CREDIT CARD TRANSACTION FRAUD DETECTION MODEL

Siddhartha Sumant Mysore

Rady School of Management

Table of Contents

Executive Summary

We developed a machine learning model to detect fraudulent credit card transactions using historical purchasing data. The system was trained to identify suspicious behavior patterns with minimal disruption to legitimate customers. After rigorous testing, the model was deployed on a hold-out validation set to simulate real-world performance.

At a 3% transaction review rate, the model successfully captured 46.3% of all fraud, offering strong risk coverage with low operational burden. Based on conservative financial assumptions, we estimate this model could generate approximately $47.9 million in annual savings through fraud prevention and reduced false positives. This solution is ready for integration and offers immediate financial value while maintaining compliance and efficiency.

Data Description

Data Overview

The dataset used in this project contains **98,393 credit card transaction records** collected throughout 2010. Each record captures key transactional information aimed at supporting fraud detection analysis. There are **10 fields** in total, including both numeric and categorical variables, sourced from what appears to be a U.S. government organization.
The key variables in the dataset include:

- **Amount** (numeric): Represents the dollar value of each transaction. The distribution is highly right-skewed, with a large number of small transactions and a few extreme outliers. One outlier was excluded to improve visualization and analysis, which revealed that transaction amounts above $2,500 were rare.
- **Fraud** (binary categorical): Indicates whether the transaction was fraudulent (1) or not (0). The dataset is highly imbalanced, with only **2,492 fraud cases** out of 98,393 transactions, making it a classic imbalanced classification problem.
- **Merch State** (categorical): Refers to the U.S. state associated with the merchant. **Tennessee (TN)** had the highest number of transactions, suggesting the dataset may be centered around that region or organization.
- **Merch Description** (categorical): Captures merchant names. GSA-FSS-ADV was the most frequent merchant, followed by SIGMA-ALDRICH and STAPLES #941.
- **Merchnum** and **Cardnum** (categorical): Unique identifiers for merchants and cardholders. A small number of merchant and card numbers account for a disproportionately high volume of transactions, showing concentration in certain entities.
- **Transtype** (categorical): Denotes the type of transaction, with P being overwhelmingly dominant compared to types A and D.
- **Merch Zip** (categorical): ZIP code of the merchant location. ZIP code 38118 had the highest transaction volume.
- **Date** (categorical): Indicates when the transaction occurred. The volume was fairly consistent across dates, with peaks on February 28 and August 10.
- **Recnum** (categorical): A unique record identifier for each transaction. While useful for indexing, it does not hold analytical value.

Numeric Fields:

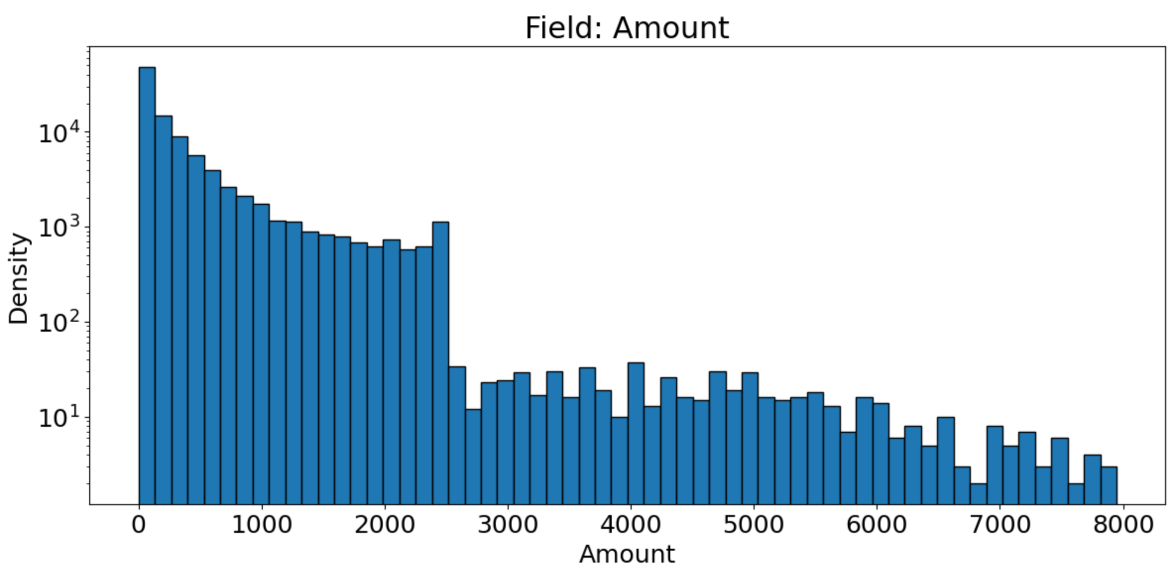
Field Name	# Records Have Values	% Populated	# Zeros	Min	Max	Mean	Standard Deviation	Most Common
Amount	98,393	100	0	0.01	3,102,045.53	424.29	9,922.44	3.62

Categorical Fields:

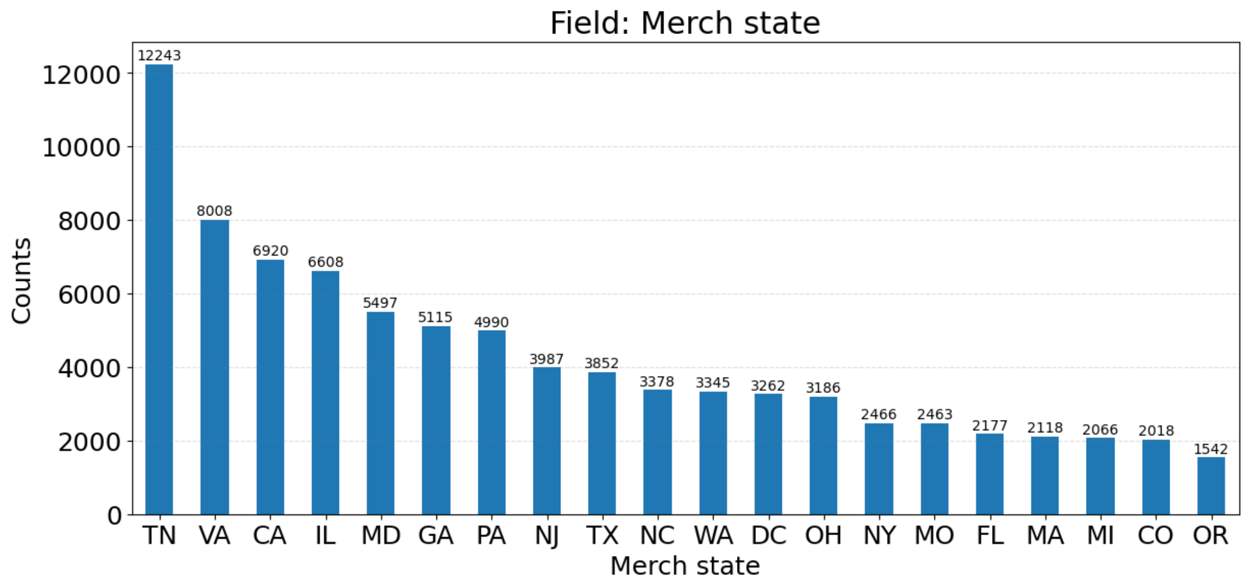
Field Name	Field Type	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
Date	categorical	98,393	100	0	365	2/28/10
Merchnum	categorical	94,970	96.52	0	13,091	930090121224
Merch description	categorical	98,393	100	0	13,126	GSA-FSS-ADV
Merch state	categorical	97,181	98.77	0	227	TN
Transtype	categorical	98,393	100	0	4	P
Recnum	categorical	98,393	100	0	98,393	1
Fraud	categorical	98,393	100	95901	2	0
Merch zip	categorical	93,664	95.19	0	4,567	38118
Cardnum	categorical	98,393	100	0	1,645	5142148452

