Application Fraud Detection Using Machine Learning Models

DSO 562

Yixuan Liu

Table of Contents

Executive Summary

In this project, we developed a machine learning-based fraud detection model for identifying fraudulent product applications. Using the LightGBM classifier, we achieved a 59.72% Fraud Detection Rate (FDR) at a 3% score cutoff on the out-of-time dataset, effectively balancing fraud detection with minimizing false positives. The cutoff score was selected to optimize overall financial savings by maximizing fraud detection while minimizing lost revenue. The model is estimated to save the business approximately $3,205,800,000 by detecting fraudulent applications efficiently without significantly affecting legitimate sales.

Description of the Data

The raw data consists of product applications with both numerical and categorical fields. Numerical fields include date and dob (date of birth), which, strictly speaking, are not numerical variables. These two fields were later transformed into datetime variables. Categorical fields include firstname, lastname, address, record, ssn, zip5, homephone, and our target variable, fraud_label. The data is well-populated, with all fields having 100% completeness (0 null values). Below are detailed descriptive statistics for each of the fields.

Field Name	Field Type	# Records Have Values	% Populated	# Zeros	Min	Max	Mean	Standard Deviation	Most Common
date	numeric	1,000,000.00	100.0%	0	20,170,101.00	20,171,231.00	20,170,667.78	344.99	20,170,816.00
dob	numeric	1,000,000.00	100.0%	0	19,000,101.00	20,161,031.00	19,517,248.66	356,887.02	19,070,626.00

Table 1. Descriptive Statistics for Numerical Fields

Field Name	Field Type	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
firstname	categorical	1,000,000.00	100.0%	0.00	78,136.00	EAMSTRMT
lastname	categorical	1,000,000.00	100.0%	0.00	177,001.00	ERJSAXA
address	categorical	1,000,000.00	100.0%	0.00	828,774.00	123 MAIN ST
record	categorical	1,000,000.00	100.0%	0.00	1,000,000.00	1
fraud_label	categorical	1,000,000.00	100.0%	985,607.00	2.00	0
ssn	categorical	1,000,000.00	100.0%	0.00	835,819.00	999999999
zip5	categorical	1,000,000.00	100.0%	0.00	26,370.00	68138
homephone	categorical	1,000,000.00	100.0%	0.00	28,244.00	9999999999

Table 2. Descriptive Statistics for Categorical Fields

Among all the field distributions, there are a few needed to be highlighted.

“fraud_label”

As our target, the field “fraud_label” needs to be scrutinized. The distribution in Figure 1 shows the counts for the target variable "fraud_label". There are two classes represented: "0" for non-fraudulent applications and "1" for fraudulent applications.

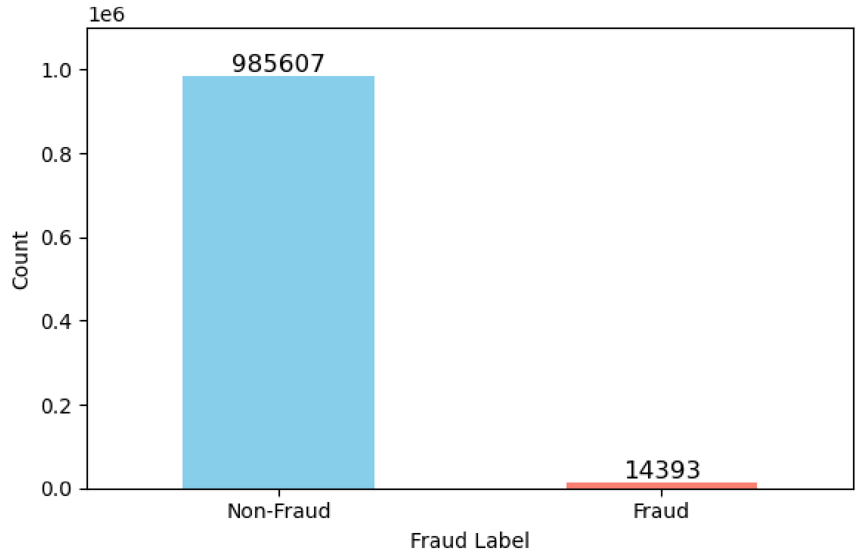


Figure 1. Distribution Plot for The Categorical Field “fraud_label”

