Identity Fraud Project Report

Prepared By
NUJOUM UNUS

Table of Contents

Executive Summary 4
Business Problem 5
Business Impact 5
Description of the Data 5
Numeric Fields Summary 5
Categorical Fields Summary 5
Key Distributions 6
SSN 6
Address 7
Zip5 7
Dob 8
Homephone 8
Key Insights from Field Distributions 9
Data Cleaning 10
General Overview 10
Handling Frivolous Identity Elements 10
Remediation Approach 10
Variable Creation 11

Summary of Variables Created 12
Feature Selection Process 13
Final Variables with Filter Score 14
# of Variables vs Performance Plot 15
Model Exploration 15
Tests Performed 17
Model Comparison Plot 22
Final Model Performance 24
Performance on train data 24
Performance on test data 25
Performance on OOT data 26
Financial curves 28
Cutoff vs Savings 28
Summary/conclusions 30
Appendix I – Data Quality Report 31
Description of the Data 31
Record 32
Date 32
SSN 33
Firstname 34
Lastname 35
Address 35
Zip5 36
Dob 36
Homephone 37
Fraud_label 37

Executive Summary

Business Problem

Application fraud is largely driven by synthetic identities. It costs the business an average of $4,000 per incident. With roughly one million applications processed annually and a fraud rate of 1.43%, that translates to 14,300 fraudulent cases and approximately **$57.2 million in annual losses**. We conducted a comprehensive analysis of historical application data, engineered targeted fraud indicators, and developed a predictive model to identify high-risk applications early and reduce exposure to fraud.

Business Impact

Our fraud detection model achieved a **58.98% detection rate** when flagging the **top 2% most suspicious applications**. In practical terms, this means we can intercept over half of the fraudulent cases by rejecting just 2% of incoming applications, resulting in **estimated annual savings of $32.05 million**.

Description of the Data

This dataset has been synthetically generated to support the development of identity fraud detection algorithms, while ensuring compliance with legal and privacy regulations. Although it does not contain real Personally Identifiable Information (PII), it has been carefully designed to reflect realistic patterns observed in actual data, including PII frequency distributions and linkage characteristics.

The dataset contains **1,000,000 records** across **10 fields**, representing application data spanning the period from **January 1, 2017 to December 31, 2017**. The synthetic nature of the dataset ensures no real individuals are represented, while maintaining structural and statistical fidelity necessary for reliable model development and validation.

Numeric Fields Summary

There are no numeric fields except perhaps date and dob, but I have included them in the categorical fields summary table.

Categorical Fields Summary

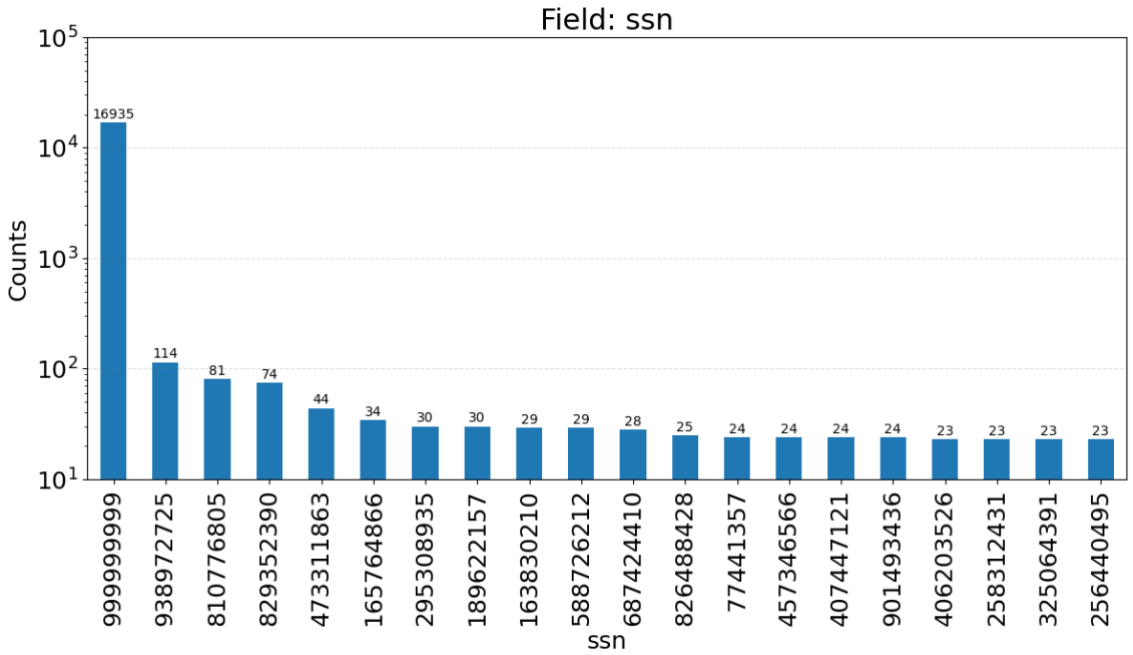
Field Name	Field Type	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
homephone	categorical	1,000,000	100.00%	0	28,244	9999999999

fraud_label	categorical	1,000,000	100.00%	985,607	2	0
lastname	categorical	1,000,000	100.00%	0	177,001	ERJSAXA
zip5	categorical	1,000,000	100.00%	0	26,370	68138
firstname	categorical	1,000,000	100.00%	0	78,136	EAMSTRMT
address	categorical	1,000,000	100.00%	0	828,774	123 MAIN ST
record	categorical	1,000,000	100.00%	0	1,000,000	1
ssn	categorical	1,000,000	100.00%	0	835,819	999999999
date	categorical	1,000,000	100.00%	0	365	20170816
dob	categorical	1,000,000	100.00%	0	42,673	19070626

Key Distributions

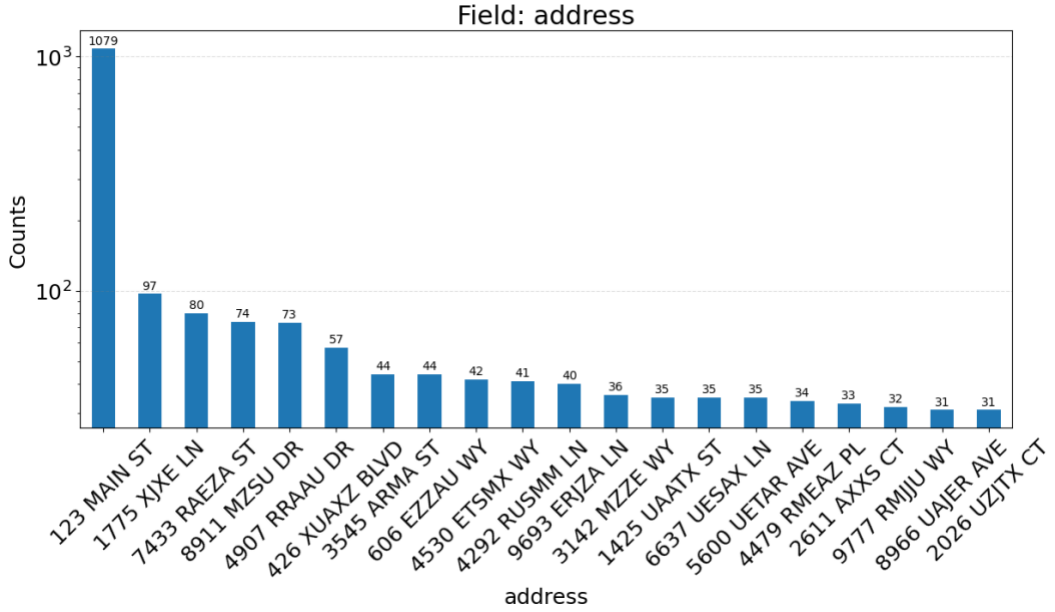
SSN

This field contains the social security number of the individual making the application.



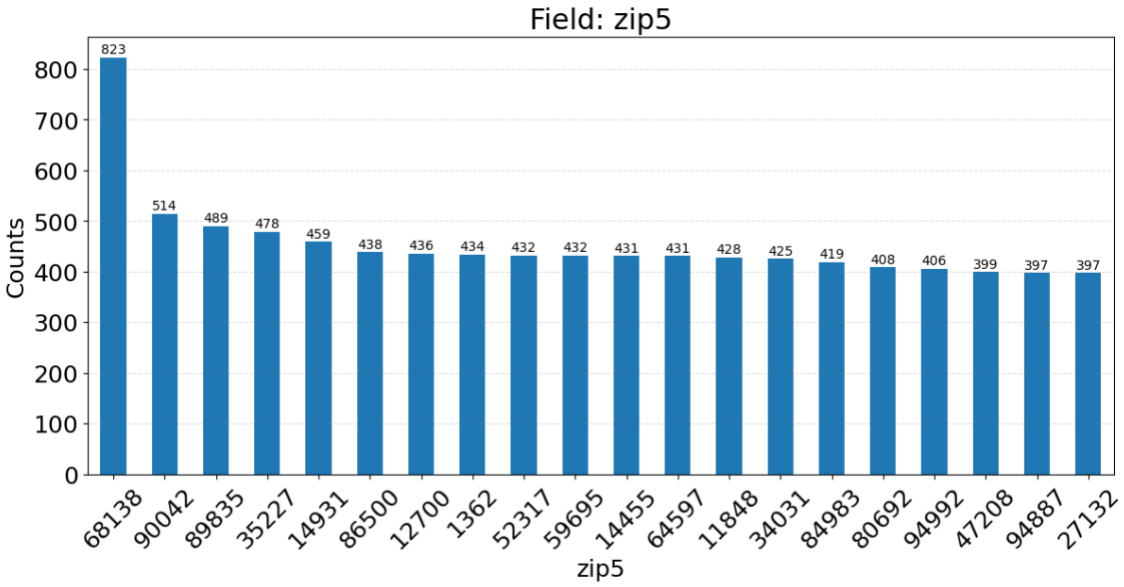
Address

The address associated with the application.