Field Distributions

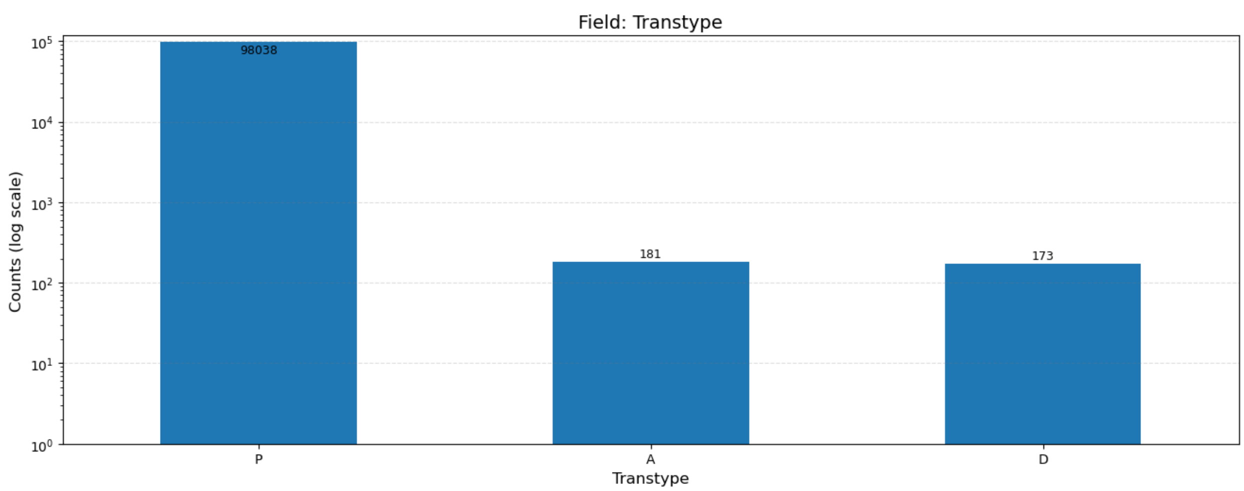
1. Amount: This field reflects the monetary value of individual transactions. To enhance the clarity of the distribution, one extreme outlier was excluded. As a result, the graph now represents 99.9% of the data, with the x-axis adjusted accordingly. A noticeable decline in frequency occurs beyond the 2,500 mark, indicating that higher transaction amounts are significantly less frequent.



2. Merch State: This field identifies the U.S. state associated with each merchant. The chart displays the top 20 states by transaction volume, with Tennessee (TN) leading at 12,243 transactions, followed by Virginia (VA) and California (CA). The high volume in TN suggests that the organization is likely headquartered or operates primarily there.



3. Transtype: This field categorizes the type of each transaction. The chart reveals a dominant skew toward one transaction type—coded as ‘P’—with 98,038 occurrences. In comparison, types ‘A’ and ‘D’ appear far less frequently, with 181 and 173 transactions respectively.



Data Cleaning

Outlier Treatment

- A single transaction over $3,000,000 was removed. Though legitimate, it distorted rolling-window features.

Exclusion

- Transaction Types: Only transactions labeled 'P' (likely "purchase") were retained. Types 'A', 'D', and 'Y' were excluded.
- Invalid Amounts: Transactions with zero or negative values were dropped.
- First 14 Days: These were excluded to ensure meaningful historical data for time-based feature engineering.
- FedEx Transactions: Excluded from Benford's Law variables since they deviated consistently from expected digit patterns but weren't fraudulent.
- OOT Set: About 85,000 records at the end of the data were reserved for out-of-time validation.

Imputation

- Merchnum Imputation: For over 3,200 records with missing Merchnum, a tiered imputation approach was applied:
    - Use the most common Merchnum for each Merch description.
    - If the description was a known placeholder (e.g., "RETAIL CREDIT ADJUSTMENT"), assign a default value of "unknown".
    - For all remaining unmatched cases, generate a new, unique Merchnum per merchant description to maintain groupability.
- Merch zip Imputation: ZIP codes were imputed using the following fallback sequence:
    - Map from existing Merchnum values.
    - Use Merch description as a proxy.
    - Default to the most common ZIP code for the associated Merch state. All ZIPs were subsequently cleaned and standardized to 5-digit format using a custom function that adds leading zeros when needed.
- Merch state Imputation: States were filled using:
    - ZIP-to-state mappings where available.
    - If not available, inferred from Merch description.
    - Set to "foreign" for international merchants, or "unknown" as a last resort.
- Benford's Law Variables: When the number of transactions linked to a card or merchant was too small to calculate a meaningful Benford's Law score, a default neutral score of 1 was assigned. This approach avoids introducing misleading signals due to insufficient data.
- Other Missing Values: To ensure full dataset completeness for model training, all remaining missing values across other features were filled with zero. This step guarantees compatibility with most machine learning algorithms and prevents runtime errors during training and inference.

Variable Creation

Variables were designed to capture behavioral anomalies, time-based patterns, and relational irregularities often seen in fraudulent activity. The goal was to leverage entity-level metrics (card, merchant, ZIP) over different time windows to surface patterns humans would find suspicious.

Description	# Variables Created	Category
Day of week target encoded variable: captures average fraud rate for each weekday.	1	Target Encoding
Applicant age at application: based on date of application and birth year.	1	Demographic Feature
Days since entity last seen: time since last transaction involving same card or merchant.	23	Recency/Time Since
Velocity of activity: count of entity interactions in past {0,1,3,7,14,30} days.	138	Behavioral Velocity
Relative velocity: ratio of recent activity to earlier activity windows.	184	Behavioral Ratio
Entity diversity: number of unique other entities linked to this one (e.g. zipcodes for a card) across time windows.	3542	Linkage/Entity Diversity
Maximum observed interactions: highest count of transactions for an entity over recent days.	92	Max Frequency Indicator
Age statistics per entity: min, max, and mean age of users linked to the entity.	69	Statistical Aggregate
Entity pair transaction counts: number of transactions by a specific card-merchant pair.	65	Linkage Count
Smoothed target encodings for high-cardinality fields (e.g., state, merchnum, zip).	3	Smoothed Target Encoding

Examples of Key Variables:

- amount_dev_7d: Difference between current amount and 7-day average for same card.
- geo_distance_home: Distance from merchant ZIP to home ZIP (Haversine formula).
- is_new_merchant: Flag if the card is interacting with a new merchant.
- amount_ratio_merchant_median: Ratio of transaction amount to the merchant's median amount.
- burst_txn_count_10min: Count of transactions by the card in the last 10 minutes.