The majority of the applications (labeled as "0") are non-fraudulent, with a count of 985,607. This represents a substantial class imbalance. Only a small fraction of the applications is fraudulent (labeled as "1"), with a count of 14,393. This imbalance between the two classes suggests that the data is highly skewed, with significantly more non-fraudulent applications than fraudulent ones.

“address”

The next noteworthy field is “address,” which shows an unusually high frequency of the value “123 MAIN ST,” appearing over 3,000 times, while all other values occur fewer than 100 times. This atypical distribution likely does not accurately reflect reality. We suspect that this unusually frequent value is most likely erroneous, a default entry, or a placeholder.

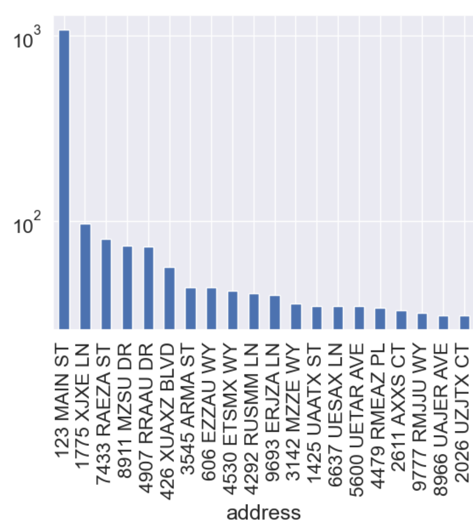


Figure 2. Distribution Plot for The Categorical Field “address”

“homephone”

Similar to the “address” field, the “homephone” field also displays an unusually frequent value of “999999999.” This pattern extends to the “ssn” and “dob” fields, where placeholder-like values appear with unusually high frequency.

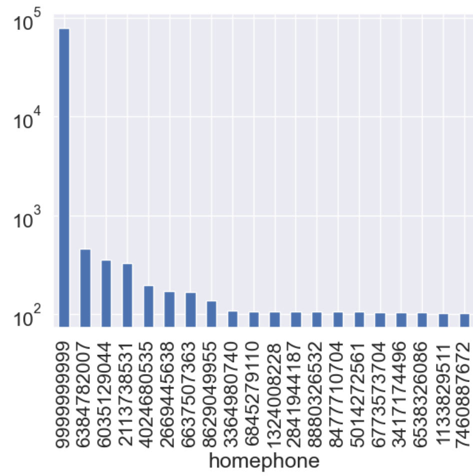


Figure 3. Distribution Plot for The Categorical Field **“**homephone”

In the subsequent sections, we will address these anomalous values.

Data Cleaning

Data Type Transformation

As noted, the fields “date” and “dob” are stored as integers in the dataset. We converted these fields to the correct datetime format to facilitate more accurate handling and analysis.

Unusual Value Treatment

During the data quality review, we identified several fields containing unusually frequent values, including “ssn”, “address”, “dob”, and “homephone”. These values are most likely erroneous, default entries, or placeholders. Treatment was applied to each of the field to better identify each record.

- **“ssn”**: The ‘ssn’ field has 16,935 instances of the value “999999999”, which we assume is used as a placeholder or default. To maintain data integrity, we replaced this value with a unique identifier from the “record” field.
- **“address”**: The address “123 MAIN ST” appears 1,079 times, indicating it may be a placeholder or erroneous entry. To minimize potential bias and enhance the accuracy of downstream analyses, we imputed these entries with unique values from the “record” field.
- **“dob”**: The date “1907-06-26” occurs 126,568 times, an unusually high frequency that suggests it is likely a placeholder rather than an actual birth date. We addressed this by imputing these occurrences with unique values from the “record” field.
- **“homephone”**: Similar to the “ssn” field, the value “9999999999” appears in 78,512 rows of the ‘homephone’ field. We treated this as a placeholder or default value and imputed it with unique identifiers from the “record” field.

