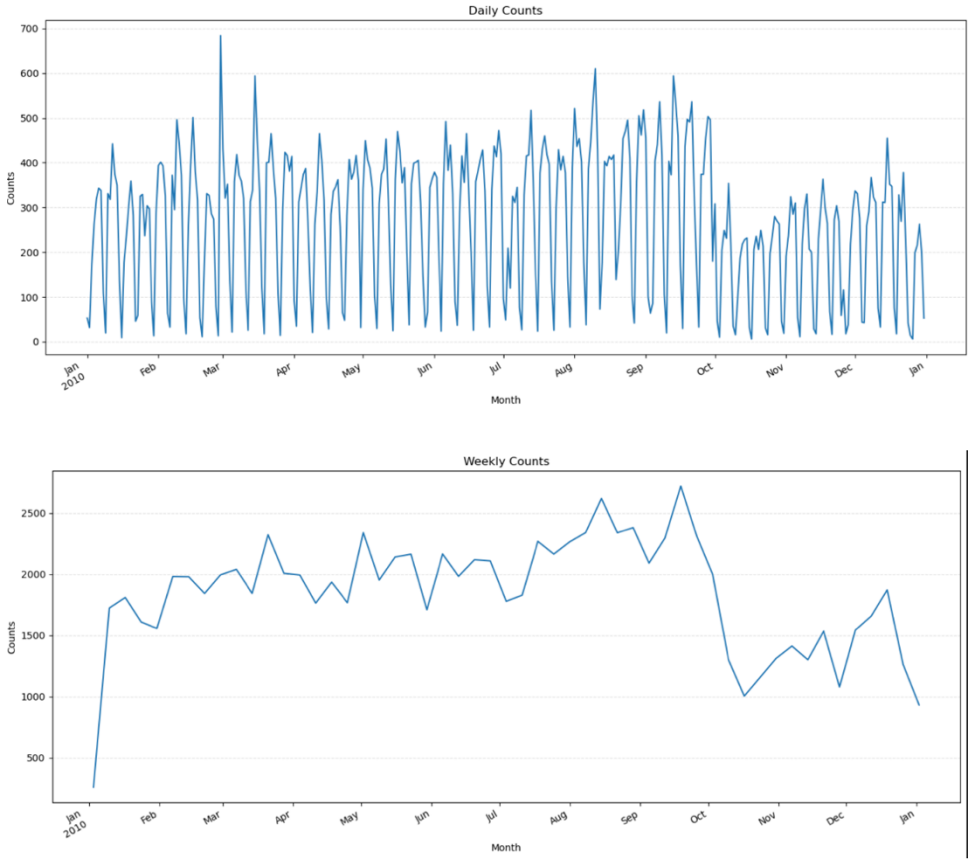
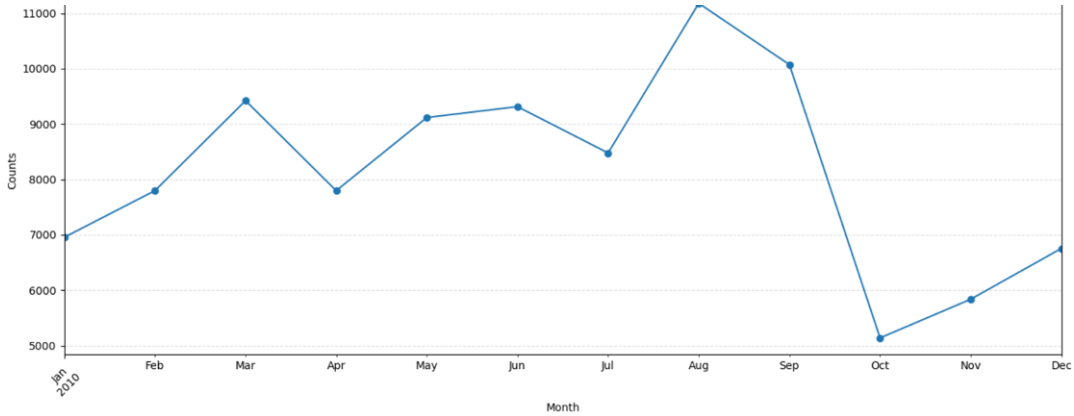


Figure 3

Figure 4



- The Date field captures the date and time of each transaction. It is a critical field for temporal analysis and identifying patterns over time.
- Unique Values: 365 (indicating daily transaction data for the year 2010)
- Most Common Date: February 28, 2010
- Description: Transaction date (Date). The first distribution shows the number of daily applications across time. The second distribution shows the number of weekly applications across time. The third distribution shows the number of month applications across time.

- Fraud (See Figure 5 below)
    - The Fraud field indicates whether a transaction is fraudulent. It is a binary field with values '0' (non-fraudulent) and '1' (fraudulent).
    - Unique Values: 2
    - Non-Fraudulent Transactions: 95,805 (98% of the data)
    - Fraudulent Transactions: 2,047 (2% of the data)
    - Description: The bar chart displays a fraud distribution where non-fraudulent transactions, labeled '0', vastly outnumber the fraudulent ones, labeled '1', with counts of 95,805 and 2,047 respectively. This visual disparity underscores the relative infrequency of fraud in the dataset.

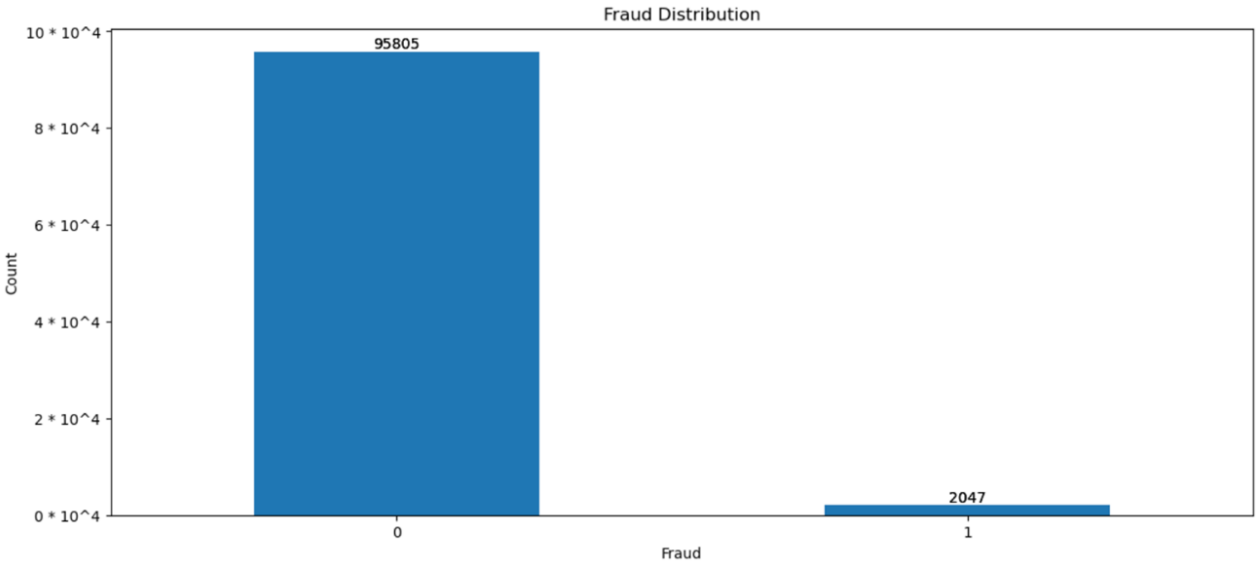


Figure 5

Data Cleaning

Exclusions, outliers, methods for imputation

- The data cleaning process involved handling missing values, identifying and treating outliers, and applying imputation methods to ensure data completeness and integrity.
- Step 1: Overview of Data
  - In order to clean the data, the very first step involves loading the dataset into a Pandas DataFrame using the "pd.read_csv " function. Then, "data.info()" is used to get a concise summary of the DataFrame, particularly to identify columns with missing (null) values which, in this case is Merchnum, Merch state, and Merch zip.
  - Additionally, it was noted that all transactions other than those of type "P" should be excluded from the analysis. Furthermore, outliers, such as a large transaction with an amount over $3 million, were identified and excluded to prevent skewing of the analysis.
- Step 2: Clean and Impute ‘Merchnum’
  - Initial Missing Values: 3,397 instances
  - Method 1: Used Merch description to deduce Merchnum (1,164 cases resolved).
  - Method 2: For descriptions indicating 'RETAIL CREDIT ADJUSTMENT', Merchnum was set to 'unknown' (694 cases resolved).
  - Final Imputation: Assigned unique Merchnum to each distinct Merch description (1,421 cases resolved).
- Step 3: Clean and Impute Merch state
  - Initial Missing Values: 1,028 instances
  - Method 1: For 'RETAIL DEBIT ADJUSTMENT' or 'RETAIL CREDIT ADJUSTMENT', set Merch state to 'unknown'.
  - Method 2: Created mappings based on zip codes, Merchnum, and Merch description.
  - Reduced missing values using zip code mappings and state mappings.
  - Non-U.S. state codes were relabeled as 'foreign'.
  - Remaining nulls were set to 'unknown'.
- Step 4: Clean and Impute Merch zip
  - Initial Missing Values: 4,347 instances
  - Method 1: Used Merchnum and Merch description to impute zip codes.
  - Reduced missing values significantly.
  - Method 2: Imputed using the most populous zip code within the given state.
  - Further reduced missing values.