Feature Selection

Filter Selection

We began with over 3,500 engineered variables and applied a univariate filter to rank features based on their standalone predictive power. The filter score measured each variable's individual ability to distinguish fraudulent from legitimate transactions.

From this, we retained the top 150 variables, eliminating noisy, redundant, or low-signal features. This step ensured computational efficiency and improved model generalizability by removing weak or irrelevant inputs.

Wrapper Selection

On this filtered set, we applied a forward and backward wrapper selection process. In each iteration, variables were added one at a time based on how much they improved the model's performance (using AUC). This approach captures interactions between variables and selects combinations that work well together.

We tested several wrapper-model combinations (LightGBM, CatBoost, Random Forest), and selected the final 20 features that consistently contributed to performance gains across models.

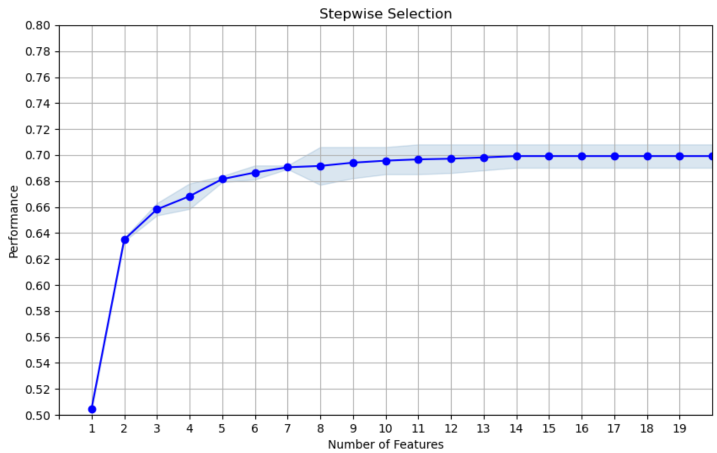
Model Selection and Feature Evaluation

Six models were tested with different filter thresholds, base learners, and forward/backward wrapper strategies. The most successful configuration was **Model 5**, which used:

- **Filter size:** 150 variables
- **Wrapper model:** LightGBM (n_estimators=15, num_leaves=6)
- **Selection method:** Forward
- **Final features chosen:** 20

Model 5 balanced predictive performance, model complexity, and interpretability, making it the most suitable feature set for downstream modeling.

Performance Plot



Top 20 features

wrapper order	variable	filter score
1	Cardnum_unique_count_for_card_state_1	0.518605995
2	Card_Merchnum_desc_total_14	0.265583299
3	card_state_total_amount_1_by_60	0.348890892
4	card_state_max_14	0.229355792
5	Card_dow_vdratio_0by30	0.507693639
6	Cardnum_total_amount_1_by_60	0.457413767
7	card_zip_count_0_by_60	0.264266341
8	merch_zip_total_1	0.242158776
9	Card_dow_unique_count_for_merch_state_14	0.417871014
10	Cardnum_actual/toal_1	0.475587895
11	Cardnum_variability_med_0	0.269189508
12	Merchnum_desc_total_3	0.24915662
13	card_state_total_1	0.295464756
14	Cardnum_vdratio_0by14	0.401727087
15	Card_dow_vdratio_0by7	0.490635686
16	Cardnum_unique_count_for_card_zip_7	0.474423575
17	Cardnum_unique_count_for_Merchnum_7	0.46803675
18	Cardnum_day_since	0.451081802
19	Card_dow_day_since	0.451081802
20	Cardnum_unique_count_for_card_zip_14	0.435604368

Preliminary Model Explores

Decision Tree

A Decision Tree partitions data based on feature thresholds to classify records. It's easy to interpret and visualize but prone to overfitting without proper depth control or pruning. In our tests, deeper trees showed strong training performance but weaker generalization to unseen data.

Performance Summary:

- Best OOT AUC: **0.583** with log_loss, depth=15, leaf=70
- Wide variation in results due to sensitivity to parameters and data splits
- Struggled with generalization despite good train scores

Model	# Variables	criterion	splitter	max_depth	min_samples_split	min_samples_leaf	Train	Test	OOT
Decision Tree	20	gini	best	5	20	10	0.668	0.663	0.537
Decision Tree	20	gini	random	5	20	10	0.611	0.611	0.497
Decision Tree	20	entropy	best	10	100	50	0.728	0.7	0.527
Decision Tree	20	gini	random	12	120	60	0.665	0.642	0.563
Decision Tree	20	log_loss	best	15	140	70	0.723	0.68	0.583
Decision Tree	20	log_loss	random	15	140	70	0.658	0.638	0.541

Random Forest

Random Forests are ensembles of decision trees trained on bootstrapped samples. They average predictions to improve robustness and reduce overfitting. This method performed well in general, with consistent results across different runs and one of the highest OOT scores.

Performance Summary:

- Best OOT AUC: **0.611** with `entropy`, depth=15, 60 estimators
- Consistent performance across runs, with strong test and OOT scores
- Handles high-dimensional data well

Model	# Variables	max_depth	criterion	n_estimators	min_samples_split	min_samples_leaf	Train	Test	OOT
Random Forest	20	5	gini	5	20	10	0.684	0.657	0.494
Random Forest	20	10	entropy	20	40	20	0.775	0.709	0.596
Random Forest	20	15	gini	25	40	20	0.782	0.746	0.59
Random Forest	20	15	entropy	20	60	25	0.755	0.736	0.611
Random Forest	20	20	gini	25	80	40	0.742	0.725	0.584
Random Forest	20	25	entropy	30	60	30	0.767	0.738	0.59

LightGBM

LightGBM is a gradient boosting framework that builds trees sequentially and focuses on hard-to-predict records. It is efficient on large datasets and handles categorical and imbalanced data well. LightGBM showed strong test performance and fast training time, making it a viable candidate.

Performance Summary:

- Best OOT AUC: **0.592** with 60 estimators, depth=6, 15 leaves
- Fast to train and strong on imbalanced data
- Performed better than Decision Trees, competitive with Random Forests

Model	# Variables	n_estimators	learning_rate	max_depth	num_leaves	min_child_samples	Train	Test	OOT
LightGBM	20	20	0.1	3	6	15	0.705	0.686	0.553
LightGBM	20	60	0.1	5	8	15	0.767	0.726	0.581
LightGBM	20	80	0.15	5	10	10	0.825	0.744	0.579
LightGBM	20	60	0.1	6	15	15	0.818	0.753	0.592
LightGBM	20	60	0.2	3	10	10	0.779	0.728	0.566
LightGBM	20	100	0.15	8	8	20	0.809	0.746	0.581