Variable Creation

During the feature engineering process, a diverse set of variables was developed to capture meaningful patterns in the data and enhance the model's ability to detect fraud. Attributes for creating new variables were carefully selected based on their counts, excluding those with low counts in the group. This selection process resulted in 14 final attributes: **"ssn," "address," "zip5," "dob," "homephone," "name," "fulladdress," "name_dob," "name_fulladdress," "name_homephone," "fulladdress_dob," "fulladdress_homephone," "dob_homephone," "homephone_name_dob."**

Day-since features were created to track the recency of applications for each entity. **Velocity features** captured the number of applications associated with a specific entity over various timeframes (0, 1, 3, 7, 14, and 30 days). To complement these, **Velocity Change** features measured rapid shifts in application activity, comparing application counts in the most recent periods (0 or 1 days) to those observed over 3, 7, 14, and 30 days.

Unique Counts variables were added to capture connections between different entities by counting unique occurrences of an entity linked to another over timeframes of 0, 1, 3, 7, 14, 30, and 60 days. **Maximum Indicators** were designed to record the maximum count of an entity over 1, 3, 7, and 30 days. Additionally, **Age Indicators** were included to represent the maximum, mean, and minimum ages when an application was made for each entity.

During this process, duplicated variables emerged when linking entities. These duplicates were eliminated, resulting in a final set of 651 variables.

This comprehensive set of variables, encompassing temporal, applicational, and behavioral metrics, was developed to capture a broad spectrum of fraud-related signals, ultimately improving the dataset's value for fraud detection models.

Below is a summary table detailing the variables created:

Variable/Variable Family	Description	# of Variables Created
Day-since	Number of days since the last application for this entity	14
Velocity	Number of applications from this entity over the past [0,1,3,7,14,30] days	84
Relative Velocity	Number of applications with that entity group seen in the recent past [0,1] days over Numberof applications with that same entity group seen in the past [3, 7, 14, 30] days	112
Unique Counts	Number of unique occurences of an entity for another particular entity over the past [0, 1, 3, 7, 14, 30, 60] days	1274
Maximum Indicator	Maximum number of counts of an entity over the past [1, 3, 7, 30] days	56

Age Indicator	Maximum, mean, and minimum age when apply for each entity	42

Table 3. Description of Each New Variable/Variable Family

Feature Selection

For this project, forward selection using the LGBMClassifier with 600 filters was chosen as the primary feature selection method to select the final 20 variables. This method was selected because it consistently delivered strong performance and stability, exceeding a score of 0.6 after only a few iterations. Forward selection allows us to incrementally add features based on their contribution to improving model performance, ensuring that only the most relevant variables are included.

The final set of selected features includes a combination of various application count metrics, and entity linkage. This feature selection process was crucial in refining the dataset, allowing the model to focus on the variables that have the highest predictive power for fraud detection.

The final 20 selected variables are listed below:

Wrapper Order	Variable	Filter Score
1	max_count_by_address_30	0.3592
2	max_count_by_ssn_7	0.2274
3	max_count_by_homephone_7	0.2322
4	zip5_unique_count_for_dob_1	0.2191
5	max_count_by_fulladdress_30	0.3599
6	homephone_unique_count_for_fulladdress_14	0.0547
7	name_fulladdress_count_30	0.0677
8	fulladdress_homephone_unique_count_for_zip5_60	0.0038
9	address_count_30	0.3326
10	max_count_by_address_7	0.3433
11	fulladdress_day_since	0.3333
12	max_count_by_fulladdress_3	0.3295
13	max_count_by_address_3	0.3294
14	address_count_14	0.3224
15	fulladdress_count_14	0.3220
16	max_count_by_address_1	0.3153
17	max_count_by_fulladdress_1	0.3153
18	address_count_7	0.3017
19	fulladdress_count_7	0.3017
20	address_count_0_by_30	0.2919