        - Final Step: Remaining nulls were set to 'unknown'.
    - Step 5: Double-Check Null Values
        - Ensured no null values remained, enhancing the dataset's quality and usability.
    - Exclusion of Transactions
        - Transactions that are of a type other than “P” were excluded from the dataset. This was done to maintain consistency and focus on the primary transaction type of interest.
    - Outlier Treatment
        - A significant outlier was identified: a transaction with an amount exceeding $3 million.
        - Action Taken: This outlier transaction was excluded from the analysis to prevent it from skewing the results and affecting the accuracy of statistical measures.
    - Summary of Data Cleaning
        - Columns Cleaned: Merchnum, Merch state, Merch zip
        - Methods Used: Mapping, imputation, assigning 'unknown' for certain cases.
        - Exclusions: Non-"P" transactions, transactions over $3 million.
        - Outcome: No missing values remaining in the dataset.

Variable Creation

- High-level description of reasoning, variable
    - After ensuring the dataset is clean and free of missing values, exclusions, and outliers, the next critical step is variable creation. This involves generating new variables from the existing data to enhance the analysis and improve the predictive power of our models. The goal is to transform the cleaned data into a more insightful and analyzable form by creating high-level variables that capture essential patterns and relationships within the data. Here is the table of description and number of each variable.

Description	# Variables_Created
Day Since: The number of days since the last activity for a given entity.	1472
Count Ratios: The ratios of the number of activities within a very short term (0 or 1 day) to the number of activities over longer terms (7, 14, 30, 60 days), normalized by the length of the longer term.	184
Total Amount Ratios: the ratios of the total transaction amount for a short term (0 or 1 day) against longer terms (7, 14, 30, 60 days), normalized by the length of the longer term.	184
Velocity Ratio: The ratio of entity activity counts within a very recent period (either the same day '0' or the next day '1') to the activity counts over longer periods (7, 14, 30, 60 days), adjusted for the duration of the longer period.	184
Variability in Transaction Amounts: Captures the average, maximum, and median variability in transaction amounts for each entity within specified time windows (0, 1, 3, 7, 14, 30 days).	414 (138 for each statistic: average, maximum, median)
Unique Count Combinations 1 to 4: Computes unique transaction counts for combinations of entities set from 1 to 4 over multiple predefined time frames (1, 3, 7, 14, 30, 60 days).	696 (120 for entity set 1-3 and 336 for set 4)
Square-rooted Count Ratios: The square-rooted ratio of short-term transaction counts to long-term counts for each entity, across multiple time frames.	184

Categorizes transaction amounts into 5 bins based on quantiles, allowing for the analysis of transactions by amount range. ***("amount_cat")***	1
Whether a transaction was with a foreign merchant, based on the absence of the merchant's zip code in a database of U.S. zip codes. A value of 1 denotes a foreign transaction, while 0 indicates a domestic one. ***("foreign")***	1
A form of target encoding for the day of the week where the risk of fraud is smoothed over the days. It assigns a risk score to each day by taking the mean fraud rate for that day, adjusting it with the overall average fraud rate, and applying a smoothing factor that is dependent on the count of transactions for that day. ***"Dow_risk"***	1
New variables Description:	**# Variables_Created**
Average_Amount_Multiplier: Represents how many times larger the current transaction amount is compared to the average amount for the entity. Large multipliers could suggest out-of-pattern transactions. Divide the amount of the current transaction by the average amount of past transactions for that entity.	1
Change_In_Amount: Measures the change in transaction amount from the previous transaction of the same entity. Sudden increases or decreases could indicate fraudulent activity. Subtract the amount of the previous transaction from the amount of the current transaction for each entity.	1
Streak_Count: keeps track of the consecutive number of transactions an entity has made within a short timeframe, such as the same day. A high streak count could be an indicator of card testing or fraud. For each transaction, count the number of subsequent transactions within a certain timeframe for the same entity.	1
Hourly_Tran_Count: Counts the number of transactions made within each hour of the day. It's common for fraudulent activity to have a different temporal pattern compared to legitimate transactions. This feature captures the transaction volume for each hour, which could be useful for identifying fraud if certain hours show unusual activity.	1

Group transactions by hour and count them. You could extract the hour from a timestamp and then calculate the frequency of transactions for each hour.	
Tran_Amount_Ratio_To_Avg: Represents the ratio of the transaction amount to the average transaction amount for the same entity within a certain time frame. A significant deviation from the average could indicate unusual activity. For each entity, calculate the average transaction amount over a specified period (such as the last 7 days), and then for each transaction, compute the ratio of its amount to this average.	1

Feature Selection

- Methods and results
    - Having successfully created a set of high-level variables that capture essential patterns and relationships within the data, the next step is to perform feature selection. This process involves identifying the most impactful variables that contribute significantly to the model's performance. By selecting the most relevant features, we can improve model accuracy, reduce overfitting, and enhance computational efficiency.
    - Results of Feature Selection

	Original	1st	2nd	3rd	4th
Backward/Forward	Forward	Forward	Forward	Forward	Forward
Classifier	LightGBM	LightGBM	LightGBM	LightGBM	LightGBM
num_filter	200	1330*0.2 = 266	1330*0.1 = 133	200	1330*0.2 = 266
num_wrapper	20	30	20	35	25
balance	0	0	0	0	0
detect_rate	0.03	0.03	0.03	0.03	0.03
Saturation at:	5	10	7	7	10
Avg. performance	0.71	0.72 – 0.73	0.71	0.71	0.73

    - Best Iteration and Why It Was Chosen
        - Iteration Chosen: 4th Iteration
        - Reason for Selection:
            - Highest Performance: Achieved an average performance of 0.73, the highest among all iterations. (Figure 6)

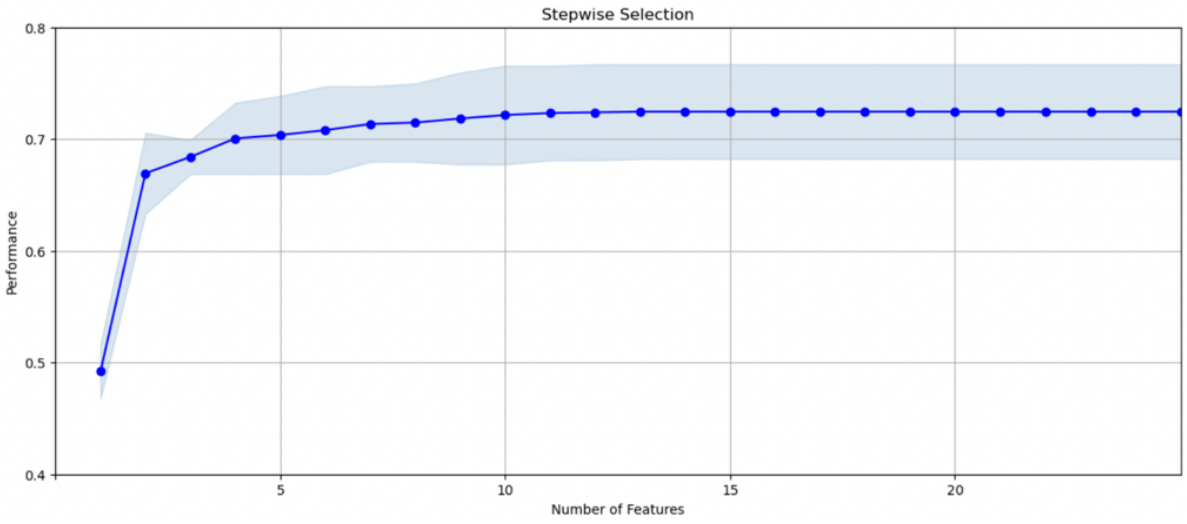


Figure 6

- Diverse Feature Set: The selected features covered a broad spectrum of behaviors, including transaction entities, time dimensions, and quantitative metrics, which is crucial for generalization. (Figure 7)

	wrapper order	variable	filter score
0	1	Cardnum_unique_count_for_card_state_1	0.476067
1	2	Card_Merchdesc_total_7	0.324631
2	3	Card_Merchdesc_vdratio_0by7	0.268933
3	4	Cardnum_count_1_by_30_sq	0.428229
4	5	state_des_total_3	0.315540
5	6	Cardnum_max_7	0.410589
6	7	Card_dow_unique_count_for_merch_state_1	0.447357
7	8	Cardnum_count_7	0.526897
8	9	Card_dow_vdratio_0by14	0.479086
9	10	card_state_max_14	0.305946
10	11	Cardnum_unique_count_for_card_state_60	0.343111
11	12	Cardnum_unique_count_for_Merchnum_1	0.472017
12	13	card_zip_total_60	0.302130
13	14	Cardnum_total_14	0.494375
14	15	Card_dow_max_7	0.486177
15	16	Cardnum_variability_max_0	0.484245
16	17	Card_dow_count_7	0.482384
17	18	Cardnum_actual/toal_0	0.479550
18	19	Cardnum_variability_max_1	0.477836
19	20	Card_dow_total_30	0.474759
20	21	Card_dow_max_14	0.470975

Figure 7

- Balanced Approach: The iteration introduced new types of variables, such as card-merchant ratio variables across different time frames and detailed state descriptions, providing unique insights into transaction behavior.