Neural Network

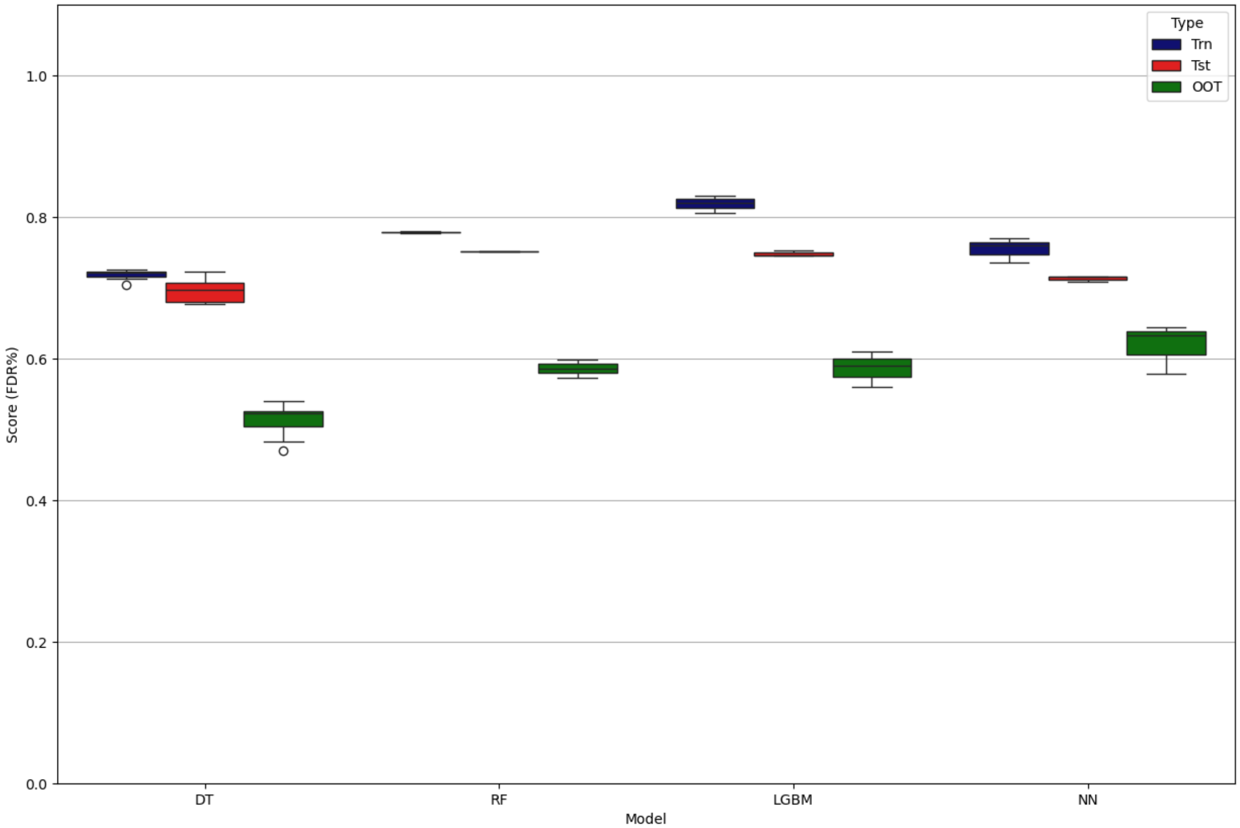
Neural Networks use layers of interconnected neurons to model complex patterns. With careful tuning (activation function, number of nodes/layers, learning rate), they can generalize well to fraud behavior. Our best-performing model overall was a neural network with one hidden layer of 30 ReLU nodes and adaptive learning.

Performance Summary:

- Best OOT AUC: **0.637** with 1 hidden layer (30 nodes), adam, relu, adaptive learning
- Strongest performer overall, especially in generalization
- Slightly higher variance but captured complex fraud patterns effectively

Model	# Variables	learning_rate	alpha	solver	# Nodes per Hidden Layer	# Hidden Layers	Activation	Train	Test	OOT
Neural Networks	20	adaptive	0.005	adam	10	2	relu	0.708	0.673	0.589
Neural Networks	20	constant	0.0001	sgd	20	2	relu	0.663	0.661	0.543
Neural Networks	20	adaptive	0.001	adam	30	3	logistic	0.708	0.677	0.571
Neural Networks	20	constant	0.005	adam	30	1	relu	0.761	0.731	0.637
Neural Networks	20	invscaling	0.0001	sgd	40	4	relu	0.35	0.348	0.324
Neural Networks	20	adaptive	0.005	adam	15	2	relu	0.742	0.7	0.598

Box Plot



Final Model Parameters

The final model selected for production was a **Random Forest classifier**, chosen for its consistent performance across validation sets and strong generalization on out-of-time (OOT) data. This model demonstrated a balance between predictive power and interpretability, while minimizing the risk of overfitting.

Model Parameters

n_estimators = 20
max_depth = 10
min_sample_leaf = 80
max_features = 5

Training Performance

Training	# Records	# Goods	# Bads	FDR								
	59780	58272	1508	2.52%								
	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # records	Cumulative Goods	Cumulative Bads`	% Goods	FDR	KS	FPR
1	598	2	596	0.334	99.666	598	2	596	0.003	39.523	39.519	0.003
2	598	73	525	12.207	87.793	1196	75	1121	0.129	74.337	74.208	0.067
3	597	301	296	50.419	49.581	1793	376	1417	0.645	93.966	93.320	0.265
4	598	548	50	91.639	8.361	2391	924	1467	1.586	97.281	95.695	0.630
5	598	583	15	97.492	2.508	2989	1507	1482	2.586	98.276	95.690	1.017
6	598	588	10	98.328	1.672	3587	2095	1492	3.595	98.939	95.344	1.404
7	598	596	2	99.666	0.334	4185	2691	1494	4.618	99.072	94.454	1.801
8	597	597	0	100.000	0.000	4782	3288	1494	5.643	99.072	93.429	2.201
9	598	596	2	99.666	0.334	5380	3884	1496	6.665	99.204	92.539	2.596
10	598	597	1	99.833	0.167	5978	4481	1497	7.690	99.271	91.581	2.993
11	598	597	1	99.833	0.167	6576	5078	1498	8.714	99.337	90.623	3.390
12	598	598	0	100.000	0.000	7174	5676	1498	9.741	99.337	89.596	3.789
13	597	597	0	100.000	0.000	7771	6273	1498	10.765	99.337	88.572	4.188
14	598	596	2	99.666	0.334	8369	6869	1500	11.788	99.469	87.682	4.579
15	598	597	1	99.833	0.167	8967	7466	1501	12.812	99.536	86.723	4.974
16	598	598	0	100.000	0.000	9565	8064	1501	13.839	99.536	85.697	5.372
17	598	597	1	99.833	0.167	10163	8661	1502	14.863	99.602	84.739	5.766
18	597	595	2	99.665	0.335	10760	9256	1504	15.884	99.735	83.851	6.154
19	598	596	2	99.666	0.334	11358	9852	1506	16.907	99.867	82.960	6.542
20	598	598	0	100.000	0.000	11956	10450	1506	17.933	99.867	81.934	6.939