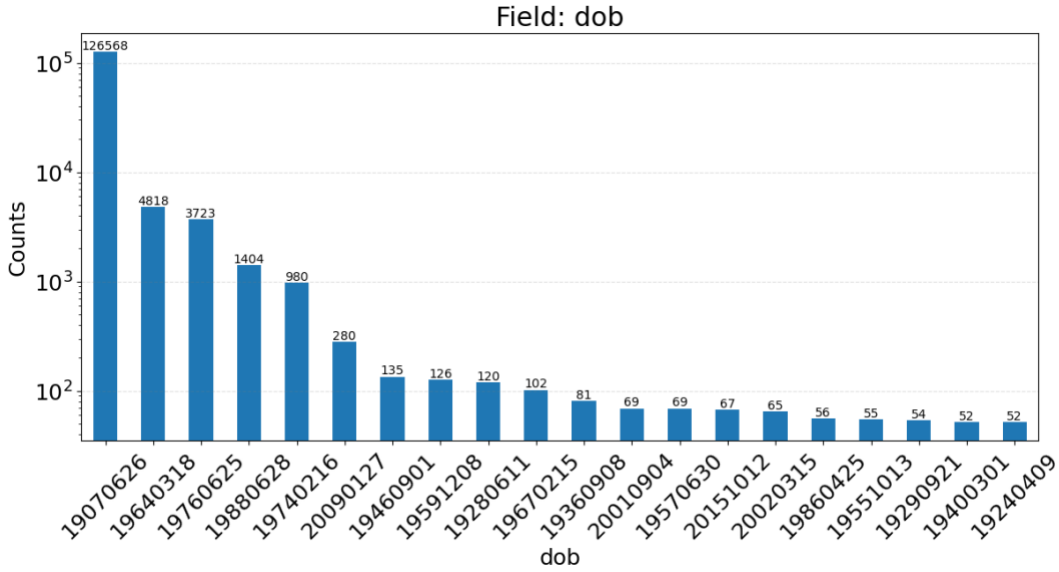
Zip5

The 5 digit zipcode tied to the application.



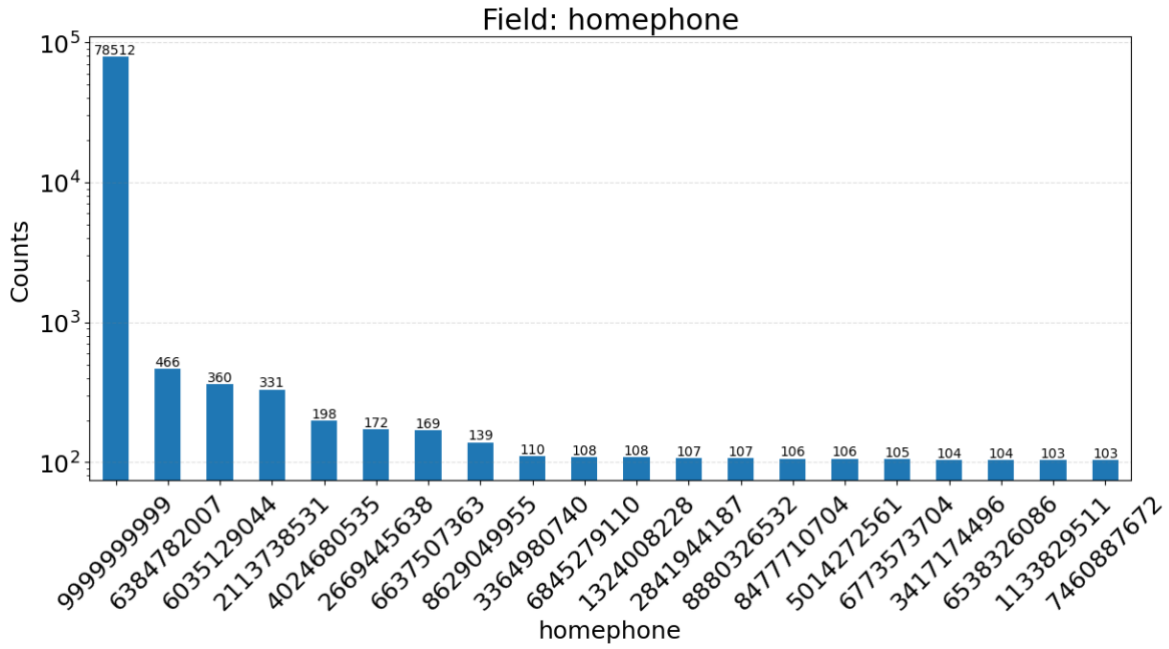
Dob

Date of birth of the individual making the application.



Homephone

Phone number of the individual making the application.



Key Insights from Field Distributions

Across multiple fields, we observe extreme repetition of specific values: 999999999 for SSN, 123 MAIN ST for address, 9999999999 for homephone, and 19070626 for DOB; each appearing thousands to hundreds of thousands of times. These outliers massively outpace other values and act as red flags for fraud. Further investigation confirms that many of these are frivolous **placeholders** used in lieu of missing or unverifiable information.

Data Cleaning

General Overview

No exclusions, imputations, or outlier treatments were applied to this dataset. However, specific attention was given to handling non-informative or *frivolous* identity elements that could distort linkage analysis and introduce false positives.

Handling Frivolous Identity Elements

Definition and Rationale

Frivolous values refer to system-generated placeholder values used to represent missing identity information in the dataset. These placeholders often follow predictable patterns and are reused across multiple records, which can lead to misleading inferences during identity resolution. Common examples include:

- **SSN**: 123456789, 999999999
- **Address**: "123 Main St"
- **Phone Number**: 1234567890

Such values are problematic because their repetition does **not** indicate a real-world connection between records. For instance, multiple records sharing the phone number 1234567890 likely reflect missing data rather than a genuine linkage. If left unaddressed, these patterns could trigger false positives in fraud detection models by incorrectly flagging records as being associated with the same individual.

Remediation Approach

To mitigate this issue, the following steps were taken:

1. **Detection**
   Frivolous values were identified through distributional analysis—values that appeared with unusually high frequency (e.g., SSN = 999999999) were flagged as synthetic placeholders.
2. **Replacement Strategy**
   - Each identified frivolous value was replaced with the corresponding record number to ensure uniqueness and prevent artificial linkages across records.
   - This method preserves record integrity without falsely signaling connections between unrelated entries.
3. **Affected Fields**
   The following fields were found to contain frivolous values and were handled accordingly:
   - SSN
   - Address
   - zip5
   - homephone
4. **Date of Birth Handling**
   A specific dummy date of birth value—**June 26, 1907**—was identified as non-informative. Records containing this value had the dob field

replaced with a randomly selected date between **1930 and 1980**, reflecting a more plausible age range for applicants while maintaining variability.

Variable Creation

Understanding Modes of Identity Fraud

To guide effective feature engineering, we first identified the primary modes through which identity fraud typically occurs. This understanding informs the types of variables that can provide meaningful signals to a fraud detection model. The two key fraud scenarios addressed in this dataset are:

- **Mode 1: One fraudster using multiple stolen identities**
  A single bad actor uses the PII (e.g., SSN, Name, Date of Birth) of multiple individuals, but submits applications with their own consistent contact details (e.g., Address, Phone Number).
  - **Key signals**: Reuse of the same address or phone number across multiple unique identities.
- **Mode 2: One stolen identity used by multiple fraudsters**
  A single victim's identity is compromised and distributed (e.g., via the dark web) to multiple actors. Each fraudster uses the same core PII but with different addresses, phone numbers, or other contact details.
  - **Key signals**: A single SSN linked to multiple distinct addresses, phone numbers, or names.

Engineered Identity Entities

To detect these patterns, we derived 22 key identity entities by combining PII and contact information. These entities form the basis for more complex feature engineering. Examples include:

Entity Name	Composition
fulladdress	Address + ZIP (zip5)
name	First Name + Last Name
name_dob	Name + Date of Birth
name_fulladdress	Name + Full Address
dob_homephone	Date of Birth + Phone Number
fulladdress_dob	Full Address + Date of Birth

fulladdress_homephone	Full Address + Phone Number
homephone_name_dob	Home Phone + Name + DOB
ssn_{var} combinations	SSN combined with various fields (e.g. name, address, dob, etc.)