Table 4. Univariate Filter Results of the Final 20 Variables

Preliminary Model Exploration

To classify fraudulent applications, we experimented with four types of machine learning models alongside a baseline logistic regression model. The models included:

- **DecisionTreeClassifier**
- **RandomForestClassifier**
- **BoostedTreeClassifier**
- **Neural Network**

Each model underwent extensive hyperparameter tuning, and the performance was compared based on their Fraud Detection Rate (FDR) across train, test, and out-of-time (OOT) datasets. The hyperparameters tested for each model are detailed in the table below. Hyperparameters that achieved the best performance by avoiding overfitting and maintaining low OOT FDR are highlighted in red.

Model	Parameter						Perofrmance		
Logistic Regression	Number of Variables	penalty	C	solver	max_iter		Train	Test	OOT
1	20	l2	1	lbfgs	1000		0.5842	0.5847	0.5504
2	20	l2	2	lbfgs	2000		0.5862	0.5819	0.5511
3	20	l2	0.1	lbfgs	2000		0.5854	0.5814	0.5501
4	20	l2	0.5	lbfgs	1000		0.5848	0.5842	0.5510
5	15	l2	1	lbfgs	1000		0.5861	0.5837	0.5516
Decision Tree	Number of Variables	critierion	splitter	max_depth	min_samples_split	min_samples_leaf	Train	Test	OOT
1	15	entropy	best	None	150	50	0.6275	0.6094	0.5712
2	15	gini	best	None	50	30	0.6334	0.6043	0.5645
3	20	entropy	best	None	10	5	0.6579	0.5996	0.5539
4	20	entropy	best	None	30	10	0.6493	0.6073	0.5575
5	20	gini	best	None	300	150	0.6181	0.6177	0.5742
6	20	gini	best	None	2	1	0.6623	0.5467	0.5079
Random Forest	Number of Variables	n_estimators	criterion	max_depth	min_samples_split	min_samples_leaf	Train	Test	OOT
1	15	100	gini	None	2	1	0.6611	0.6025	0.5649
2	15	100	entropy	None	150	50	0.6206	0.6236	0.5903
3	20	50	gini	None	100	10	0.6312	0.6262	0.5915
4	20	50	gini	20	100	10	0.6327	0.6097	0.5909
5	20	100	gini	30	20	5	0.6478	0.6186	0.5863
6	20	100	entropy	20	20	5	0.6383	0.6195	0.5893
Boosted Tree	Number of Variables	num_leaves	max_depth	learning_rate	n_estimators		Train	Test	OOT
1	15	31	-1	0.1	100		0.6291	0.6207	0.5883
2	15	50	-1	0.01	200		0.6232	0.6239	0.5919
3	15	100	30	0.01	400		0.6347	0.6184	0.5873
4	20	31	-1	0.01	200		0.6231	0.6255	0.5943
5	20	50	30	0.05	600		0.6425	0.6204	0.5858
6	20	50	-1	0.01	500		0.6297	0.6266	0.5915
Neural Network	Number of Variables	hidden_layer_sizes	activation	learning_rate	learning_rate_init	max_iter	Train	Test	OOT
1	15	(100,0)	relu	constant	0.001	200	0.6195	0.6168	0.5872
2	15	(150,0)	relu	adaptive	0.01	300	0.6141	0.6148	0.5798
3	15	(200,0)	relu	constant	0.0001	500	0.6156	0.6197	0.5852
4	20	(1000,0)	relu	adaptive	0.0001	2000	0.6191	0.6181	0.5900
5	20	(1000,0)	tanh	adaptive	0.0001	1000	0.6108	0.6179	0.5812
6	20	(500,0)	relu	constant	0.0005	1000	0.6208	0.6175	0.5902

Table 5. Hyperparameters Tested for Each Model Type and Their Performance

Key findings include:

- **LightGBM (LGBM)** was the best-performing model, demonstrating strong generalization across all datasets.
- **Neural Networks** performed well but had slightly lower OOT generalization compared to LGBM.
- **Random Forests** performed well on training data but struggled with generalization.
- **Decision Trees** performed worse than Random Forests across training, test, and OOT data.
- **Logistic Regression** showed decent training and test results but suffered from significant drops in OOT performance.