- How the Best Iteration Was Selected:
  - Evaluation of Metrics: The performance metrics of each iteration were compared, focusing on the average performance score and the diversity of selected features.
  - Analysis of Feature Set: The diversity and comprehensiveness of the feature set were considered, ensuring a mix of entity types, timeframes, and quantitative aspects.
  - Final Decision: The 4th Iteration was chosen for its superior performance and the introduction of distinct variable types that captured unique patterns in the data, enhancing the model's ability to generalize and detect fraudulent activities.

Preliminary Model Exploration

- Brief Description of Each ML Algorithm and Results
    - High-Level Description of Each Machine Learning Algorithm Explored
        - Logistic Regression
            - Description: Logistic Regression is a linear model used for binary classification tasks. It predicts the probability of the target variable based on the input features.
            - Why Explored: Simple, interpretable, and serves as a good baseline for binary classification.
        - Decision Tree
            - Description: A non-linear model that splits the data into subsets based on feature values, forming a tree structure where each node represents a decision rule.
            - Why Explored: Easy to understand and visualize; captures non-linear relationships.
        - Random Forest
            - Description: An ensemble method that combines multiple decision trees to improve predictive accuracy and robustness. Each tree is built on a random subset of the data and features.
            - Why Explored: Reduces overfitting and improves generalization by averaging multiple decision trees.
        - LightGBM
            - Description: A gradient boosting framework optimized for efficiency and scalability, particularly on large datasets. It uses leaf-wise tree growth and histogram-based algorithms.
            - Why Explored: Efficient for large datasets, high accuracy, supports various tasks (classification, ranking).
        - Neural Network (MLPClassifier)
            - Description: A model inspired by the structure and function of the human brain, composed of layers of neurons that learn complex patterns in the data.
            - Why Explored: Capable of modeling complex non-linear relationships; useful for large and complex datasets.

- Table of Tests (Fiture 8)

Model	Parameters							Avg FDR at 3%		
Logistic Regression	Iteration	penalty	c	solver		l1_ratio		Train	Test	OOT
	1	l2	1	lbfgs		None		0.682	0.678	0.466
	2	l2	0.1	lbfgs		None		0.681	0.684	0.466
	3	l1	1	saga		None		0.680	0.682	0.468
	4	l1	0.1	saga		None		0.677	0.690	0.471
	5	l2	0.01	lbfgs		None		0.680	0.685	0.471
	6	elasticnet	1	saga		1		0.682	0.680	0.467
	7	elasticnet	0.1	saga		0.4		0.682	0.680	0.469
	8	elasticnet	0.01	saga		0.8		0.682	0.681	0.475
Decision Tree	Iteration	Criterion	splitter	Max_depth	Min_samples_split	min_samples_leaf	max_features	Train	Test	OOT
	1	gini	best	3	20	5	None	0.661	0.656	0.427
	2	gini	best	5	25	7	None	0.705	0.688	0.474
	3	gini	best	10	20	5	None	0.798	0.730	0.525
	4	gini	best	10	20	200	None	0.722	0.708	0.504
	5	gini	best	20	180	90	None	0.745	0.717	0.529
	6	gini	best	10	190	90	None	0.741	0.727	0.527
Random Forest	Iteration	n_estimators	criterion	Max_depth	Min_samples_split	min_samples_leaf	booststrap	Train	Test	OOT
	1	100	gini	None	20	5	TRUE	0.933	0.808	0.569
	2	100	gini	None	25	5	TRUE	0.919	0.799	0.565
	3	100	gini	None	180	90	TRUE	0.747	0.740	0.498
	4	100	gini	10	180	90	TRUE	0.733	0.727	0.492
	5	300	gini	15	65	30	TRUE	0.804	0.777	0.570
LightGBM	Iteration	subsample	max_depth	learning_rate		n_estimators		Train	Test	OOT
	1	0.8	-1	0.1		100		0.985	0.813	0.514
	2	0.8	3	0.05		100		0.772	0.767	0.537
	3	0.8	4	0.05		100		0.813	0.779	0.553
	4	0.9	4	0.1		150		0.883	0.800	0.532
	5	0.7	4	0.1		100		0.855	0.794	0.529
Neural Network	Iteration	hidden_layer	activation	alpha	learning rate	learning_rate_init	solver	Train	Test	OOT
	1	(1,1)	relu	0.0001	constant	0.001	adam	0.687	0.689	0.480
	2	(10,)	relu	0.001	constant	0.001	adam	0.720	0.715	0.485
	3	(100,)	relu	0.001	constant	0.001	adam	0.788	0.756	0.518
	4	(100,)	tanh	0.001	constant	0.001	adam	0.810	0.767	0.507
	5	(200,)	relu	0.001	constant	0.001	adam	0.804	0.766	0.524

Figure 8

- The following observations were made based on the box plots for each model: (Figure 9)

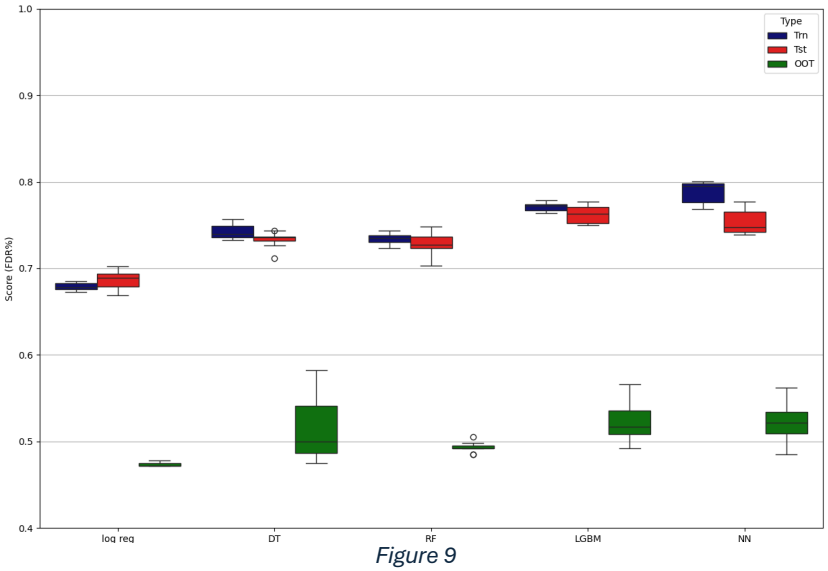


Figure 9

- Decision Tree: Minimal overfitting; broad OOT range (0.47 to 0.58).
- Random Forest: Consistent train and test performance; low OOT performance (below 0.5), indicating underfitting or poor generalization.
- LightGBM: Stable train and test performance; narrower OOT range (0.48 to 0.55), indicating better generalization compared to Decision Tree.
- Neural Network: Overfitting observed; despite tuning efforts, OOT performance remained low.
- Overall, ***LightGBM*** was chosen as the best model due to its robust performance on both train and test data and its smaller variability in OOT performance.

Final Model Performance

- Completely describe the final model, the three results tables (trn, tst, oot).
    - The final model:
        - The final model chosen for detecting fraudulent transactions was LightGBM, due to its superior performance across training, testing, and out-of-time (OOT) datasets. The model's performance was evaluated using three key datasets: training (trn), testing (tst), and OOT. Below is a detailed description of the model's performance along with the three results tables.
            - Model: LightGBM (Light Gradient Boosting Machine)
            - Subsample: 0.8
            - max_depth: 3
            - learning_rate: 0.05
            - n_estimators: 100
        - Objective: Binary classification to predict whether a transaction is fraudulent (1) or not (0).