Summary of Variables Created

Feature Type	Description	Count
Target Encoded	Fraud rate by day of week (dow_risk)	1
Recency Variables	Days since the entity was last seen (entity_days_since)	22
Velocity Variables	Count of how often the same entity appeared in the last 0/1/3/7/14/30 days (entity_count_x)	132
Relative Velocity Variables	Short-term vs long-term velocity comparisons (e.g. 1-day count vs 7-day count)	176
Cross-Entity Linkage Variables	Number of unique entity2 linked to the same entity1 within time windows (entity1_unique_count_for_entity2_t)	2772
Maximum Count Variables	Maximum number of times an entity appeared in rolling windows (max_count_by_entity_d)	88
Age at Application	Applicant's age at time of application (age_when_apply)	1
Age Indicators	Max, min, and mean age seen for each entity over time	66

Motivation and Methodology Behind Feature Engineering

The feature generation process began with the 22 base entities derived from core identity attributes. From these, we developed several categories of dynamic variables:

- **Velocity & Recency**: Capturing how frequently and recently an entity appeared within defined time windows (e.g., 1, 7, or 30 days), to detect abnormal application patterns or intense short-term reuse.
- **Cross-Entity Linkage**: Measuring how many distinct values one entity is linked to—for example, how many addresses are associated with the same SSN. This helps detect synthetic identities or fraud rings.
- **Relative Velocity**: Comparing recent entity frequency (e.g., 1-day) to broader baselines (e.g., 30-day) to spot sudden spikes in activity.
- **Max Frequency**: Tracking the highest occurrence of an entity across different windows to flag excessive historical usage.
- **Age Consistency Checks**: Identifying inconsistencies in age over time by examining the range of application ages associated with a single entity—e.g., an SSN that has been linked to both a 22-year-old and a 68-year-old raises suspicion.

Purpose and Impact

Together, these engineered features are designed to uncover key markers of identity fraud: frequent or sudden reuse of PII, inconsistent personal details, widespread sharing of contact info, and temporal anomalies. These signals are critical inputs to machine learning models tasked with distinguishing legitimate applications from fraudulent ones.

Feature Selection Process

Following the feature engineering stage, we were left with thousands of candidate variables. While this high-dimensional feature set offered the potential for strong signal extraction, it also introduced the risk of model degradation due to the curse of dimensionality, a well-known issue where increasing the number of features can lead to overfitting and reduced generalizability.

To address this, we implemented a three-step feature selection framework to isolate the most predictive and stable features for our final model:

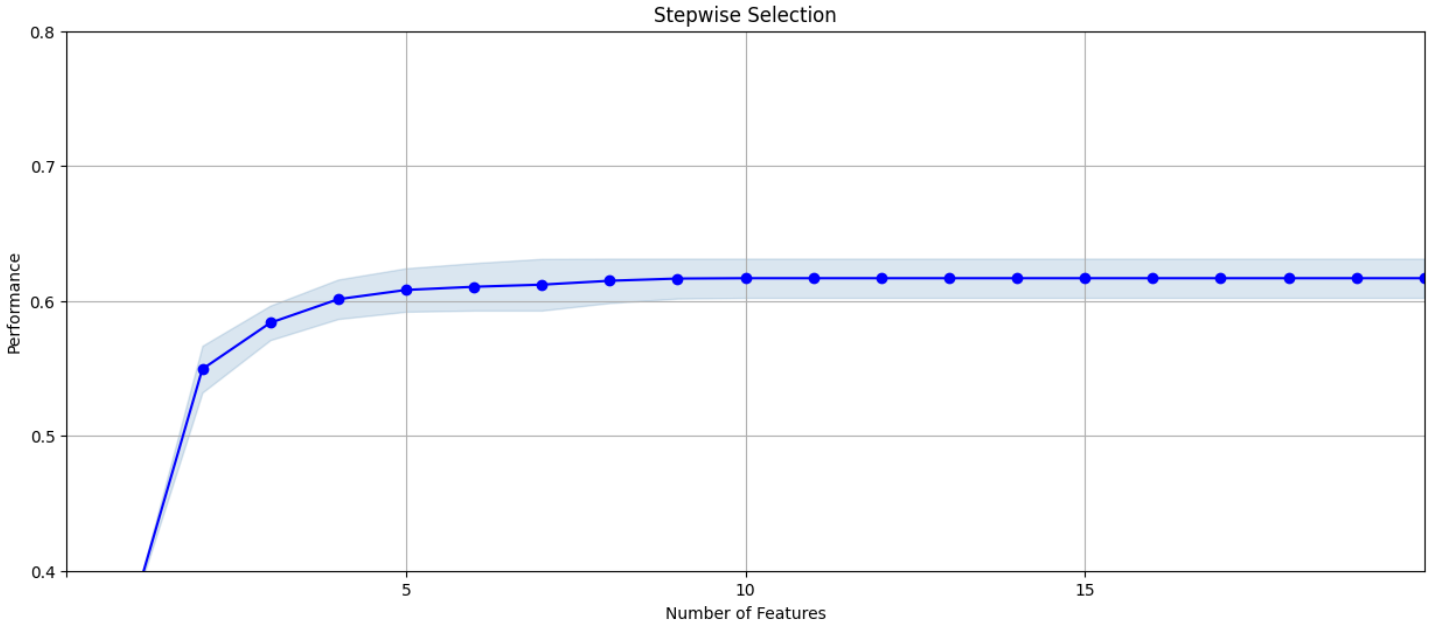
1. **Filter Step:**
   We began with a univariate filtering approach, ranking all candidate features using the **Kolmogorov-Smirnov (K-S) statistic**, which measures each variable's ability to distinguish between good and bad classes. The top ~20% of variables from this ranking were selected as inputs for the next stage.
2. **Wrapper Step:**
   A wrapper-based strategy was then employed using non-linear models such as **LightGBM (LGBM)** to evaluate feature subsets through **greedy search techniques**—both forward and backward selection. The objective was to optimize performance based on the **Fraud Detection Rate at 3% (FDR@3%)**. Multiple configurations were tested, varying:
   - The number of variables selected from the filter stage
   - Types of wrapper models
   - Selection strategies (forward vs. backward)
   - Final variable counts for evaluation
3. **Model Evaluation and Final Selection:**
   For each combination, we reviewed the **wrapper saturation graph**, which plots model performance (FDR@3%) against the number of selected features. The configuration yielding the **highest FDR@3% at the lowest feasible dimensionality** was chosen. From this optimal wrapper run, the **top 20 features** (ranked by multivariate importance) were selected to form the final input set for our model.

Final Variables with Filter Score

wrapper order	variable	filter score
1	max_count_by_address_30	0.36
2	max_count_by_fulladdress_7	0.34
3	max_count_by_fulladdress_7	0.33
4	address_day_since	0.33
5	fulladdress_count_30	0.33
6	fulladdress_day_since	0.33
7	address_count _13	0.33
8	fulladdress_count	0.33
9	max_count_by_address_3	0.33
10	address_count_14	0.32
11	fulladdress_count_14	0.32
12	max_count_by_address_1	0.31
13	fulladdress_count_7	0.30
14	address_count_7	0.30
15	address_count_0_by_30	0.29
16	max_count_by_ssn_dob_7	0.23

17	ssn_count_30	0.23
18	max_count_by_homephone_3	0.22
19	zip5_count_1	0.22
20	max_count_by_homephone_30	0.22

of Variables vs Performance Plot



Model Exploration

The objective of this phase was to evaluate a range of non-linear machine learning models and identify the optimal hyperparameter configurations that maximize performance—measured by **Fraud Detection Rate at 3% (FDR@3%)**, while minimizing the risk of overfitting.

The process followed these key steps:

1. **Model Exploration and Tuning:**
   Multiple non-linear model families were explored (e.g., gradient boosting, random forests, etc.). For each model type, an extensive hyperparameter grid was defined and systematically tested.

2. **Repeated Training for Robustness:**
   Each hyperparameter configuration was evaluated by training the model **10 times**, capturing the average FDR@3% across **training, validation (test), and out-of-time (OOT)** datasets. This repetition ensured that the results were robust to variance in training splits.
3. **Performance Tracking:**
   A comprehensive results table was constructed, documenting each hyperparameter combination along with its corresponding FDR@3% scores across all datasets.
4. **Best Configuration per Model Type:**
   For each model type, the hyperparameter set yielding the most favorable balance between high test/OOT performance and low overfitting was selected as the **final configuration** for that model.
5. **Final Model Selection:**
   A **box plot** was generated to visualize the distribution of FDR@3% scores across training, test, and OOT datasets for each model type. This comparative visualization guided the selection of the **final model**, which demonstrated both high predictive performance and stability across data splits.