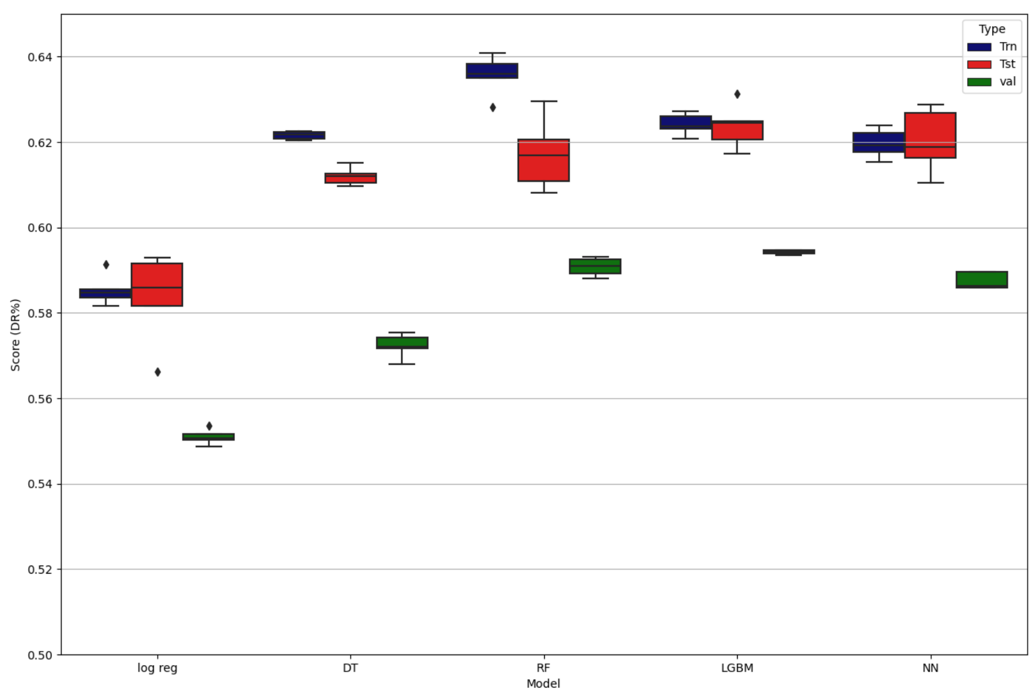


Figure 4. Performance Plot of Each Model Type

Final Model Performance

Based on the performance plots obtained in our previous exploration, we have decided to proceed with the LightGBM Classifier (LGBMClassifier) as our final model, utilizing the optimal hyperparameters identified. The model is configured with the following hyperparameters:

- Number of Leaves: The maximum number of leaves per tree is set to 31.
- Maximum Depth: There is no limit on the maximum depth of each individual tree.
- Learning Rate: The learning rate for boosting is set to 0.01.
- Number of Estimators: The model uses 200 trees (estimators) for boosting.

With these hyperparameters, we have produced three results tables corresponding to the training, testing, and out-of-time (OOT) datasets, respectively. We sorted the predicted probability of a application being fraudulent in a descending order and examined different population bin percentages to determine our fraud detection cutoff (i.e., what percentage of applications above cutoff is considered fraud).