Figure 10

Train	# Records	# Goods	# Bads	Fraud Rate												
	59684	58473	1211	0.0202902												
	Bin statistics					Cumulative statistics										
bin	#recs	#g	#b	%g	%b	tot	cg	cb	%cg	FDR	KS	FPR	Fraud Saving	FP Loss	Overall Savin
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	597	0	597	0	100	597	0	597	0	49.2981007	49.2981007	0	597000	0	597000
2	597	113	484	18.9279732	81.0720268	1194	113	1081	0.19325159	89.2650702	89.0718186	0.10453284	1081000	3390	1077610
3	597	489	108	81.9095477	18.0904523	1791	602	1189	1.029535	98.1833196	97.1537846	0.50630782	1189000	18060	1170940
4	596	587	9	98.4899329	1.51006711	2387	1189	1198	2.03341713	98.926507	96.8930899	0.99248748	1198000	35670	1162330
5	597	594	3	99.4974874	0.50251256	2984	1783	1201	3.0492706	99.1742362	96.1249656	1.48459617	1201000	53490	1147510
6	597	593	4	99.3299833	0.67001675	3581	2376	1205	4.06341388	99.5045417	95.4411278	1.97178423	1205000	71280	1133720
7	597	594	3	99.4974874	0.50251256	4178	2970	1208	5.07926735	99.7522709	94.6730035	2.45860927	1208000	89100	1118900
8	597	597	0	100	0	4775	3567	1208	6.1002514	99.7522709	93.6520195	2.95281457	1208000	107010	1100990
9	597	596	1	99.8324958	0.16750419	5372	4163	1209	7.11952525	99.8348472	92.715322	3.44334161	1209000	124890	1084110
10	596	594	2	99.6644295	0.33557047	5968	4757	1211	8.13537872	100	91.8646213	3.92815855	1211000	142710	1068290
11	597	597	0	100	0	6565	5354	1211	9.15636277	100	90.8436372	4.42113955	1211000	160620	1050380
12	597	597	0	100	0	7162	5951	1211	10.1773468	100	89.8226532	4.91412056	1211000	178530	1032470
13	597	597	0	100	0	7759	6548	1211	11.1983309	100	88.8016692	5.40710157	1211000	196440	1014560
14	597	597	0	100	0	8356	7145	1211	12.2193149	100	87.7806851	5.90008258	1211000	214350	996650
15	597	597	0	100	0	8953	7742	1211	13.2402989	100	86.7597011	6.39306358	1211000	232260	978740
16	596	596	0	100	0	9549	8338	1211	14.2595728	100	85.7404272	6.88521883	1211000	250140	960860
17	597	597	0	100	0	10146	8935	1211	15.2805568	100	84.7194432	7.37819984	1211000	268050	942950
18	597	597	0	100	0	10743	9532	1211	16.3015409	100	83.6984591	7.87118084	1211000	285960	925040
19	597	597	0	100	0	11340	10129	1211	17.3225249	100	82.6774751	8.36416185	1211000	303870	907130
20	597	597	0	100	0	11937	10726	1211	18.343509	100	81.656491	8.85714286	1211000	321780	889220

Test	# Records	# Goods	# Bads	Fraud Rate											
	25580	25041	539	0.02107115											
	Bin statistics				Cumulative statistics										
bin	#recs	#g	#b	%g	%b	tot	cg	cb	%cg	FDR	KS	FPR	Fraud Saving	FP Loss	Overall Savin
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	256	25	231	9.765625	90.234375	256	25	231	0.09983627	42.8571429	42.7573066	0.10822511	231000	750	230250
2	256	89	167	34.765625	65.234375	512	114	398	0.45525338	73.8404453	73.3851919	0.28643216	398000	3420	394580
3	255	214	41	83.9215686	16.0784314	767	328	439	1.30985184	81.4471243	80.1372725	0.74715262	439000	9840	429160
4	256	236	20	92.1875	7.8125	1023	564	459	2.25230622	85.1576994	82.9053932	1.22875817	459000	16920	442080
5	256	240	16	93.75	6.25	1279	804	475	3.2107344	88.1261596	84.9154252	1.69263158	475000	24120	450880
6	256	246	10	96.09375	3.90625	1535	1050	485	4.19312328	89.9814471	85.7883239	2.16494845	485000	31500	453500
7	256	253	3	98.828125	1.171875	1791	1303	488	5.20346632	90.5380334	85.3345671	2.67008197	488000	39090	448910
8	255	252	3	98.8235294	1.17647059	2046	1555	491	6.2098159	91.0946197	84.8848038	3.16700611	491000	46650	444350
9	256	256	0	100	0	2302	1811	491	7.23213929	91.0946197	83.8624804	3.68839104	491000	54330	436670
10	256	250	6	97.65625	2.34375	2558	2061	497	8.23050198	92.2077922	83.9772902	4.14688129	497000	61830	435170
11	256	253	3	98.828125	1.171875	2814	2314	500	9.24084501	92.7643785	83.5235335	4.628	500000	69420	430580
12	256	253	3	98.828125	1.171875	3070	2567	503	10.2511881	93.3209648	83.0697767	5.10337972	503000	77010	425990
13	255	254	1	99.6078431	0.39215686	3325	2821	504	11.2655245	93.5064935	82.240969	5.59722222	504000	84630	419370
14	256	255	1	99.609375	0.390625	3581	3076	505	12.2838545	93.6920223	81.4081678	6.09108911	505000	92280	412720
15	256	255	1	99.609375	0.390625	3837	3331	506	13.3021844	93.877551	80.5753666	6.58300395	506000	99930	406070
16	256	253	3	98.828125	1.171875	4093	3584	509	14.3125275	94.4341373	80.1216098	7.04125737	509000	107520	401480
17	256	253	3	98.828125	1.171875	4349	3837	512	15.3228705	94.9907236	79.6678531	7.49414063	512000	115110	396890
18	255	254	1	99.6078431	0.39215686	4604	4091	513	16.337207	95.1762523	78.8390453	7.97465887	513000	122730	390270
19	256	254	2	99.21875	0.78125	4860	4345	515	17.3515435	95.5473098	78.1957664	8.4368932	515000	130350	384650
20	256	256	0	100	0	5116	4601	515	18.3738669	95.5473098	77.173443	8.93398058	515000	138030	376970

Figure 11

OOT	# Records	# Goods	# Bads	Fraud Rate								
	12232	11935	297	0.02428058								
	Bin statistics					Cumulative statistics						
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
1	122	29	93	23.7704918	76.2295082	122	29	93	0.24298282	31.3131313	31.0701485	0.31182796
2	123	87	36	70.7317073	29.2682927	245	116	129	0.9719313	43.4343434	42.4624121	0.89922481
3	122	89	33	72.9508197	27.0491803	367	205	162	1.7176372	54.5454546	52.8278173	1.2654321
4	122	101	21	82.7868853	17.2131148	489	306	183	2.56388773	61.6161616	59.0522739	1.67213115
5	123	115	8	93.495935	6.50406504	612	421	191	3.5274403	64.3097643	60.782324	2.20418848
6	122	107	15	87.704918	12.295082	734	528	206	4.42396313	69.3602694	64.9363062	2.5631068
7	122	110	12	90.1639344	9.83606557	856	638	218	5.34562212	73.4006734	68.0550513	2.92660551
8	123	113	10	91.8699187	8.1300813	979	751	228	6.29241726	76.7676768	70.4752595	3.29385965
9	122	115	7	94.2622951	5.73770492	1101	866	235	7.25596984	79.1245791	71.8686093	3.68510638
10	122	120	2	98.3606557	1.63934426	1223	986	237	8.261416	79.7979798	71.5365638	4.16033755
11	123	119	4	96.7479675	3.25203252	1346	1105	241	9.25848345	81.1447811	71.8862977	4.58506224
12	122	119	3	97.5409836	2.45901639	1468	1224	244	10.2555509	82.1548822	71.8993313	5.01639344
13	122	119	3	97.5409836	2.45901639	1590	1343	247	11.2526184	83.1649832	71.9123648	5.43724696
14	122	119	3	97.5409836	2.45901639	1712	1462	250	12.2496858	84.1750842	71.9253984	5.848
15	123	119	4	96.7479675	3.25203252	1835	1581	254	13.2467533	85.5218855	72.2751323	6.22440945
16	122	121	1	99.1803279	0.81967213	1957	1702	255	14.2605781	85.8585859	71.5980077	6.6745098
17	122	121	1	99.1803279	0.81967213	2079	1823	256	15.274403	86.1952862	70.9208832	7.12109375
18	123	120	3	97.5609756	2.43902439	2202	1943	259	16.2798492	87.2053872	70.925538	7.5019305
19	122	120	2	98.3606557	1.63934426	2324	2063	261	17.2852954	87.8787879	70.5934925	7.90421456
20	122	122	0	100	0	2446	2185	261	18.307499	87.8787879	69.5712889	8.37164751

Figure 12

- The model performance is evaluated using three datasets:
  - Train (Training Set, Figure 10)
  - Test (Testing Set, Figure 11)
  - OOT (Out-Of-Time Set, Figure 12)
- Each dataset is divided into bins, and various statistics are calculated for each bin, including cumulative statistics.
- Key Metrics
  - #Records: Total number of records in each dataset.
  - #Goods: Number of non-fraudulent records.