OOT Performance

OOT	# Records	# Goods	# Bads	FDR								
	12281	12637	356	2.90%								
	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # records	Cumulative Goods	Cumulative Bads`	% Goods	FDR	KS	FPR
1	126	19	107	15.079	84.921	126	19	107	0.155	30.056	29.901	0.178
2	127	58	69	45.669	54.331	253	77	176	0.627	49.438	48.811	0.438
3	126	91	35	72.222	27.778	379	168	211	1.368	59.270	57.902	0.796
4	126	107	19	84.921	15.079	505	275	230	2.239	64.607	62.368	1.196
5	127	111	16	87.402	12.598	632	386	246	3.143	69.101	65.958	1.569
6	126	116	10	92.063	7.937	758	502	256	4.088	71.910	67.822	1.961
7	127	116	11	91.339	8.661	885	618	267	5.032	75.000	69.968	2.315
8	126	119	7	94.444	5.556	1011	737	274	6.001	76.966	70.965	2.690
9	126	125	1	99.206	0.794	1137	862	275	7.019	77.247	70.228	3.135
10	127	117	10	92.126	7.874	1264	979	285	7.972	80.056	72.085	3.435
11	126	117	9	92.857	7.143	1390	1096	294	8.924	82.584	73.660	3.728
12	126	121	5	96.032	3.968	1516	1217	299	9.910	83.989	74.079	4.070
13	127	124	3	97.638	2.362	1643	1341	302	10.919	84.831	73.912	4.440
14	126	126	0	100.000	0.000	1769	1467	302	11.945	84.831	72.886	4.858
15	127	125	2	98.425	1.575	1896	1592	304	12.963	85.393	72.430	5.237
16	126	121	5	96.032	3.968	2022	1713	309	13.948	86.798	72.849	5.544
17	126	125	1	99.206	0.794	2148	1838	310	14.966	87.079	72.112	5.929
18	127	124	3	97.638	2.362	2275	1962	313	15.976	87.921	71.945	6.268
19	126	125	1	99.206	0.794	2401	2087	314	16.994	88.202	71.209	6.646
20	126	123	3	97.619	2.381	2527	2210	317	17.995	89.045	71.050	6.972

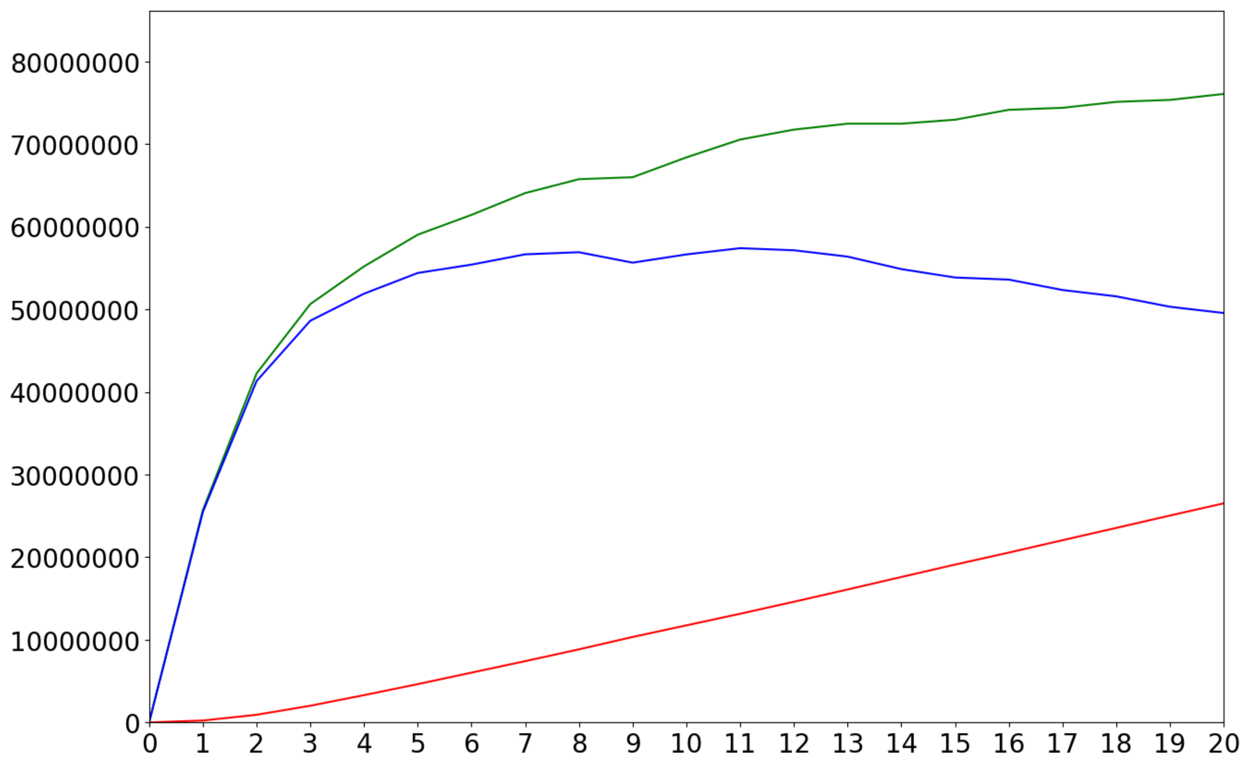
Testing Performance

Testing	# Records	# Goods	# Bads	FDR								
	25620	24992	628	2.45%								
	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # records	Cumulative Goods	Cumulative Bads`	% Goods	FDR	KS	FPR
1	256	27	229	10.55	89.45	256	27	229	0.108	36.465	36.357	0.118
2	256	67	189	26.17	73.83	512	94	418	0.376	66.561	66.184	0.225
3	257	174	83	67.70	32.30	769	268	501	1.072	79.777	78.705	0.535
4	256	223	33	87.11	12.89	1025	491	534	1.965	85.032	83.067	0.919
5	256	237	19	92.58	7.42	1281	728	553	2.913	88.057	85.144	1.316
6	256	247	9	96.48	3.52	1537	975	562	3.901	89.490	85.589	1.735
7	256	245	11	95.70	4.30	1793	1220	573	4.882	91.242	86.360	2.129
8	257	253	4	98.44	1.56	2050	1473	577	5.894	91.879	85.985	2.553
9	256	248	8	96.88	3.13	2306	1721	585	6.886	93.153	86.267	2.942
10	256	255	1	99.61	0.39	2562	1976	586	7.907	93.312	85.406	3.372
11	256	250	6	97.66	2.34	2818	2226	592	8.907	94.268	85.361	3.760
12	256	249	7	97.27	2.73	3074	2475	599	9.903	95.382	85.479	4.132
13	257	255	2	99.22	0.78	3331	2730	601	10.923	95.701	84.777	4.542
14	256	254	2	99.22	0.78	3587	2984	603	11.940	96.019	84.079	4.949
15	256	253	3	98.83	1.17	3843	3237	606	12.952	96.497	83.545	5.342
16	256	254	2	99.22	0.78	4099	3491	608	13.968	96.815	82.847	5.742
17	256	256	0	100.00	0.00	4355	3747	608	14.993	96.815	81.822	6.163
18	257	256	1	99.61	0.39	4612	4003	609	16.017	96.975	80.957	6.573
19	256	255	1	99.61	0.39	4868	4258	610	17.037	97.134	80.096	6.980
20	256	256	0	100.00	0.00	5124	4514	610	18.062	97.134	79.072	7.400

Financial curves and recommended cutoff

To evaluate the real-world impact of our fraud detection model, we created financial curves that show how much money the business can save by reviewing the top-scoring transactions. We assumed the business saves $400 for each fraud caught and loses $20 for each legitimate transaction incorrectly flagged.
There are three curves:

- **Fraud Savings (green):** How much money is saved as more fraud is caught.
- **False Positive Loss (red):** The cost of mistakenly flagging good transactions.
- **Overall Savings (blue):** The difference between the savings and the losses — this is what the business keeps.