Training	# Records					# Goods			# Bads				Fraud Rate		
	583,455					575,105			8,350				0.0143		
	Bin Statistics					Cumulative Statistics							Financial Statistics		
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Cumulative Goods	% Cumulative Bads(FDR)	KS	FPR	Fraud Savings	FP Loss	Overall Savings
1	5,835	1,173	4,662	20.10	79.90	5,835	1,173	4,662	0.20	55.83	55.63	0.25	18,648,000	117,300	18,530,700
2	5,834	5,416	418	92.84	7.16	11,669	6,589	5,080	1.15	60.84	59.69	1.30	20,320,000	658,900	19,661,100
3	5,835	5,669	166	97.16	2.84	17,504	12,258	5,246	2.13	62.83	60.69	2.34	20,984,000	1,225,800	19,758,200
4	5,834	5,760	74	98.73	1.27	23,338	18,018	5,320	3.13	63.71	60.58	3.39	21,280,000	1,801,800	19,478,200
5	5,835	5,770	65	98.89	1.11	29,173	23,788	5,385	4.14	64.49	60.35	4.42	21,540,000	2,378,800	19,161,200
6	5,834	5,786	48	99.18	0.82	35,007	29,574	5,433	5.14	65.07	59.92	5.44	21,732,000	2,957,400	18,774,600
7	5,835	5,789	46	99.21	0.79	40,842	35,363	5,479	6.15	65.62	59.47	6.45	21,916,000	3,536,300	18,379,700
8	5,834	5,799	35	99.40	0.60	46,676	41,162	5,514	7.16	66.04	58.88	7.46	22,056,000	4,116,200	17,939,800
9	5,835	5,798	37	99.37	0.63	52,511	46,960	5,551	8.17	66.48	58.31	8.46	22,204,000	4,696,000	17,508,000
10	5,835	5,804	31	99.47	0.53	58,346	52,764	5,582	9.17	66.85	57.68	9.45	22,328,000	5,276,400	17,051,600
11	5,834	5,798	36	99.38	0.62	64,180	58,562	5,618	10.18	67.28	57.10	10.42	22,472,000	5,856,200	16,615,800
12	5,835	5,796	39	99.33	0.67	70,015	64,358	5,657	11.19	67.75	56.56	11.38	22,628,000	6,435,800	16,192,200
13	5,834	5,788	46	99.21	0.79	75,849	70,146	5,703	12.20	68.30	56.10	12.30	22,812,000	7,014,600	15,797,400
14	5,835	5,807	28	99.52	0.48	81,684	75,953	5,731	13.21	68.63	55.43	13.25	22,924,000	7,595,300	15,328,700
15	5,834	5,802	32	99.45	0.55	87,518	81,755	5,763	14.22	69.02	54.80	14.19	23,052,000	8,175,500	14,876,500
16	5,835	5,798	37	99.37	0.63	93,353	87,553	5,800	15.22	69.46	54.24	15.10	23,200,000	8,755,300	14,444,700
17	5,834	5,796	38	99.35	0.65	99,187	93,349	5,838	16.23	69.92	53.68	15.99	23,352,000	9,334,900	14,017,100
18	5,835	5,794	41	99.30	0.70	105,022	99,143	5,879	17.24	70.41	53.17	16.86	23,516,000	9,914,300	13,601,700
19	5,834	5,801	33	99.43	0.57	110,856	104,944	5,912	18.25	70.80	52.55	17.75	23,648,000	10,494,400	13,153,600
20	5,835	5,815	20	99.66	0.34	116,691	110,759	5,932	19.26	71.04	51.78	18.67	23,728,000	11,075,900	12,652,100

Table 6. Model Results for Training Dataset (Top 20 Percent Score Cutoff)