Tests Performed

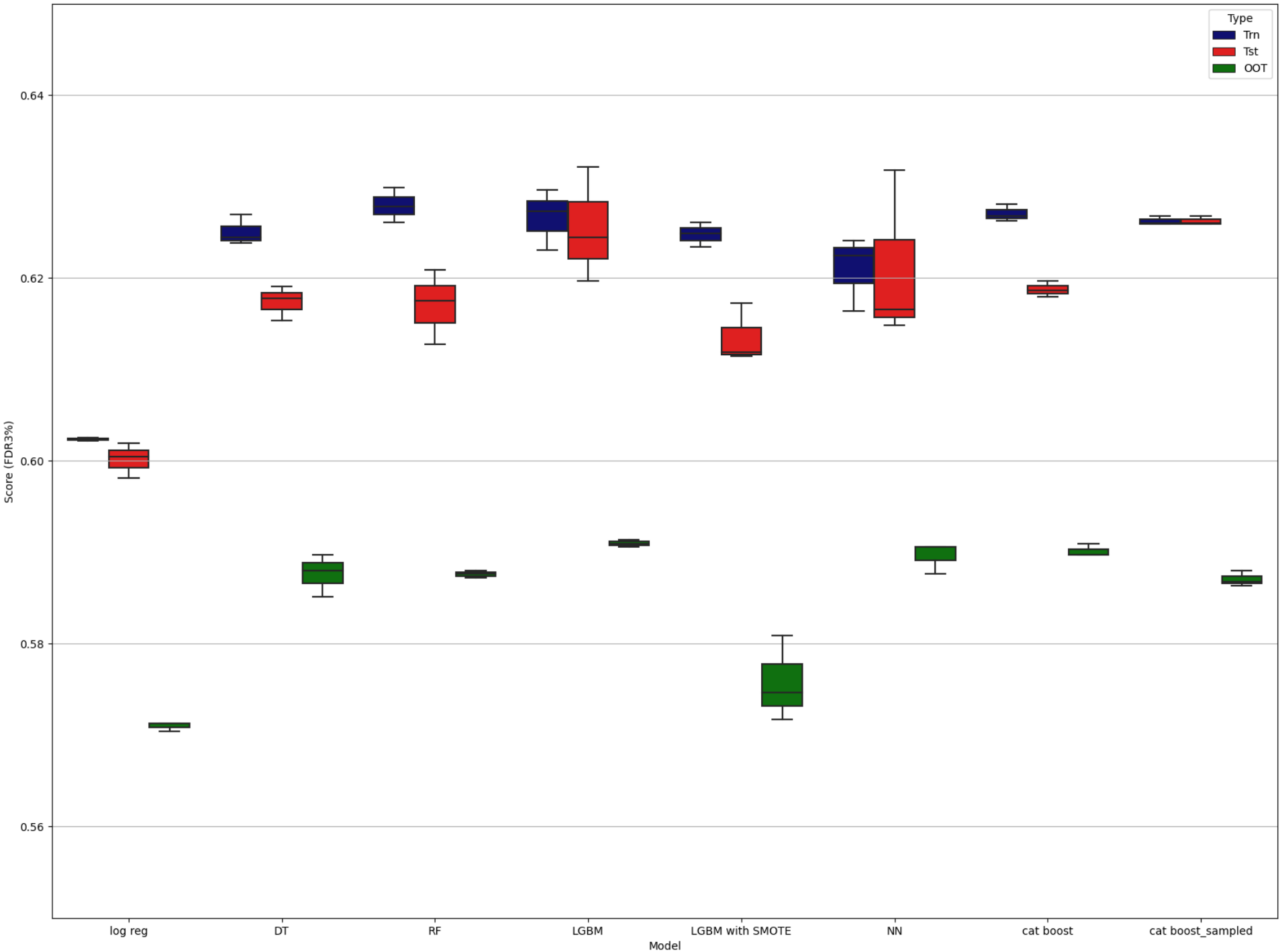
Model	Parameters						Average FDR @ 3%		
Logistic Regression	# Variables	Penalty	Solver	L1 ratio			Train	Test	OOT
1	20	None	Lbfgs	None			0.60	0.60	0.57
Decision Tree	# Variables	Criterion	Splitter	Max depth	Min samples	Min Samples leaf	Train	Test	OOT
1*	20	Gini	Random	None	200	100	0.63	0.62	0.58
2	20	Gini	Best	None	200	100	0.64	0.61	0.58
3	15	Gini	Random	None	200	100	0.63	0.62	0.58
4	20	Gini	Random	20	200	100	0.63	0.62	0.58
5	20	Gini	Random	10	150	75	0.60	0.61	0.57
Random Forest	# Variables	Max depth	N estimators	Max Features	Min Samples split	Min Samples leaves	Train	Test	OOT
1	20	12	20	8	60	30	0.62	0.63	0.59

2*	20	12	50	Log2	60		30		0.62	0.62	0.59
3	20	22	50	Log2	45		22		0.64	0.62	0.59
4	20	15	50	Sqrt	45		22		0.63	0.62	0.59
5	20	15	30	12	50		25		0.63	0.62	0.59
LGBM	# variables	N estimators	Learning rate	Max depth	Num leaves		Min Child samples		Train	Test	OOT
1	20	500	0.1	3	5		30		0.63	0.62	0.59
2	20	1000	0.05	3	5		30		0.62	0.63	0.59
3*	20	1000	0.05	5	5		30		0.63	0.63	0.59
4	20	1000	0.05	5	5		10		0.63	0.62	0.59
Neural Network	# Variables	Solver	Activation Function	Hidden Layer Sizes	Alpha	Learning Rate Init		Leaning Rate	Train	Test	OOT
1*	20	Adam	Relu	(10, 10)	0.005	0.002		Adaptive	0.62	0.63	0.59

2	20	Adam	Relu	(10,10)	0.005	0.007	Adaptive	0.62	0.62	0.58
3	20	Adam	Relu	(20,20)	0.005	0.01	Adaptive	0.62	0.63	0.59
4	20	Adam	Tanh	(20,20)	0.005	0.01	Adaptive	0.62	0.62	0.58
CatBoost	# Variables	N_estimators	Learning rate	Depth	L2 leaf reg	Loss Function		Train	Test	OOT
1	20	20	0.1	8	5	Logloss		0.62	0.61	0.58
2*	20	40	0.1	5	8	Logloss		0.62	0.62	0.59
3	20	80	0.05	4	8	Logloss		0.62	0.61	0.58
4	20	20	0.1	4	1	Logloss		0.61	0.61	0.57
XGBoost	# Variables	Max_depth	Min_child_weight	Gamma	Learning_rate	N_estimators		Train	Test	OOT
1	20	5	75	0.01	0.05	500		0.63	0.62	0.59
2	20	5	75	0.01	0.001	1000		0.61	0.62	0.58

3	20	5	75	0.01	0.1	250	0.63	0.62	0.59
4*	20	8	75	0.01	0.1	250	0.63	0.63	0.59

Model Comparison Plot


Type
Trn
Tst
OOT
0.64
0.62
0.60
0.58
0.56
Score (FDR3%)
log reg
DT
RF
LGBM
LGBM with SMOTE
NN
cat boost
cat boost_sampled
Model

Final Model Performance

Based on a comprehensive evaluation of model performance across training (Trn), test (Tst), and out-of-time (OOT) datasets, **LightGBM (LGBM)** was selected as the final model for deployment.

LGBM achieved:

- **Test FDR@3% Mean:** 0.6254 (Std: 0.0063)
- **OOT FDR@3% Mean:** 0.5909 (Std: 0.0004)

These results reflect a strong balance between predictive power and generalization, with **minimal overfitting** and **low performance variance**, particularly on the OOT dataset—critical for real-world fraud detection. While models like CatBoost and Random Forest showed competitive train/test scores, LGBM demonstrated **the most consistent and stable performance across all data splits**.

Given its robustness, efficiency, and scalability, **LGBM was selected as the optimal model** for final deployment.

Performance on train data

Training	# Records		# Goods		# Bads		Fraud Rate					
	583,454		575,105		8,349		1.43 %					
	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
0	0	0	0	0	0.00	0.00	0	0	0	0.00	0.00	0.00
1	1	5,835	1,103	4,732	18.90	81.10	5,835	1,103	4,732	0.19	56.35	56.15
2	2	5,834	5,395	439	92.48	7.52	11,669	6,498	5,171	1.13	61.57	60.44
3	3	5,835	5,737	98	98.32	1.68	17,504	12,235	5,269	2.13	62.74	60.61
4	4	5,834	5,780	54	99.07	0.93	23,338	18,015	5,323	3.13	63.38	60.25
5	5	5,835	5,772	63	98.92	1.08	29,173	23,787	5,386	4.14	64.13	60.00
6	6	5,834	5,789	45	99.23	0.77	35,007	29,576	5,431	5.14	64.67	59.53
7	7	5,835	5,766	69	98.82	1.18	40,842	35,342	5,500	6.15	65.49	59.35
8	8	5,834	5,792	42	99.28	0.72	46,676	41,134	5,542	7.15	65.99	58.84
9	9	5,835	5,786	49	99.16	0.84	52,511	46,920	5,591	8.16	66.58	58.42
10	10	5,834	5,803	31	99.47	0.53	58,345	52,723	5,622	9.17	66.94	57.78

11	11	5,835	5,800	35	99.40	0.60	64,180	58,523	5,657	10.18	67.36	57.18
12	12	5,834	5,793	41	99.30	0.70	70,014	64,316	5,698	11.18	67.85	56.67
13	13	5,835	5,791	44	99.25	0.75	75,849	70,107	5,742	12.19	68.37	56.18
14	14	5,835	5,804	31	99.47	0.53	81,684	75,911	5,773	13.20	68.74	55.54
15	15	5,834	5,796	38	99.35	0.65	87,518	81,707	5,811	14.21	69.20	54.99
16	16	5,835	5,796	39	99.33	0.67	93,353	87,503	5,850	15.22	69.66	54.44
17	17	5,834	5,795	39	99.33	0.67	99,187	93,298	5,889	16.22	70.12	53.90
18	18	5,835	5,803	32	99.45	0.55	105,022	99,101	5,921	17.23	70.50	53.27
19	19	5,834	5,792	42	99.28	0.72	110,856	104,893	5,963	18.24	71.01	52.76
20	20	5,835	5,799	36	99.38	0.62	116,691	110,692	5,999	19.25	71.43	52.18