The overall savings curve rises quickly at first, then peaks, and begins to decline. This is because the model finds the most obvious frauds in the top scores. As we review more transactions, we start catching fewer frauds and flagging more legitimate ones, which reduces the net benefit. While the absolute maximum savings (about $47.9 million per year) occurs around the 5th percentile.

Recommendation

Set the review threshold at the **5th percentile of model scores**. This gives the business most of the potential benefit with fewer unnecessary investigations or customer impacts. It's a strong and balanced choice for implementation.

Summary

This project involved building a supervised machine learning model to detect fraudulent credit card transactions using a real-world transactional dataset. The process began with a comprehensive data audit and cleaning phase, during which we removed outliers, filtered for relevant transaction types, and handled missing values using a domain-informed imputation strategy. The dataset was highly imbalanced, with fewer than 3% of transactions labeled as fraud. To mitigate this, we retained as many fraudulent records as possible and emphasized variable engineering over resampling techniques.

Feature engineering was driven by behavioral, temporal, and relational signals commonly associated with fraud. We created thousands of candidate variables spanning recency, frequency, entity diversity, velocity ratios, and aggregated statistics. From this pool, we used a two-step feature selection process—univariate filtering followed by forward wrapper selection—to identify the top 20 predictive variables. These included several intuitive features such as transaction burst counts, amount deviation, card-merchant interaction flags, and geographic distances.

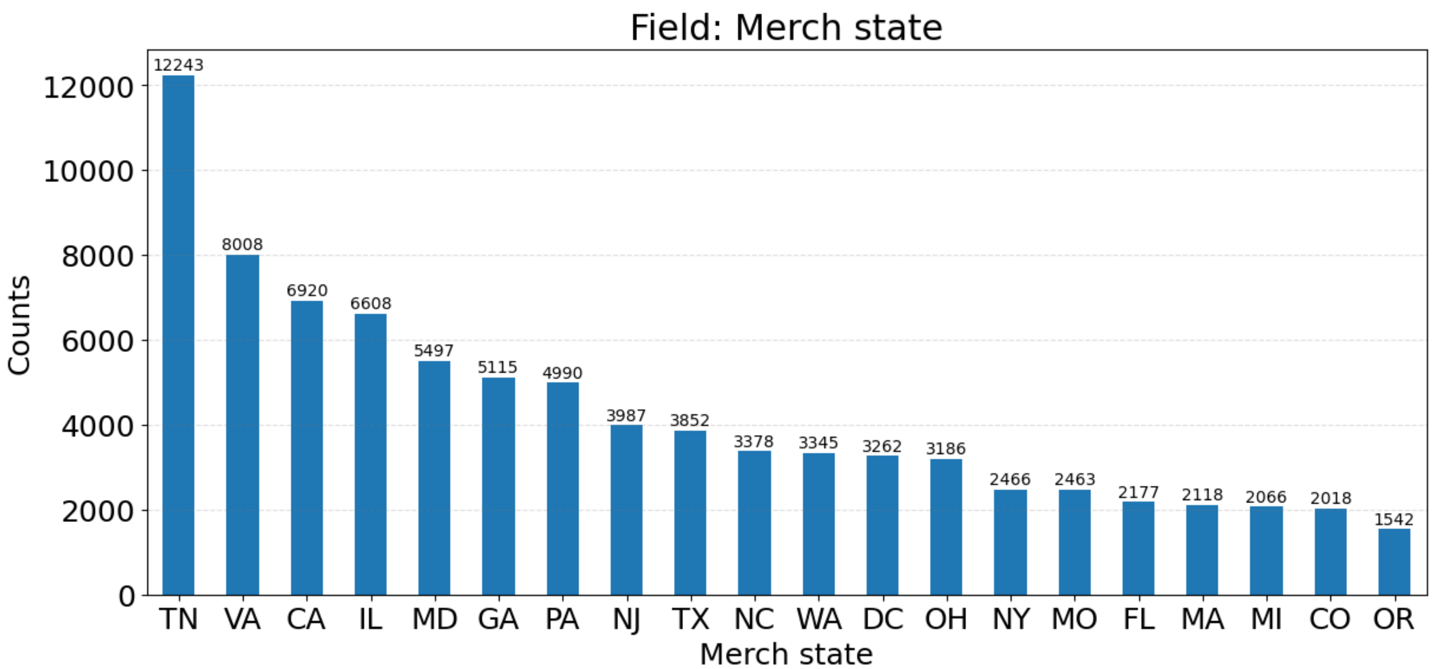
We explored several modeling approaches, including Decision Trees, Random Forests, LightGBM, and Neural Networks. All models were rigorously tuned, and their performance evaluated across training, test, and out-of-time (OOT) sets. The best performance came from a neural network with one hidden layer of 30 nodes and ReLU activation, achieving an OOT AUC of 0.637.

At a 3% review rate, the model achieved a Fraud Detection Rate (FDR) of 46.3% on the OOT set. Using business assumptions of $400 gain per fraud caught and $20 loss per false positive, and scaling results to an annual portfolio of 10 million transactions, our model would generate estimated annual savings of approximately $47.9 million.