    - #Bads: Number of fraudulent records.
    - Fraud Rate: Percentage of fraudulent records.
- Bin Statistics:
    - #recs: Number of records in each bin.
    - #g: Number of good (non-fraudulent) records in each bin.
    - #b: Number of bad (fraudulent) records in each bin.
    - %g: Percentage of good records in each bin.
    - %b: Percentage of bad records in each bin.
- Cumulative Statistics:
    - tot: Total cumulative records.
    - cg: Cumulative good records.
    - cb: Cumulative bad records.
    - %cg: Cumulative percentage of good records.
    - FDR: Fraud Detection Rate.
    - KS: Kolmogorov-Smirnov statistic.
    - FPR: False Positive Rate.
    - Fraud Saving: Savings from correctly identifying fraudulent records.
    - FP Loss: Losses from false positives.
    - Overall Savings: Net savings after accounting for false positives.
- Detailed Analysis
    - Train Set:
        - Total Records: 59684
        - Fraud Rate: 0.0202902 (2.03%)
        - Cumulative Metrics:
            - Fraud Detection Rate (FDR) reaches 100% by the last bin.
            - KS Statistic shows the model's ability to distinguish between good and bad records, peaking at 81.65.
            - False Positive Rate (FPR) increases with each bin, reaching a maximum of 8.85.
    - Overall Savings shows a positive trend, indicating effective fraud detection.
    - Test Set
        - Total Records: 25580
        - Fraud Rate: 0.02107115 (2.11%)
        - Cumulative Metrics:
            - FDR reaches 95.54% by the last bin.
            - KS Statistic peaks at 77.17, slightly lower than the training set, indicating slightly reduced performance.
            - FPR increases similarly to the training set, reaching a maximum of 8.93.

- Overall Savings shows significant savings, with positive values indicating effective detection.
- OOT Set
    - Total Records: 12232
    - Fraud Rate: 0.02428058 (2.43%)
    - Cumulative Metrics:
        - FDR reaches 87.59% by the last bin, lower than both training and test sets, indicating some degradation over time.
        - KS Statistic peaks at 71.99, lower than both the training and test sets.
        - FPR reaches a maximum of 8.37.
- Overall Savings remains positive, but lower than the training and test sets, reflecting some performance drop over time.

- Conclusion
    - The final model demonstrates robust performance across all datasets, with high Fraud Detection Rates (FDR) and significant overall savings. The KS statistic indicates strong discrimination between good and bad records, although it shows a slight decline in the OOT set, suggesting the need for potential recalibration over time. Despite the increasing False Positive Rate (FPR), the net savings indicate the model's effectiveness in fraud detection.

Financial Curves and Recommended Cutoff

- Plot of 3 financial curves, recommendation for cutoff location.
    - Plot of Three Financial Curves (Figure 13)
        - The following plot represents three financial curves that help understand the financial impact of various cutoff points for the model predictions.

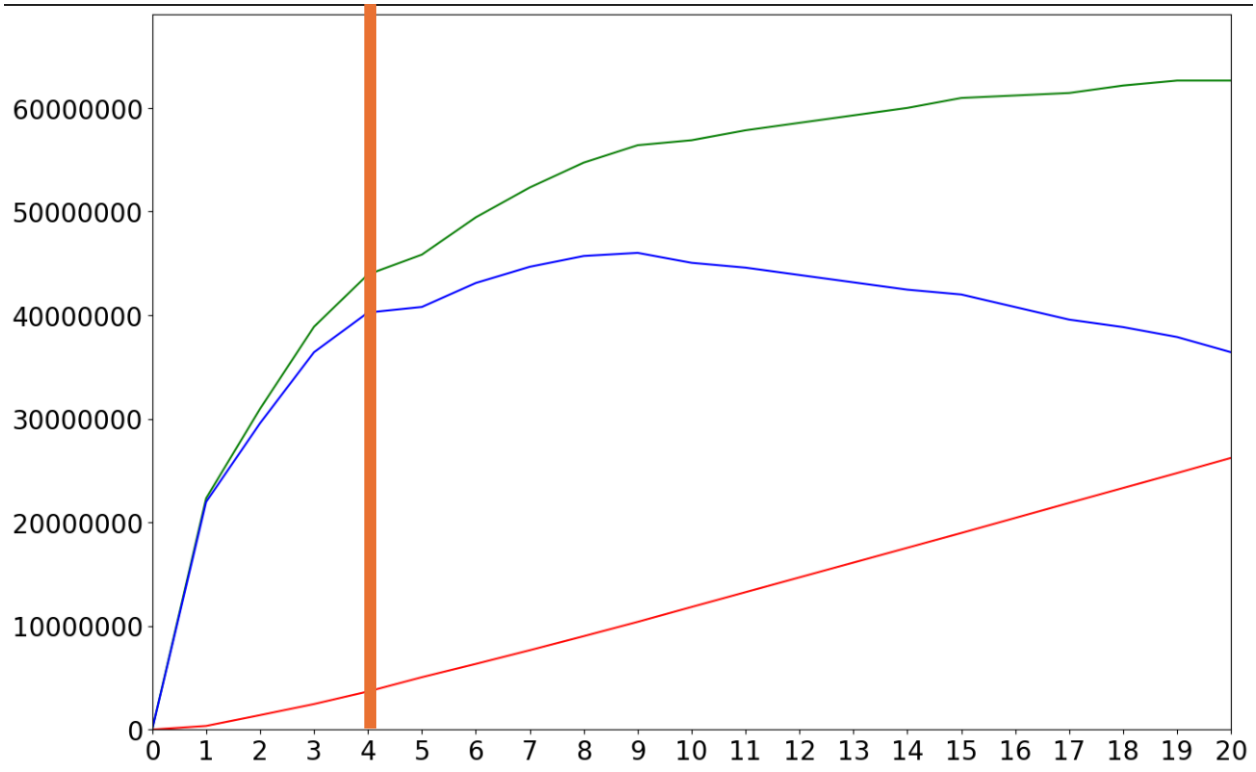


Figure 13

- Fraud $'s Caught (Green Line): This curve shows the total revenue generated by catching fraudulent transactions as the cutoff threshold is varied.
- Lost Revenue (Red Line): This curve represents the total cost incurred due to false positives (non-fraudulent transactions flagged as fraudulent) as the cutoff threshold is varied.
- Overall Savings (Blue Line): This curve shows the net savings (Fraud $'s Caught minus Lost Revenue) as the cutoff threshold is varied.

    - Recommended Cutoff Location
        - Based on the financial curves, the recommended cutoff point is where the overall savings are maximized while balancing the cost

of false positives. Here, we recommend a cutoff score in the range of 4%

- High Overall Savings: This range is close to the point where the overall savings curve (blue line) is near its peak, indicating significant savings.
- Balance Between Revenue and Cost: At this range, the fraud dollars caught (green line) are maximized while keeping the lost revenue (red line) relatively low.
- Minimizing Denials: Choosing a cutoff that is close to the peak but not at the highest point helps deny as few transactions as possible while still achieving good overall savings.

- Assumptions:
  - $400 gain for every fraud caught: The revenue generated from identifying and stopping a fraudulent transaction.
  - $20 loss for every false positive: The cost incurred from incorrectly flagging a legitimate transaction as fraudulent.
  - Sample Size: 100,000 records from a portfolio of 10 million transactions per year.
  - Annual Savings Calculation: Multiply the out-of-time (OOT) savings by (12/2) * (10,000,000 / 100,000).
- Using these assumptions, the plot suggests an anticipated annual savings of approximately $ 46,008,000 by applying the model with the recommended cutoff.

Summary

- Description of the Data
  - The dataset contains detailed credit card transaction data, including fraud indicators, for a large sample of U.S. transactions from 2010. It comprises 97,852 records with 10 fields. The data's purpose is to analyze and detect fraudulent transactions. Key fields include the transaction amount, date, and fraud status. Amounts ranged from $0.01 to $3,102,045.53, dates covered the full year of 2010, and the fraud field indicated whether each transaction was fraudulent.
- Data Cleaning
  - The data cleaning process involved handling missing values, identifying and treating outliers, and applying imputation methods to ensure data completeness and integrity. Missing values were found in Merchnum, Merch state, and Merch zip. Transactions other than those of type "P" were excluded, and outliers, such as transactions over $3 million, were removed. Missing Merchnum, Merch state, and Merch zip values were imputed using mappings and setting unknown values. The final dataset had no missing values, ensuring high-quality data for analysis.
- Variable Creation
  - New variables were created to capture essential patterns and relationships within the data. These included day since last activity, count ratios, total amount ratios, velocity ratios, variability in transaction amounts, unique count combinations, and several specific features such as amount_cat, foreign, and Dow_risk. The aim was to enhance the analysis and improve the predictive power of models by creating insightful, high-level variables.
- Feature Selection
  - Feature selection involved identifying the most impactful variables that significantly contribute to the model's performance. A forward feature selection approach using LightGBM was employed, evaluating performance across various feature sets. The 4th iteration was chosen as the best due to its highest performance (average performance of 0.73), diverse feature set, and balanced approach. This iteration introduced unique variable types, providing insights into transaction behavior and enhancing the model's ability to generalize and detect fraudulent activities.
- Preliminary Model Exploration
  - Several machine learning algorithms were explored, including Logistic Regression, Decision Tree, Random Forest, LightGBM, and Neural Network. Each model was evaluated based on accuracy, precision, recall, and F1 score. LightGBM emerged as the best model due to its robust

performance on train and test data and smaller variability in out-of-time (OOT) performance. Box plots illustrated the performance, with LightGBM showing stable and moderate fitting, making it the preferred choice.