Performance on test data

Test	**# Records**		**# Goods**		**# Bads**		**Fraud Rate**					
	250,053		**246,395**		**3,658**		**1.46%**					
	Bin Statistics					**Cumulative Statistics**						
Population Bin %	**# Records**	**# Goods**	**# Bads**	**% Goods**	**% Bads**	**Total # Records**	**Cumulative Goods**	**Cumulative Bads**	**% Goods**	**% Bads (FDR)**	**KS**	**FPR**
0	0	0	0	0.00	0.00	0	0	0	0.00	0.00	0.00	0.00
1	1	2,501	467	2034.00	18.67	81	2,501	467	2034.00	0.19	56.36	56.17
2	2	2,500	2,315	185.00	92.60	7	5,001	2,782	2219.00	1.13	61.49	60.36
3	3	2,501	2,466	35.00	98.60	1	7,502	5,248	2254.00	2.13	62.45	60.33
4	4	2,500	2,479	21.00	99.16	1	10,002	7,727	2275.00	3.14	63.04	59.90
5	5	2,501	2,483	18.00	99.28	1	12,503	10,210	2293.00	4.14	63.54	59.39
6	6	2,500	2,486	14.00	99.44	1	15,003	12,696	2307.00	5.15	63.92	58.77
7	7	2,501	2,482	19.00	99.24	1	17,504	15,178	2326.00	6.16	64.45	58.29

8	8	2,500	2,479	21.00	99.16	1	20,004	17,657	2347.00	7.16	65.03	57.87
9	9	2,501	2,480	21.00	99.16	1	22,505	20,137	2368.00	8.17	65.61	57.44
10	10	2,500	2,485	15.00	99.40	1	25,005	22,622	2383.00	9.18	66.03	56.85
11	11	2,501	2,489	12.00	99.52	0	27,506	25,111	2395.00	10.19	66.36	56.17
12	12	2,500	2,485	15.00	99.40	1	30,006	27,596	2410.00	11.20	66.78	55.58
13	13	2,501	2,488	13.00	99.48	1	32,507	30,084	2423.00	12.21	67.14	54.93
14	14	2,500	2,483	17.00	99.32	1	35,007	32,567	2440.00	13.21	67.61	54.39
15	15	2,501	2,489	12.00	99.52	0	37,508	35,056	2452.00	14.22	67.94	53.72
16	16	2,500	2,479	21.00	99.16	1	40,008	37,535	2473.00	15.23	68.52	53.29
17	17	2,501	2,489	12.00	99.52	0	42,509	40,024	2485.00	16.24	68.86	52.62
18	18	2,501	2,485	16.00	99.36	1	45,010	42,509	2501.00	17.25	69.30	52.05
19	19	2,500	2,488	12.00	99.52	0	47,510	44,997	2513.00	18.26	69.63	51.37
20	20	2,501	2,490	11.00	99.56	0	50,011	47,487	2524.00	19.27	69.94	50.67

Performance on OOT data

OOT	**# Records**		**# Goods**		**# Bads**		**Fraud Rate**					
	166,493		**164,107**		**2,386**		**1.43%**					
	Bin Statistics					**Cumulative Statistics**						
Population Bin %	**# Records**	**# Goods**	**# Bads**	**% Goods**	**% Bads**	**Total # Records**	**Cumulative Goods**	**Cumulative Bads**	**% Goods**	**% Bads (FDR)**	**KS**	**FPR**
0	0	0	0	0.00	0.00	0	0	0	0.00	0.00	0.00	0.00

1	1	1,665	390	1275.00	23.42	77	1,665	390	1275.00	0.24	53.44	53.20
2	2	1,665	1,556	109.00	93.45	7	3,330	1,946	1384.00	1.19	58.01	56.82
3	3	1,665	1,639	26.00	98.44	2	4,995	3,585	1410.00	2.18	59.09	56.91
4	4	1,665	1,651	14.00	99.16	1	6,660	5,236	1424.00	3.19	59.68	56.49
5	5	1,665	1,653	12.00	99.28	1	8,325	6,889	1436.00	4.20	60.18	55.99
6	6	1,665	1,655	10.00	99.40	1	9,990	8,544	1446.00	5.21	60.60	55.40
7	7	1,665	1,649	16.00	99.04	1	11,655	10,193	1462.00	6.21	61.27	55.06
8	8	1,664	1,650	14.00	99.16	1	13,319	11,843	1476.00	7.22	61.86	54.64
9	9	1,665	1,648	17.00	98.98	1	14,984	13,491	1493.00	8.22	62.57	54.35
10	10	1,665	1,653	12.00	99.28	1	16,649	15,144	1505.00	9.23	63.08	53.85
11	11	1,665	1,653	12.00	99.28	1	18,314	16,797	1517.00	10.24	63.58	53.34
12	12	1,665	1,661	4.00	99.76	0	19,979	18,458	1521.00	11.25	63.75	52.50
13	13	1,665	1,653	12.00	99.28	1	21,644	20,111	1533.00	12.25	64.25	51.99
14	14	1,665	1,654	11.00	99.34	1	23,309	21,765	1544.00	13.26	64.71	51.45
15	15	1,665	1,656	9.00	99.46	1	24,974	23,421	1553.00	14.27	65.09	50.82
16	16	1,665	1,657	8.00	99.52	0	26,639	25,078	1561.00	15.28	65.42	50.14
17	17	1,665	1,655	10.00	99.40	1	28,304	26,733	1571.00	16.29	65.84	49.55
18	18	1,665	1,655	10.00	99.40	1	29,969	28,388	1581.00	17.30	66.26	48.96
19	19	1,665	1,655	10.00	99.40	1	31,634	30,043	1591.00	18.31	66.68	48.37

20	20	1,665	1,653	12.00	99.28	1	33,299	31,696	1603.00	19.31	67.18	47.87

Financial curves

Cutoff vs Savings

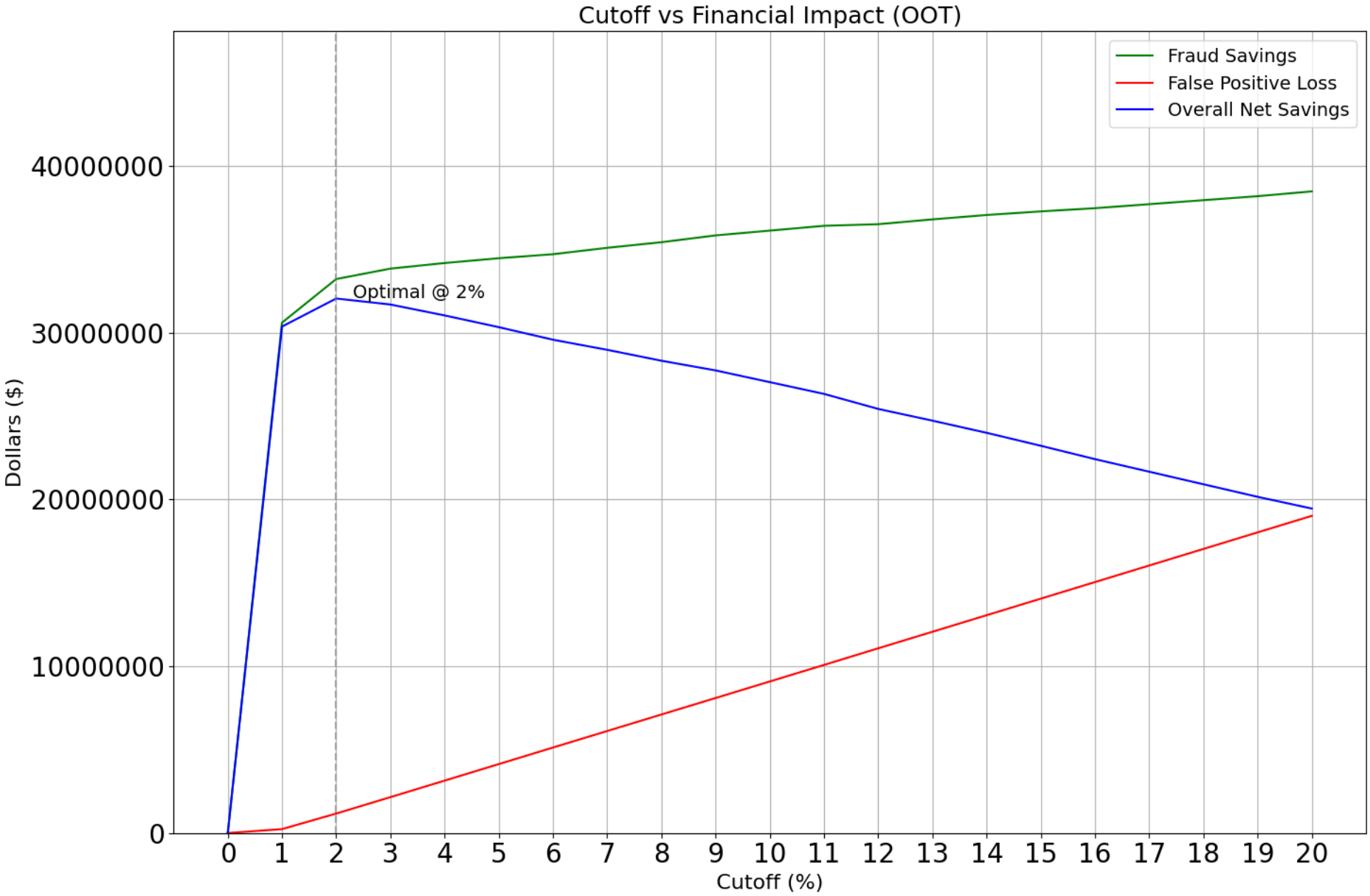
To evaluate the business value of fraud detection, we modeled the trade-off between fraud savings and false positive costs at various score cutoffs. We assumed an **average fraud loss of $4,000** and an **average false positive cost of $100** per rejected legitimate application.
The analysis shows that **maximum net savings of $32.05 million** occurs at the **2% cutoff**, where fraud detection is high and false positive losses remain relatively low. Beyond this point, while fraud savings continue to rise marginally, the cost of false positives increases sharply, leading to diminishing overall returns.
This validates the 2% cutoff as the optimal operational threshold, balancing fraud mitigation with customer experience and cost efficiency.