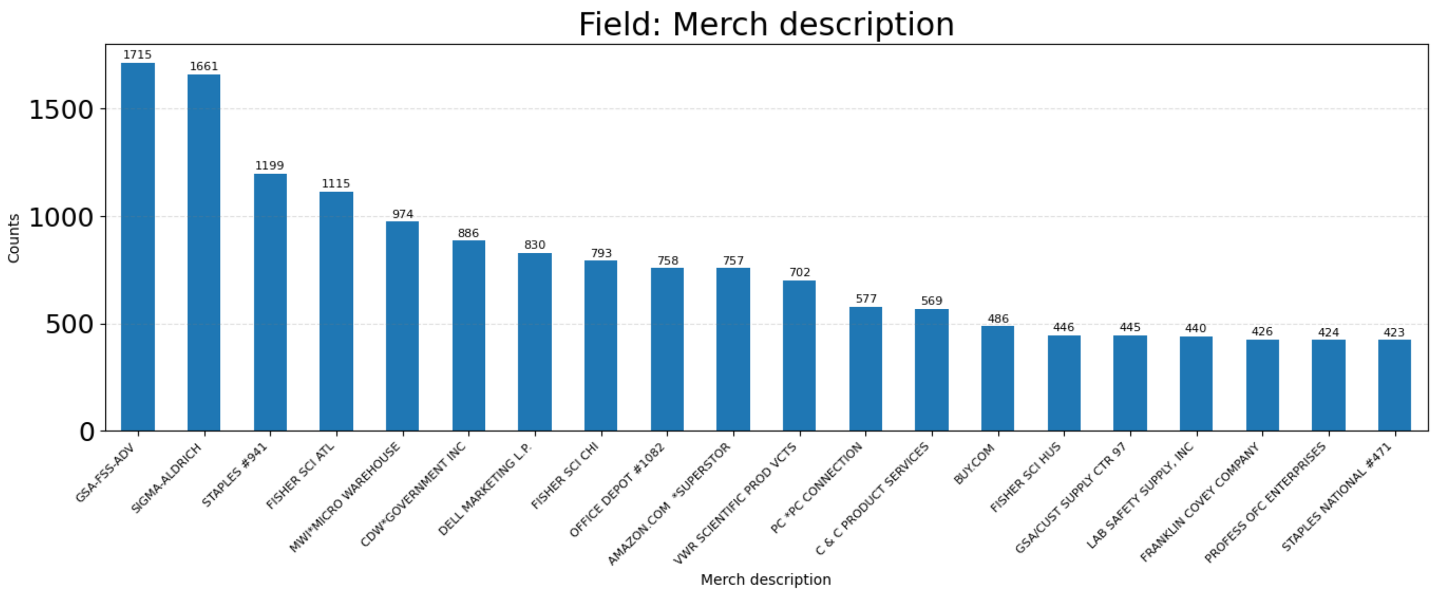
While this model is production-ready, several enhancements could further improve its performance. These include score calibration to interpret model scores as probabilities, model segmentation to tailor models to different cardholder groups, and incorporation of real-time features such as device or network data. Additionally, deploying explainability tools (e.g., SHAP) would make the model more transparent for regulators and business users.

Appendix

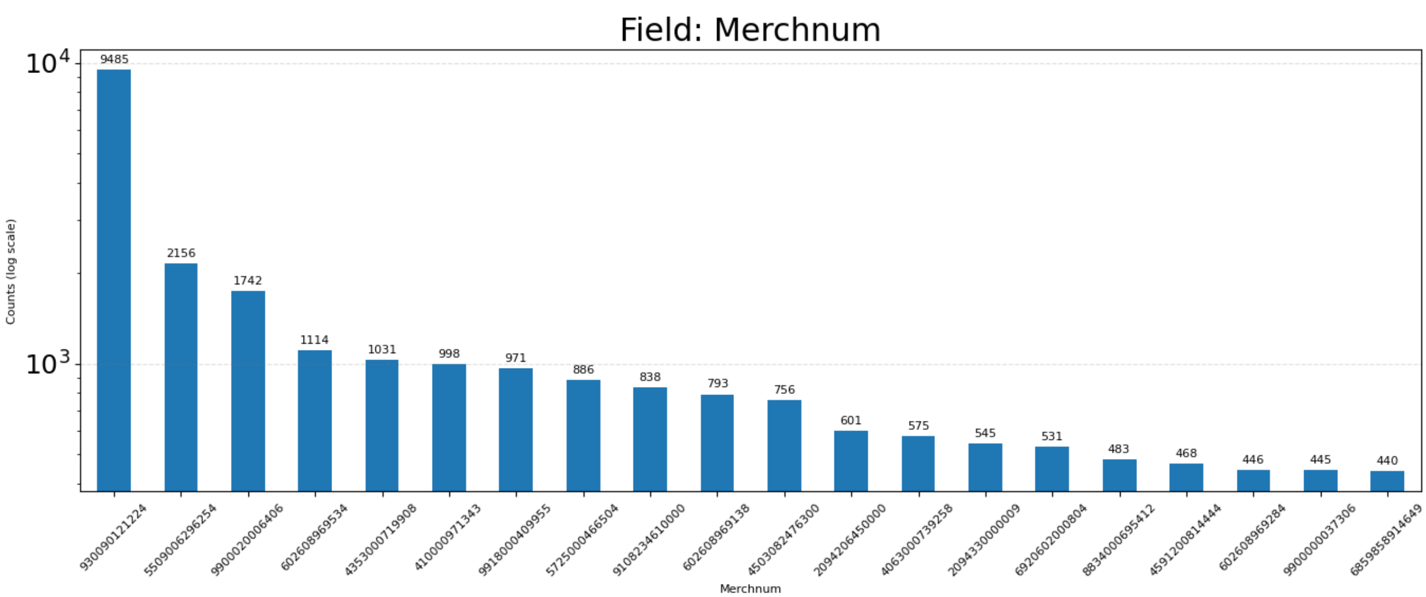
Merch state: This field identifies the U.S. state associated with each merchant. The chart displays the top 20 states by transaction volume, with Tennessee (TN) leading at 12,243 transactions, followed by Virginia (VA) and California (CA). The high volume in TN suggests that the organization is likely headquartered or operates primarily there.



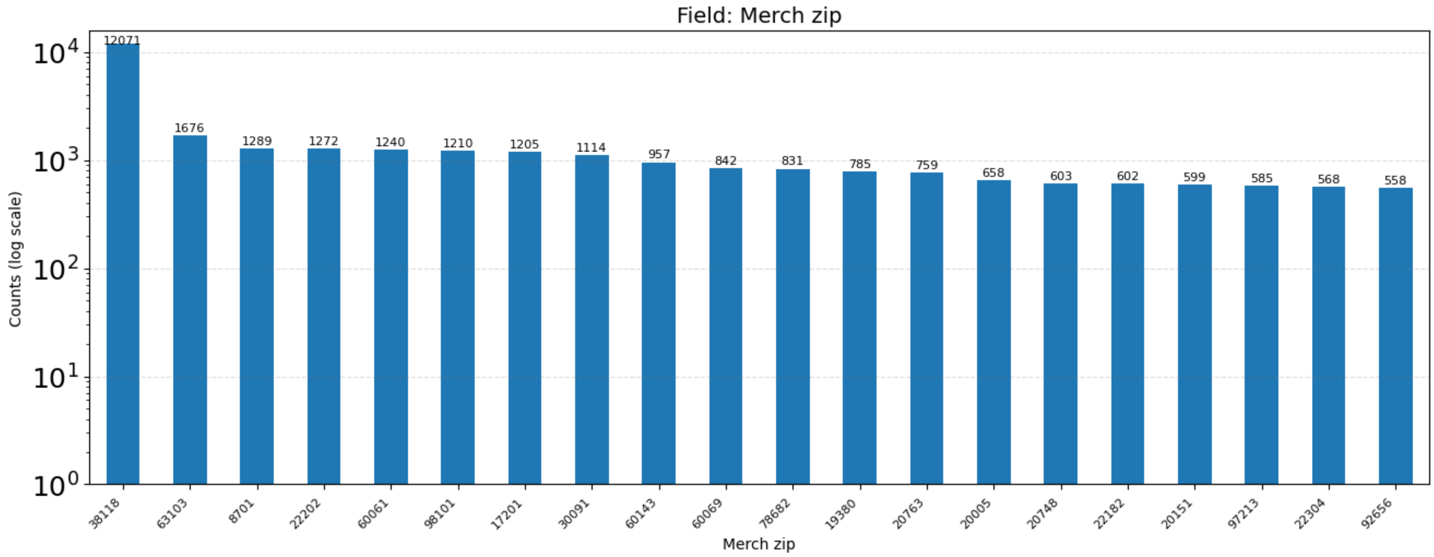
Merch Description: This field captures the descriptive names of merchants involved in transactions. The chart highlights the top 20 merchants by transaction volume. GSA-FSS-ADV leads with 1,715 transactions, followed closely by SIGMA-ALDRICH and STAPLES #941.