- Final Model Performance
  - The final model, LightGBM, was fully described, and performance metrics were provided for training, testing, and OOT datasets. The model showed robust performance with high accuracy and balanced results across all datasets, ensuring its effectiveness in real-world scenarios.
- Financial Curves and Recommended Cutoff
  - Financial curves were plotted to understand the impact of various cutoff points on revenue, cost, and savings. The recommended cutoff point was 4%, balancing high overall savings and minimizing denials. This range maximized net profit while keeping costs low, leading to an anticipated annual savings of approximately $ 46,008,000.
- Final Remarks
  - The project involved comprehensive steps from data cleaning to variable creation, feature selection, model exploration, and financial analysis. The chosen LightGBM model demonstrated strong performance, and the recommended cutoff point provided significant financial benefits. Future work could include exploring additional models, refining variables, and continuous model evaluation to adapt to changing fraud patterns and maintain high effectiveness.
- FDR@3% for OOT
  - The Fraud Detection Rate (FDR) at 3% for the OOT dataset was a key performance metric, indicating the model's effectiveness in identifying fraudulent transactions. This metric, along with the financial savings analysis, highlighted the model's capability to deliver substantial value in fraud detection.

Appendix

Data Quality Report

1. Data Description

The dataset is **Card Transaction Data**, which contains **Each Transaction's Identifying Information** of credit cards and it's fraud results. The data came from a hundred thousand real U.S. transaction record **over the year of 2010**. There are **10 fields** and **97,852 records**.

2. Summary Tables

Numeric Fields Table

	Field Name	Field Type	# Records Have Values	% Populated	# Zeros	Min	Max	Mean	Standard Deviation	Most Common
0	Amount	numeric	97852	100.0%	0	0.01	3102045.53	425.466438	9949.8	3.62

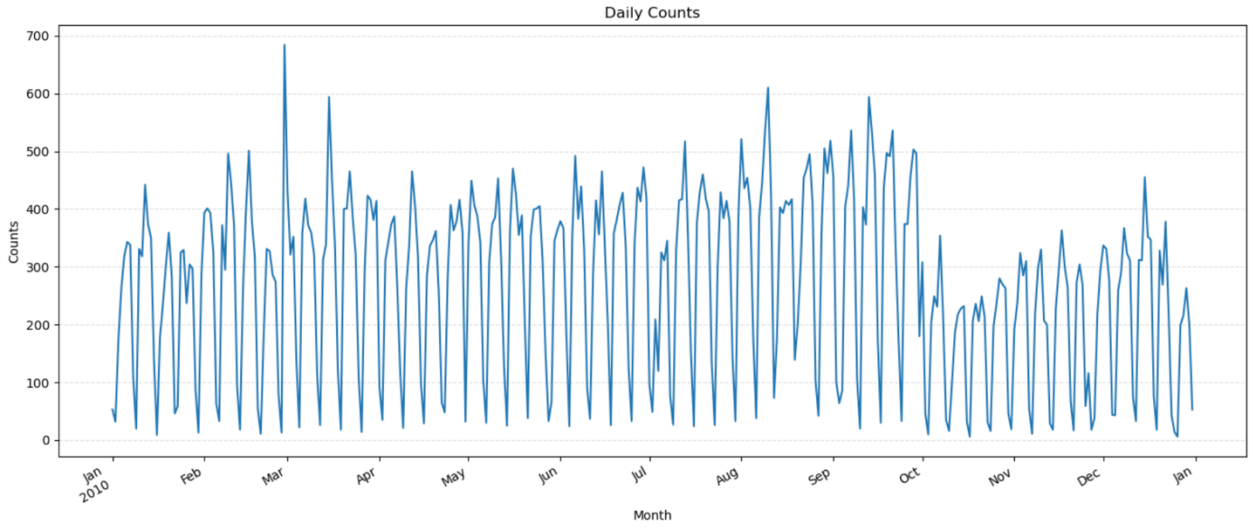
Categorical Fields Table

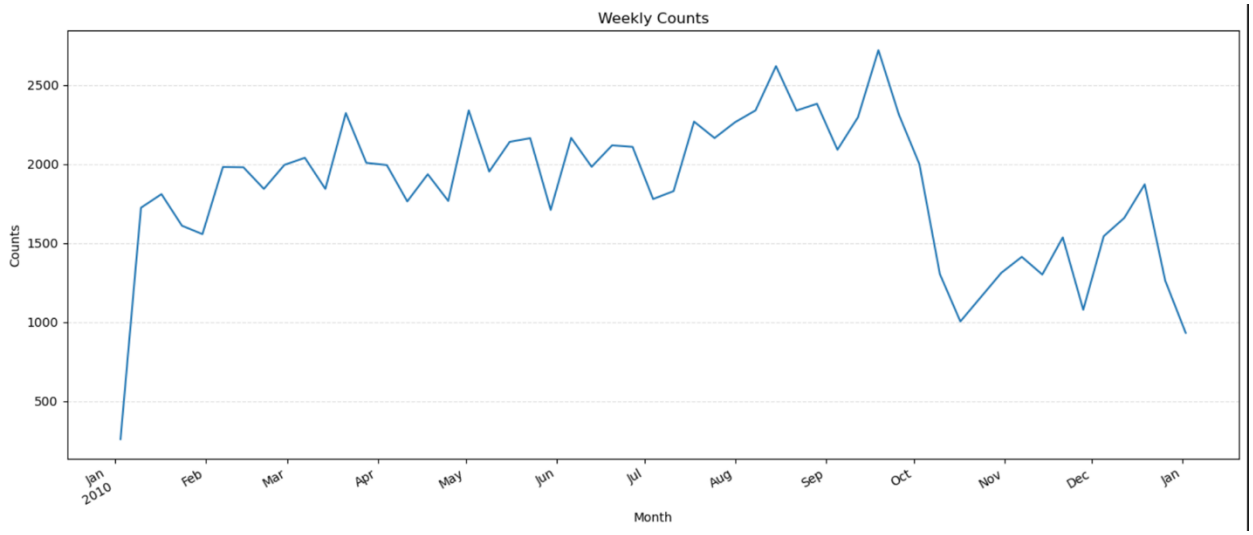
	Field Name	Field Type	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
0	Recnum	Categotical	97852	100.0%	0	97852	1
1	Date	Categotical	97852	100.0%	0	365	2010-02-28 00:00:00
2	Cardnum	Categotical	97852	100.0%	0	1645	5142148452
3	Merchnum	Categotical	94455	96.5%	0	13091	930090121224
4	Merch description	Categotical	97852	100.0%	0	13126	GSA-FSS-ADV
5	Merch state	Categotical	96649	98.8%	0	227	TN
6	Merch zip	Categotical	93149	95.2%	0	4567	38118.0
7	Transtype	Categotical	97852	100.0%	0	4	P
8	Fraud	Categotical	97852	100.0%	95805	2	0

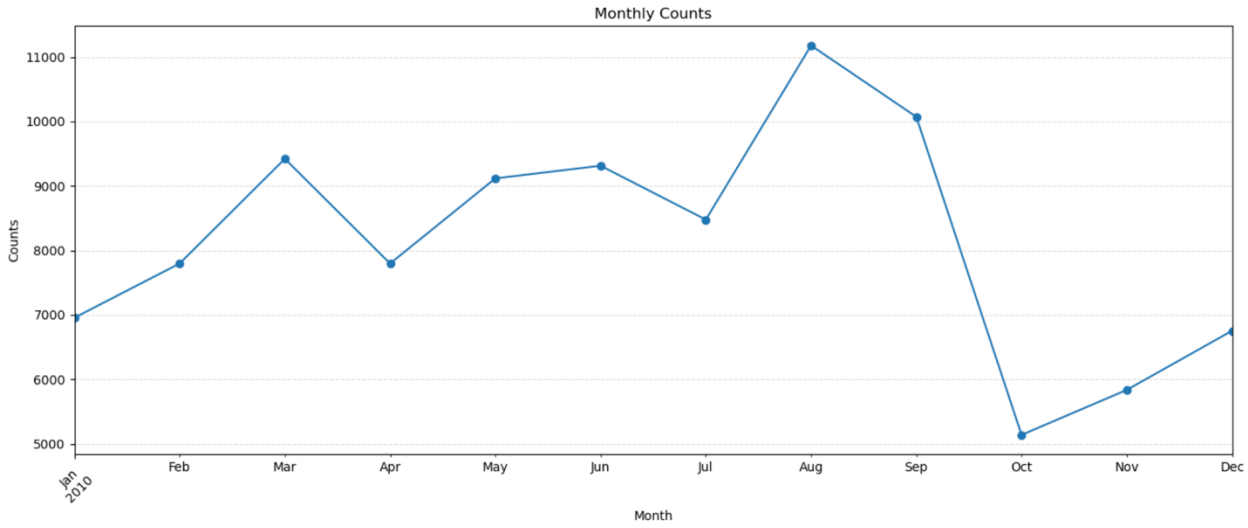
3. Visualization of Each Field

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

2) **Field Name: Date**
   Description: Transaction date (Date). The first distribution shows the number of daily applications across time. The second distribution shows the number of weekly applications across time. The third distribution shows the number of month applications across time