% Cutoff	Fraud Savings ($)	FP Loss ($)	Overall Savings ($)
0	0	0	0
1	30,600,000	234,000	30,366,000
2	33,216,000	1,167,600	32,048,400
3	33,840,000	2,151,000	31,689,000
4	34,176,000	3,141,600	31,034,400
5	34,464,000	4,133,400	30,330,600
6	34,704,000	5,126,400	29,577,600
7	35,088,000	6,115,800	28,972,200
8	35,424,000	7,105,800	28,318,200
9	35,832,000	8,094,600	27,737,400

Financial Impact Curve

The chart below reveals that the **optimal cutoff is at 2%**, where the model delivers **maximum net savings of approximately $32 million**. Beyond this point, while fraud savings continue to increase gradually, the steep rise in false positive losses leads to diminishing overall returns.

This validates 2% as the recommended operational threshold, balancing fraud prevention effectiveness with financial efficiency and customer retention.


Cutoff vs Financial Impact (OOT)
Fraud Savings
False Positive Loss
Overall Net Savings
Optimal @ 2%
Dollars ($)
0
10000000
20000000
30000000
40000000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Cutoff (%)

Summary/conclusions

The project began with a thorough data quality audit of a dataset containing 1 million loan applications, approximately **1.43%** of which were labeled as fraud. Several fields—including SSN, address, home phone, and date of birth—contained **placeholder or frivolous values**, which were identified and imputed to improve data integrity. The dataset spanned a full calendar year and was representative of real-world application patterns.

Feature engineering was guided by an understanding of synthetic identity fraud patterns, resulting in the creation of **thousands of candidate features**. To avoid the **curse of dimensionality**, we employed a two-stage feature selection strategy:

1. **Univariate filtering** using the Kolmogorov-Smirnov (K-S) statistic to shortlist the top 20% of features.
2. **Wrapper-based greedy search** (forward/backward) using models like LGBM to select the **top 20 multivariate features** that maximized FDR@3% on OOT data.

Next, we tuned hyperparameters across multiple non-linear models and evaluated performance across **train, test, and out-of-time (OOT)** datasets using repeated runs. The results were visualized via boxplots for each model to compare generalization, variance, and overfitting.

LightGBM (LGBM) was selected as the final model due to its strong and stable performance across splits. We then binned predictions into **5% population intervals**, enabling us to monitor model lift and cumulative fraud detection across segments.

To determine the ideal score cutoff for operational use, we generated a **Cutoff vs. Financial Impact** curve based on OOT data, scaled to annual volumes. This analysis incorporated an **average fraud loss of $4,000** and a **false positive cost of $100**. The curve revealed that a **2% cutoff** yields **maximum net financial savings (~$32 million annually)**, while minimizing customer disruption from false positives. Thus, our final recommendation is to **deny the top 2% of applications ranked most suspicious** by the model.

Statement of Model Performance

At the 2% recommended cutoff:

- **FDR@2% (OOT): 58.98%**
- This means the model successfully identifies **58.98% of all fraudulent applications** by reviewing just the top 2% of scored applications.
- The projected **annual financial savings** from this strategy are approximately **$32.05 million**, based on realistic fraud loss and false positive cost assumptions.

Appendix I – Data Quality Report

Description of the Data

This dataset has been synthetically generated to support the development of identity fraud detection algorithms, while ensuring compliance with legal and privacy regulations. Although it does not contain real Personally Identifiable Information (PII), it has been carefully designed to reflect realistic patterns observed in actual data, including PII frequency distributions and linkage characteristics.

The dataset contains **1,000,000 records** across **10 fields**, representing application data spanning the period from **January 1, 2017 to December 31, 2017**. The synthetic nature of the dataset ensures no real individuals are represented, while maintaining structural and statistical fidelity necessary for reliable model development and validation.

Numeric Fields Summary

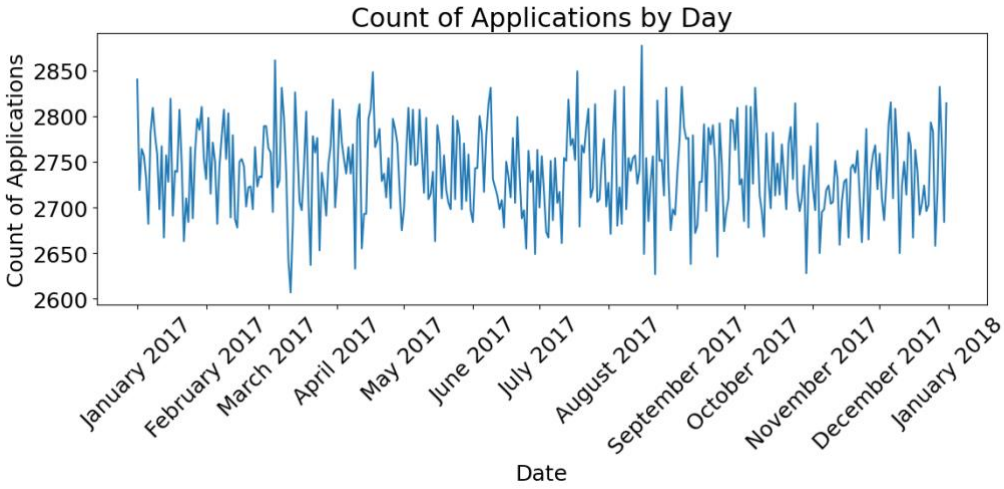
There are no numeric fields except perhaps date and dob, but I have included them in the categorical fields summary table.

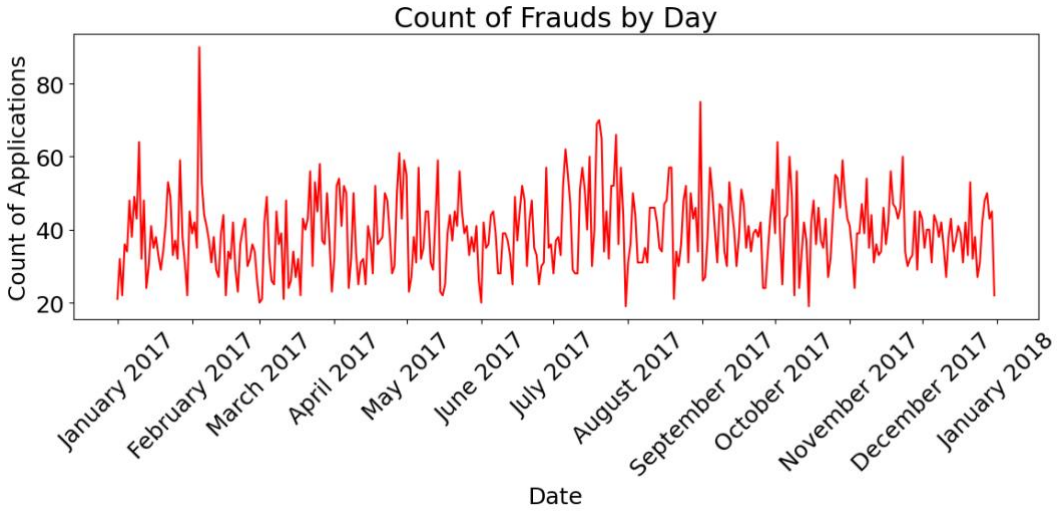
Categorical Fields Summary

Field Name	Field Type	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
homephone	categorical	1,000,000	100.00%	0	28,244	9999999999
fraud_label	categorical	1,000,000	100.00%	985,607	2	0
lastname	categorical	1,000,000	100.00%	0	177,001	ERJSAXA
zip5	categorical	1,000,000	100.00%	0	26,370	68138
firstname	categorical	1,000,000	100.00%	0	78,136	EAMSTRMT
address	categorical	1,000,000	100.00%	0	828,774	123 MAIN ST
record	categorical	1,000,000	100.00%	0	1,000,000	1
ssn	categorical	1,000,000	100.00%	0	835,819	999999999
date	categorical	1,000,000	100.00%	0	365	20170816
dob	categorical	1,000,000	100.00%	0	42,673	19070626