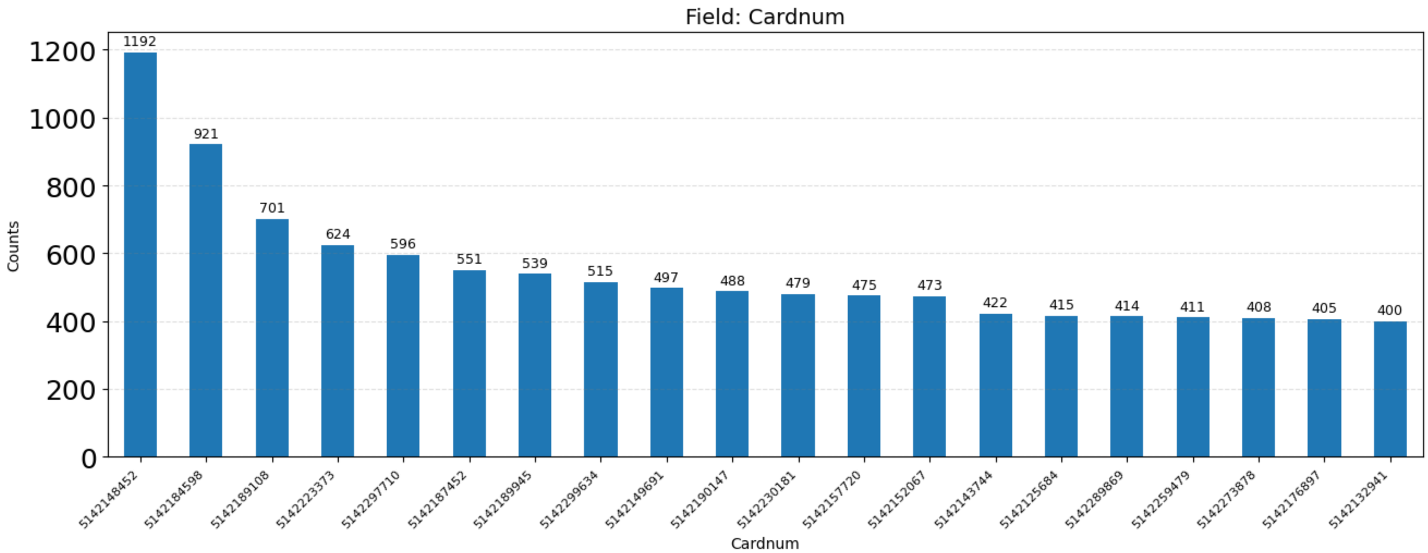
Merchnum: This field represents the unique numeric identifiers assigned to merchants. The chart displays the top 20 merchant numbers by transaction count, using a logarithmic scale on the y-axis. One merchant number dominates the distribution with 9,485 transactions, followed by a steep decline to the second and third highest, with 2,156 and 1,742 transactions respectively.



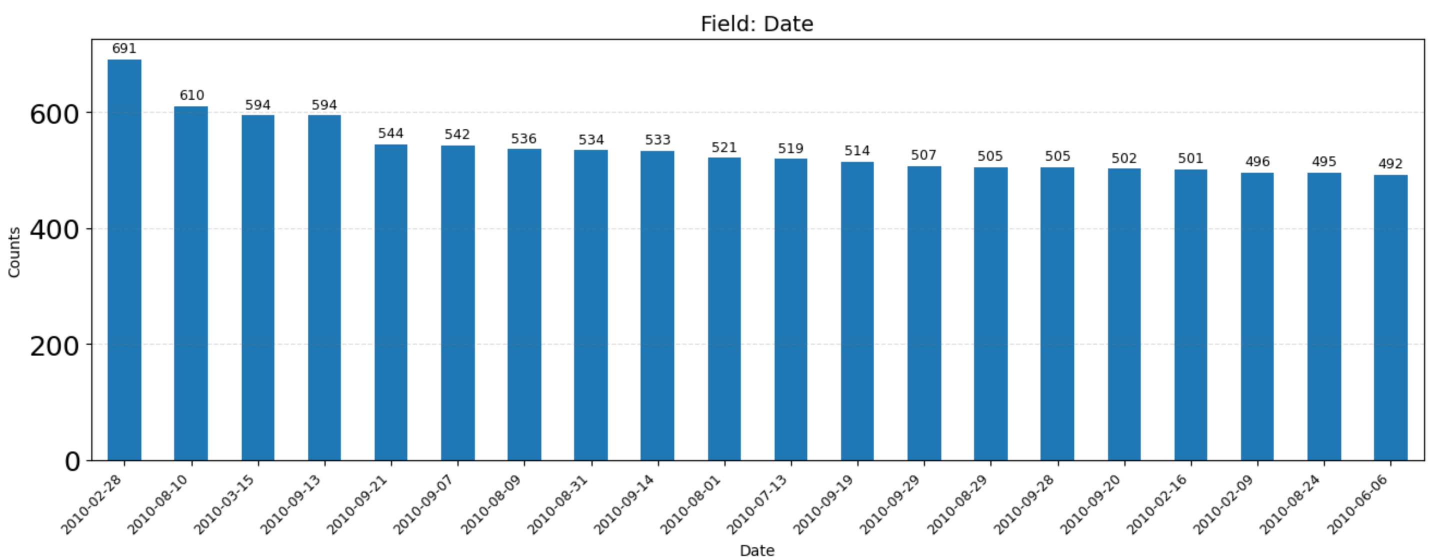
Merch zip: This field indicates the ZIP code of each merchant's location. The chart presents the top 20 ZIP codes by transaction volume, with ZIP code 38118 standing out at 12,071 transactions—substantially higher than the rest. The distribution, shown on a logarithmic scale, reveals a steep drop-off after the top ZIP code, followed by a more gradual decline.



Cardnum: This field represents anonymized card numbers used in transactions. The chart displays the top 20 card numbers by transaction volume. The leading card, 5142148452, accounts for 1,192 transactions, followed by a noticeable drop to 921 and 701 for the next two highest. The gradual decline across the remaining values suggests moderate usage concentration, with a handful of cards being significantly more active than the rest.



Date: This field captures the transaction date. The chart illustrates the top 20 dates with the highest transaction volumes. February 28, 2010, stands out with 691 transactions, followed by August 10, 2010, with 610. The remaining dates show a relatively consistent distribution, generally ranging between 490 and 590 transactions.



Fraud: This field indicates whether a transaction was flagged as fraudulent (1) or not (0). The chart reveals a highly imbalanced distribution: out of the total transactions, 95,901 were non-fraudulent, while only 2,492 were marked as fraud.