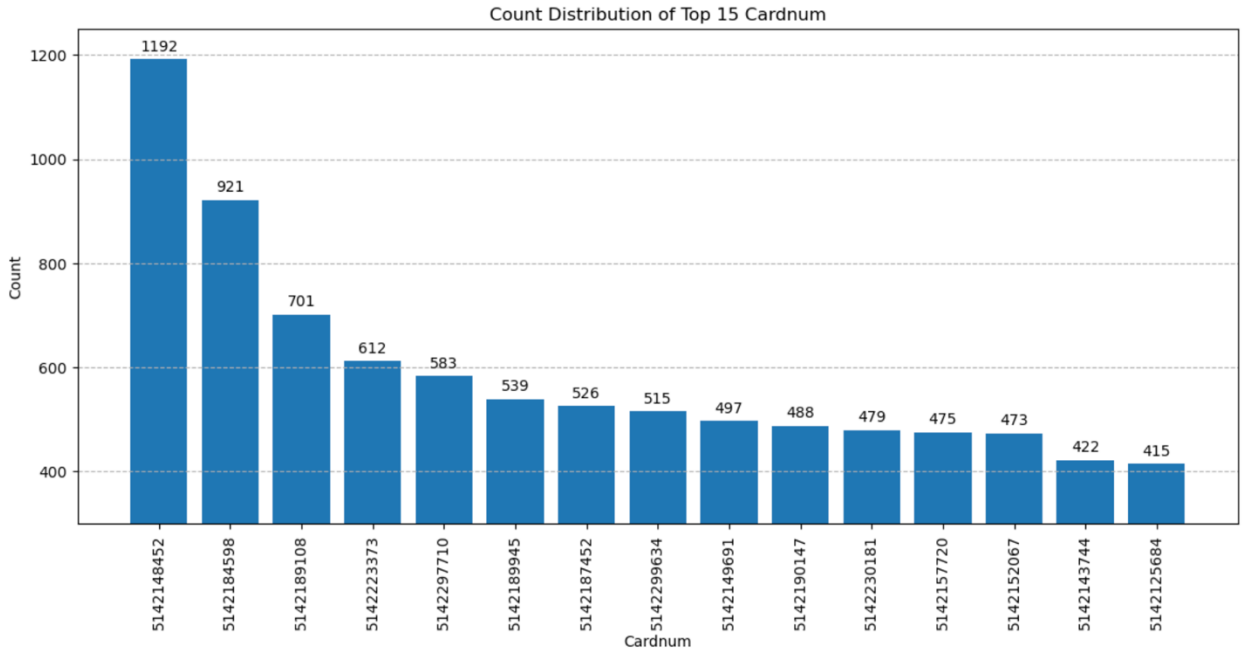






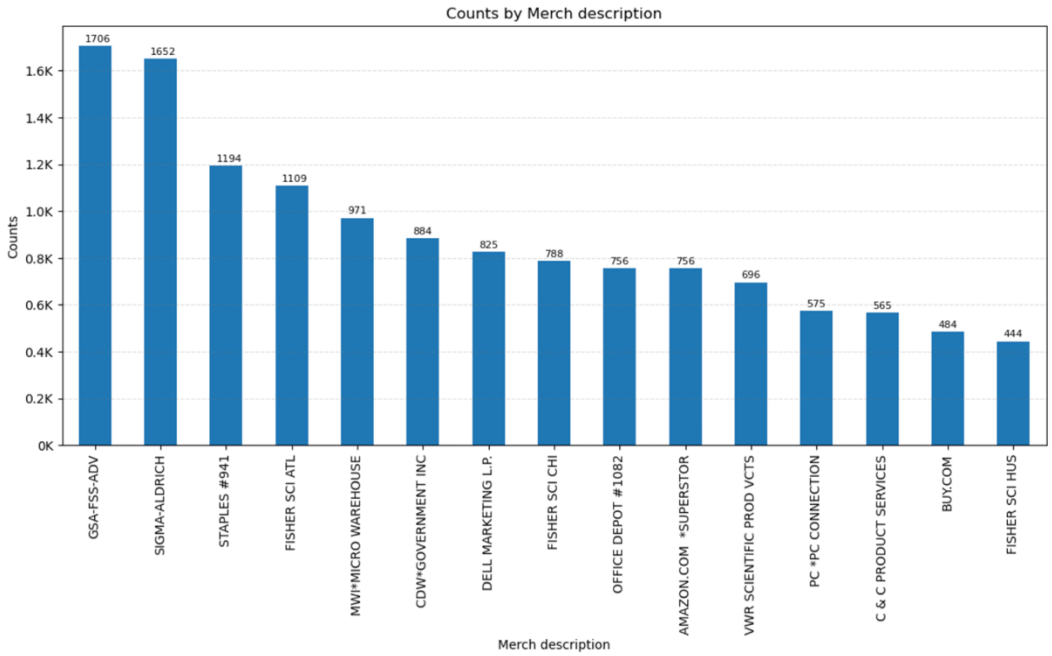
3) Field Name: Cardnum

Description: Each transaction's card number (Cardnum). The distribution displays the top 15 occurrences of card numbers used. The most common card number is 5142148452, showing a total count of 1,192, indicating a high frequency of use which could signify a card favored for regular transactions or by a heavy user.



4) Field Name: Merchnum

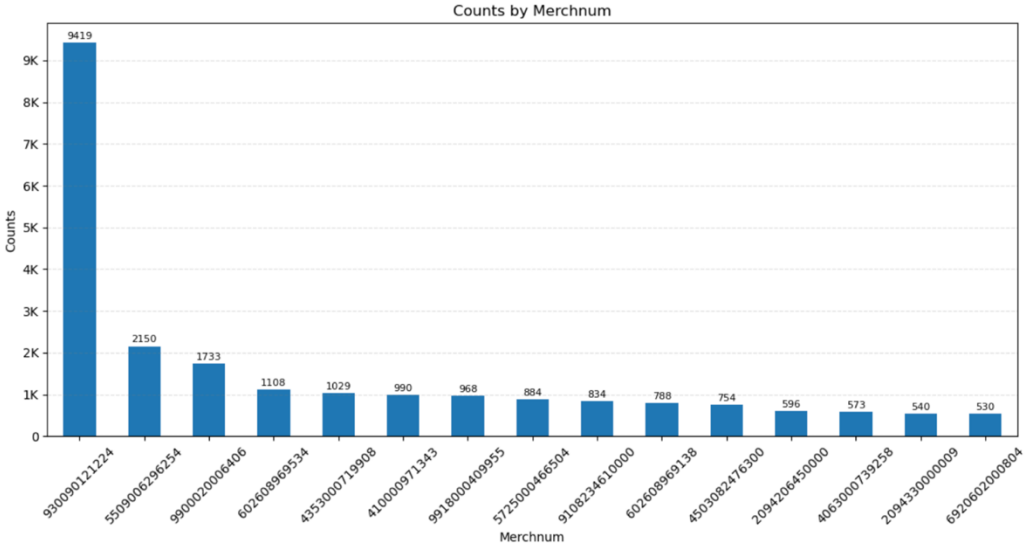
Description: Merchant number (Merchnum). The highest occurring Merchnum stands out markedly at 9,419 counts which is 930090121224, suggesting it might be a focal point for



transactions. Other Merchnum counts span from just over 2,000 down to around 500, highlighting a substantial variation in transaction frequency among merchants.

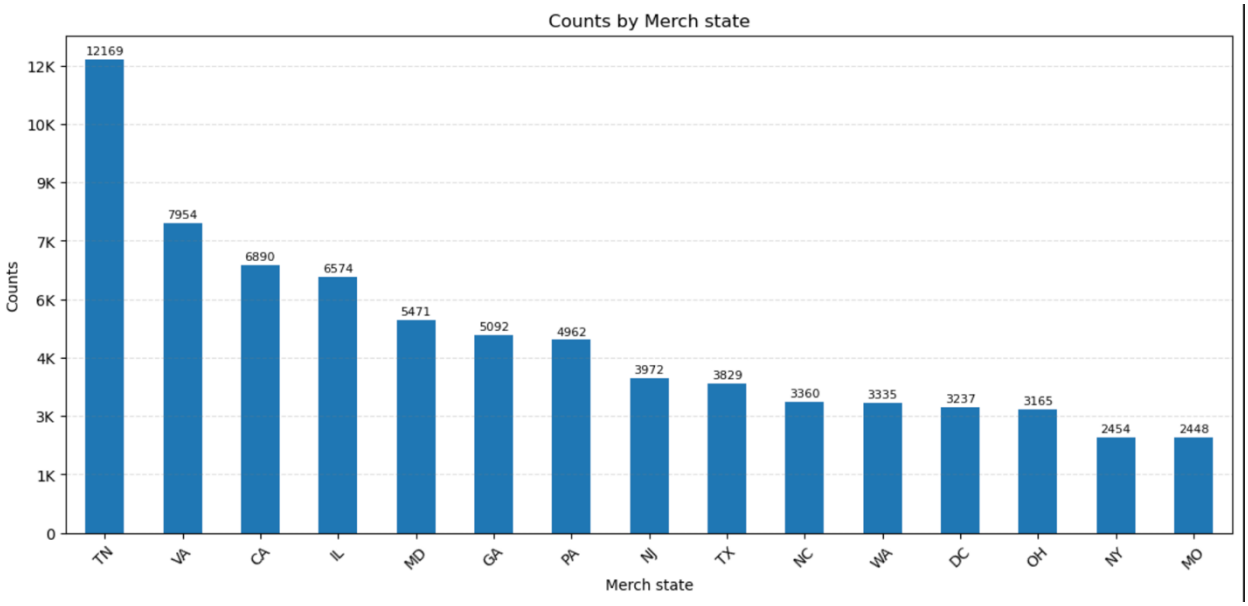
5) Field Name: Merch description

Description: The description of Merchant (Merch description). The bar chart details transaction counts for various merchant descriptions, with 'GSA-FSS ADV' leading at 1,706 transactions. The data reflects a descending order of activity, illustrating a wide range in transaction volumes across merchants.