Field Distributions

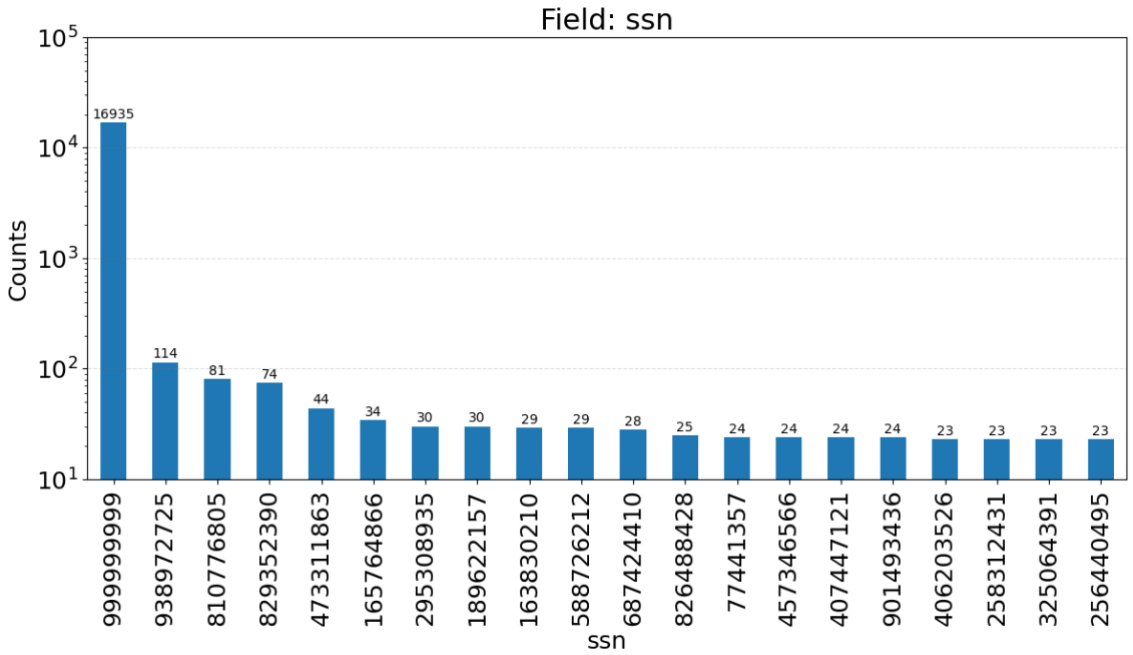
Record
Unique identifier of each application in the dataset.
Date
This field refers to the date the application was submitted.





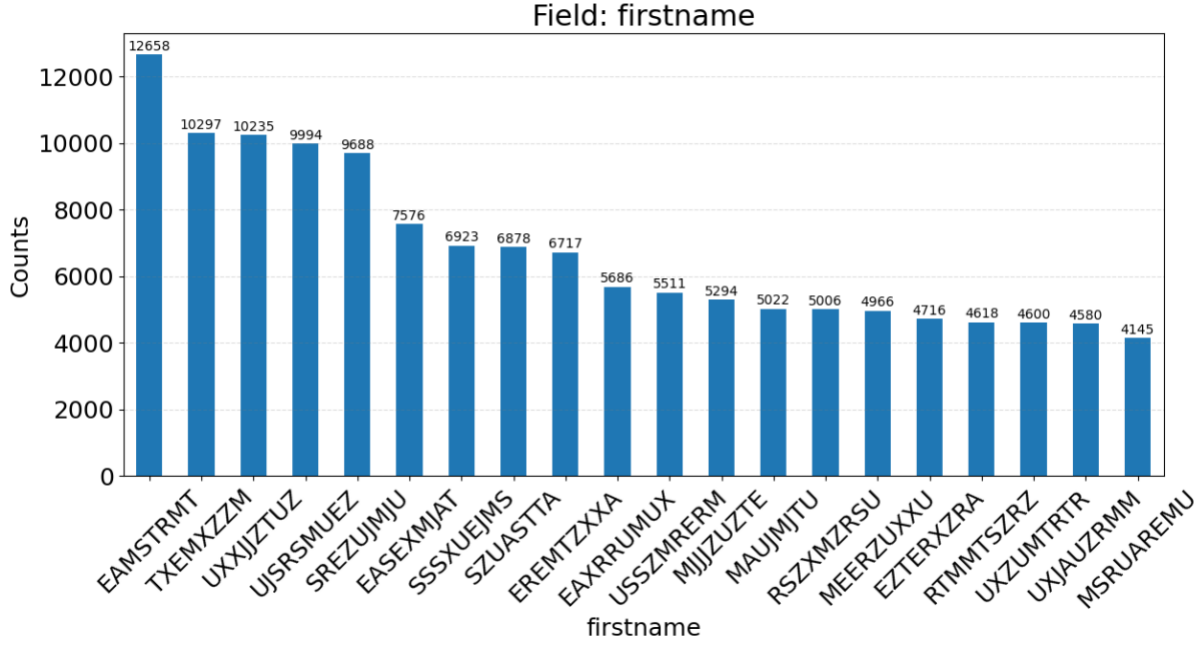
SSN

This field contains the social security number of the individual making the application.

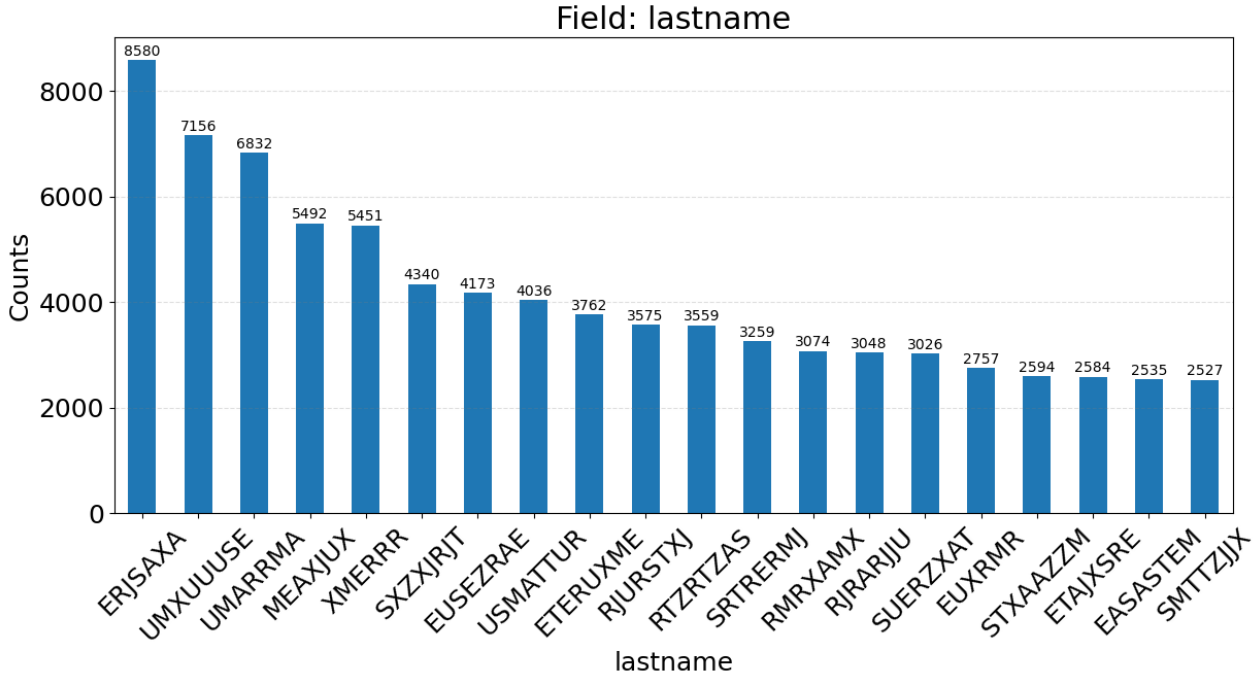


Firstname

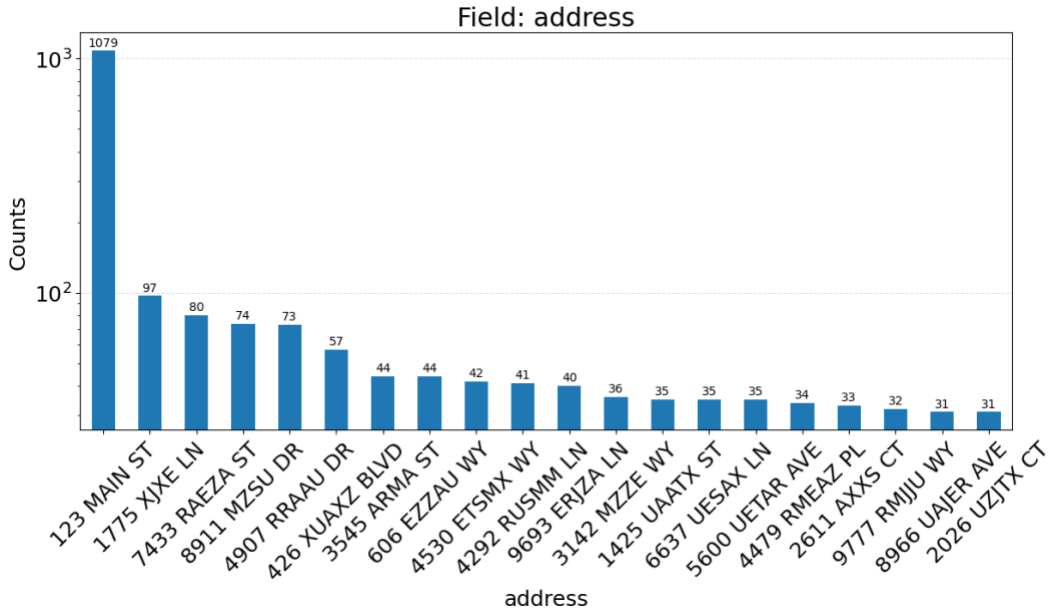
The first name of the person making the application.



Lastname
The last name of the person making the application.

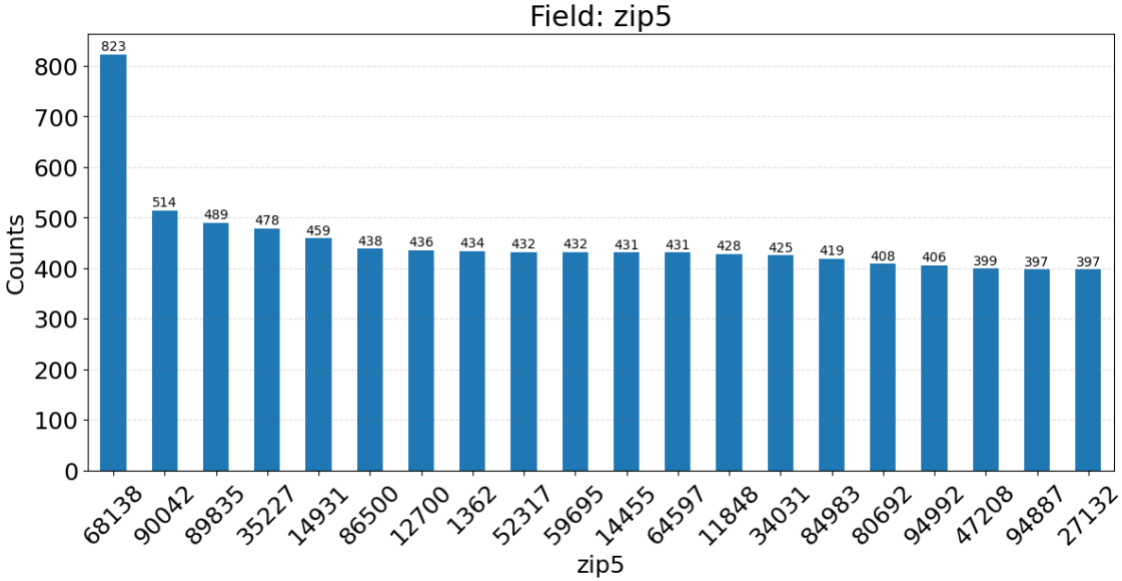


Address
The address associated with the application.



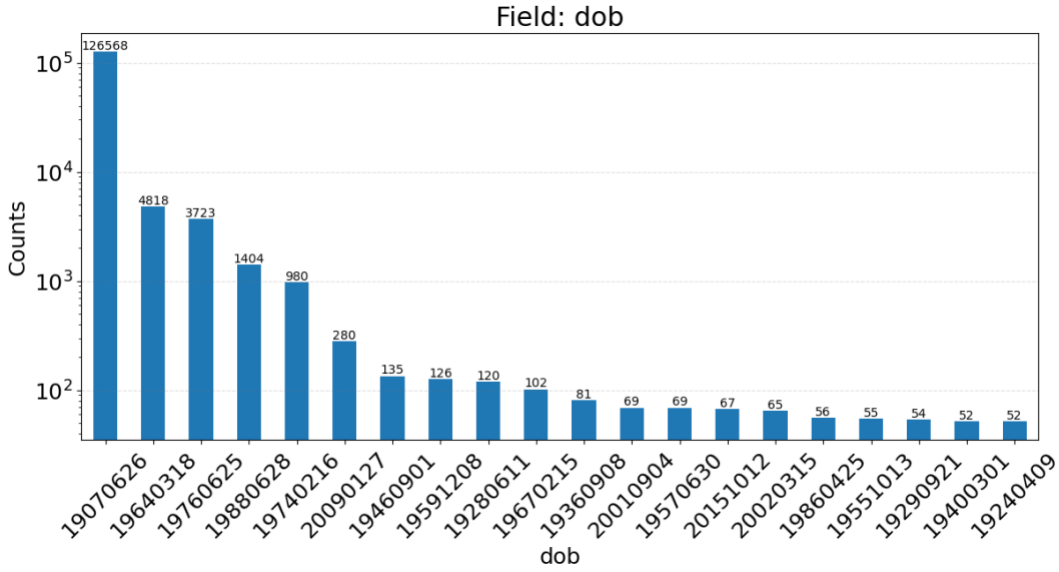
Zip5

The 5 digit zipcode tied to the application.



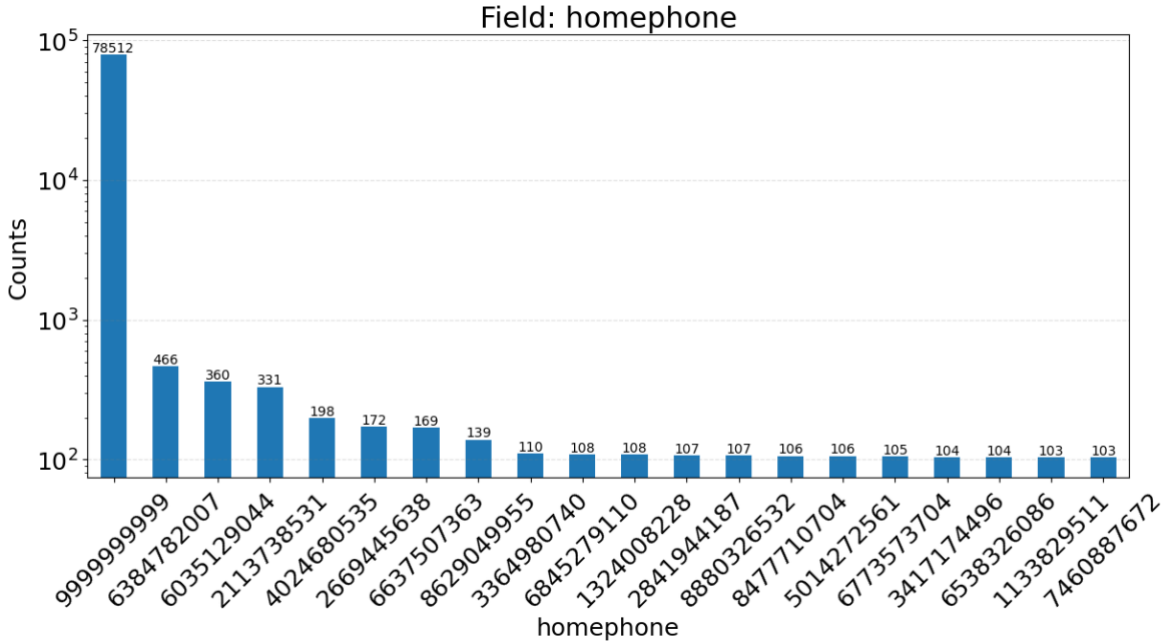
Dob

Date of birth of the individual making the application.



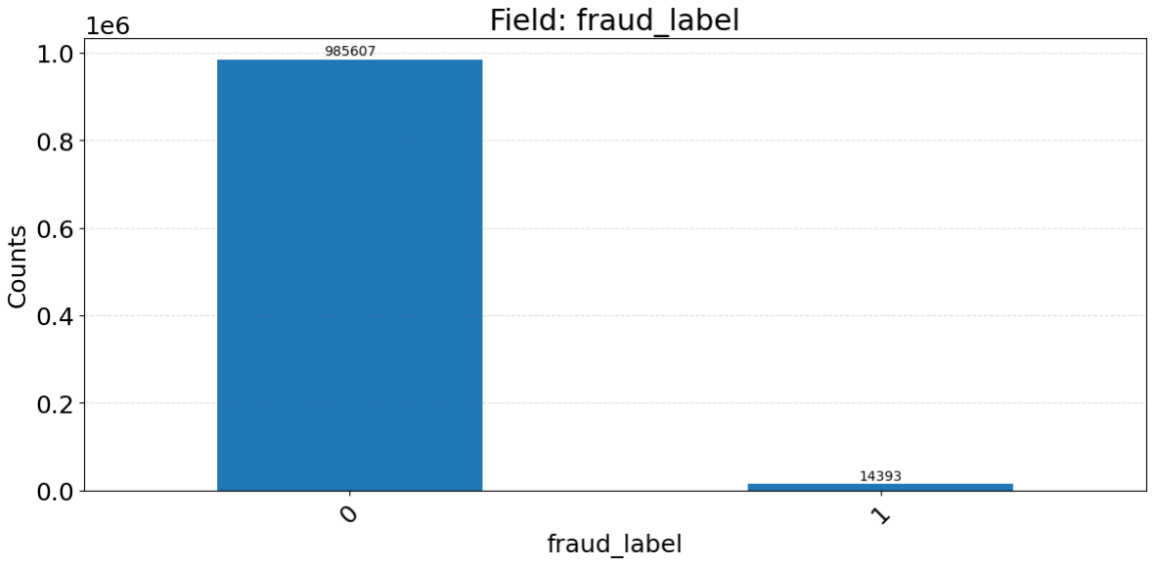
Homephone

Phone number of the individual making the application.



Fraud_label

Label denoting whether the application was fraudulent or not.


Field: fraud_label
1e6
1.0
0.8
0.6
0.4
0.2
0.0
Counts
985607
14393
0
1
fraud_label