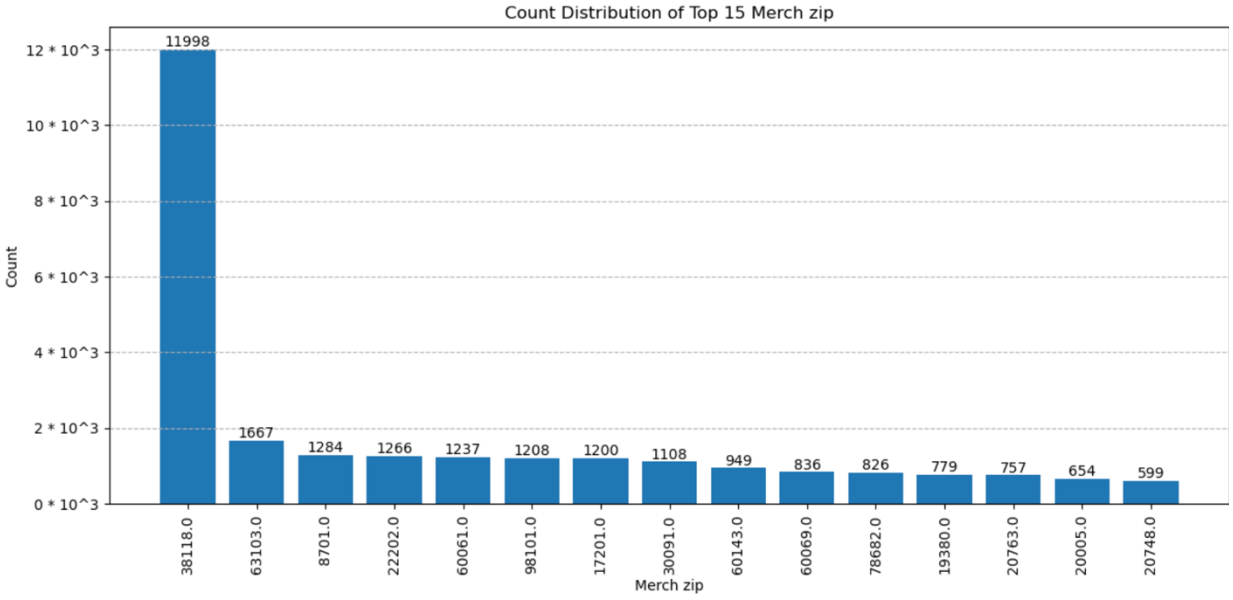
6) **Field Name: Merch State**

Description: Transaction's State (Merch State). The bar chart presents a count of transactions by merchant state, showing Tennessee (TN) with the highest at over 12,000. A clear trend of decreasing transaction counts is visible across the states displayed.



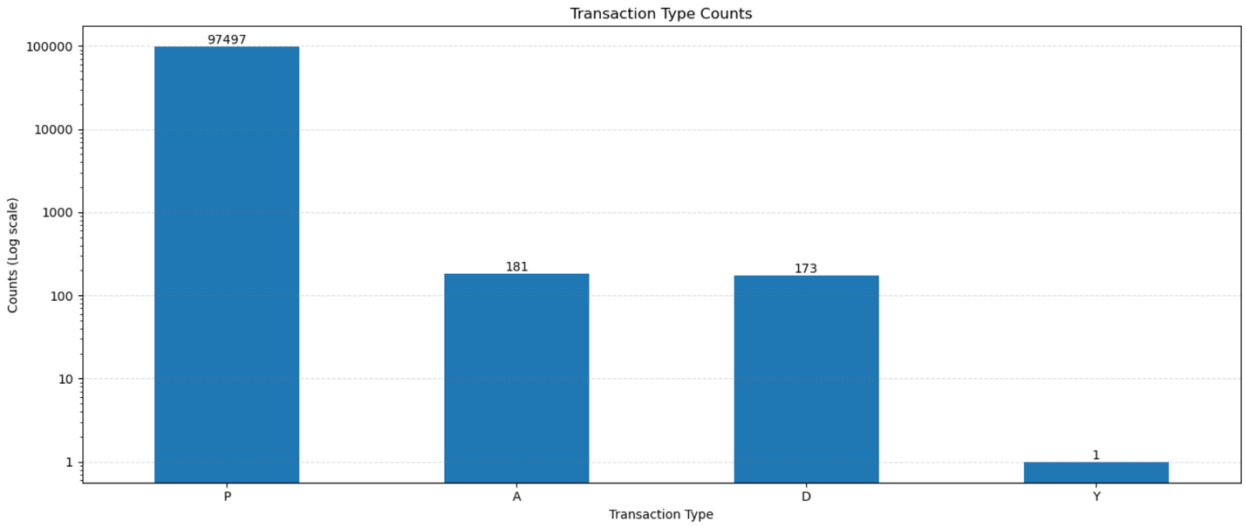
7) **Field Name: Merch zip**

Description: Transaction's zip code. This bar chart outlines the count distribution for the top 15 merchant ZIP codes. The ZIP code 38130 leads significantly, with nearly 12,000 counts, indicating a possible hotspot of commercial activity.



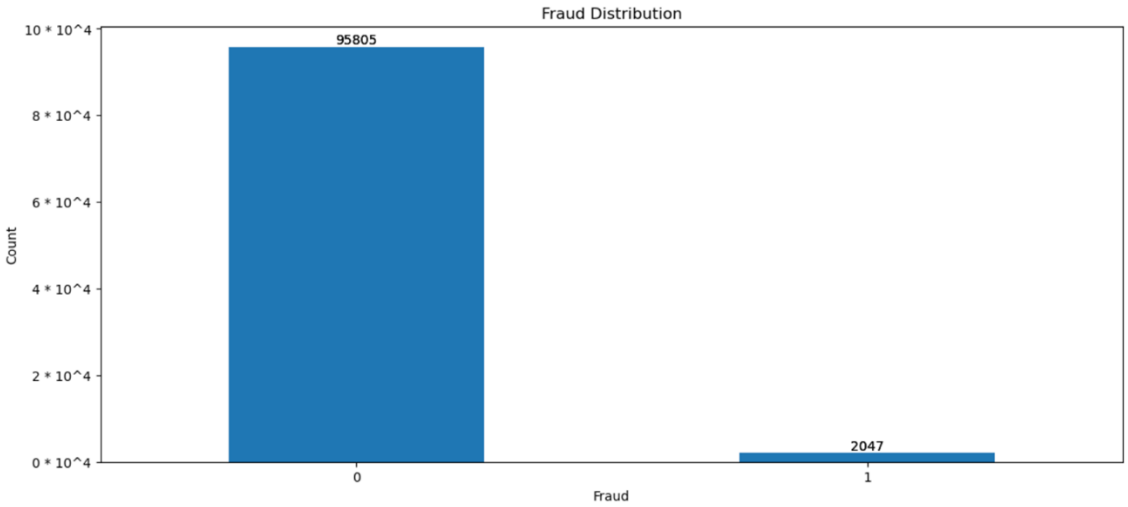
8) **Field Name: Transtype**

Description: Transaction's Type (Transtype). The bar chart illustrates the counts of different transaction types on a logarithmic scale. Type 'P' transactions dominate the chart with 97,497 occurrences, substantially outnumbering the others. Types 'A' and 'D' have a comparable presence, while 'Y' is scarcely represented with a single transaction.



9) **Field Name: Fraud**

Description: The bar chart displays a fraud distribution where non-fraudulent transactions, labeled '0', vastly outnumber the fraudulent ones, labeled '1', with counts of 95,805 and 2,047 respectively. This visual disparity underscores the relative infrequency of fraud in the dataset



10) Field Name: Amount

Description: Each transaction's amount (Amount). This histogram displays the distribution of the "Amount" variable using a logarithmic scale on the y-axis to accommodate the wide range of values. The majority of the data is heavily concentrated at the lower end, near zero, indicating a highly skewed distribution with a sharp peak at the first bin. Additionally, there is a smaller but noticeable increase at the highest end of the range, suggesting the presence of outliers or a long tail to the right of the distribution.