Test	# Records					# Goods			# Bads				Fraud Rate		
	250,053					246,396			3,657				0.0146		
	Bin Statistics					Cumulative Statistics							Financial Statistics		
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Cumulative Goods	% Cumulative Bads(FDR)	KS	FPR	Fraud Savings	FP Loss	Overall Savings
1	2,501	527	1,974	21.07	78.93	2,501	527	1,974	0.21	53.98	53.76	0.27	7,896,000	52,700	7,843,300
2	2,500	2,294	206	91.76	8.24	5,001	2,821	2,180	1.14	59.61	58.47	1.29	8,720,000	282,100	8,437,900
3	2,501	2,420	81	96.76	3.24	7,502	5,241	2,261	2.13	61.83	59.70	2.32	9,044,000	524,100	8,519,900
4	2,500	2,456	44	98.24	1.76	10,002	7,697	2,305	3.12	63.03	59.91	3.34	9,220,000	769,700	8,450,300
5	2,501	2,482	19	99.24	0.76	12,503	10,179	2,324	4.13	63.55	59.42	4.38	9,296,000	1,017,900	8,278,100
6	2,500	2,480	20	99.20	0.80	15,003	12,659	2,344	5.14	64.10	58.96	5.40	9,376,000	1,265,900	8,110,100
7	2,501	2,488	13	99.48	0.52	17,504	15,147	2,357	6.15	64.45	58.30	6.43	9,428,000	1,514,700	7,913,300
8	2,500	2,478	22	99.12	0.88	20,004	17,625	2,379	7.15	65.05	57.90	7.41	9,516,000	1,762,500	7,753,500
9	2,501	2,483	18	99.28	0.72	22,505	20,108	2,397	8.16	65.55	57.38	8.39	9,588,000	2,010,800	7,577,200
10	2,500	2,487	13	99.48	0.52	25,005	22,595	2,410	9.17	65.90	56.73	9.38	9,640,000	2,259,500	7,380,500
11	2,501	2,472	29	98.84	1.16	27,506	25,067	2,439	10.17	66.69	56.52	10.28	9,756,000	2,506,700	7,249,300
12	2,500	2,485	15	99.40	0.60	30,006	27,552	2,454	11.18	67.10	55.92	11.23	9,816,000	2,755,200	7,060,800
13	2,501	2,487	14	99.44	0.56	32,507	30,039	2,468	12.19	67.49	55.30	12.17	9,872,000	3,003,900	6,868,100
14	2,500	2,479	21	99.16	0.84	35,007	32,518	2,489	13.20	68.06	54.86	13.06	9,956,000	3,251,800	6,704,200
15	2,501	2,488	13	99.48	0.52	37,508	35,006	2,502	14.21	68.42	54.21	13.99	10,008,000	3,500,600	6,507,400
16	2,500	2,490	10	99.60	0.40	40,008	37,496	2,512	15.22	68.69	53.47	14.93	10,048,000	3,749,600	6,298,400
17	2,501	2,483	18	99.28	0.72	42,509	39,979	2,530	16.23	69.18	52.96	15.80	10,120,000	3,997,900	6,122,100
18	2,501	2,485	16	99.36	0.64	45,010	42,464	2,546	17.23	69.62	52.39	16.68	10,184,000	4,246,400	5,937,600
19	2,500	2,482	18	99.28	0.72	47,510	44,946	2,564	18.24	70.11	51.87	17.53	10,256,000	4,494,600	5,761,400
20	2,501	2,490	11	99.56	0.44	50,011	47,436	2,575	19.25	70.41	51.16	18.42	10,300,000	4,743,600	5,556,400

Table 7. Model Results for Testing Dataset (Top 20 Percent Score Cutoff)

OOT	# Records					# Goods			# Bads				Fraud Rate		
	166,492					164,106			2,386				0.0143		
	Bin Statistics					Cumulative Statistics							Financial Statistics		
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Cumulative Goods	% Cumulative Bads(FDR)	KS	FPR	Fraud Savings	FP Loss	Overall Savings
1	1,665	423	1,242	25.41	74.59	1,665	423	1,242	0.26	52.05	51.80	0.34	4,968,000	42,300	4,925,700
2	1,665	1,549	116	93.03	6.97	3,330	1,972	1,358	1.20	56.92	55.71	1.45	5,432,000	197,200	5,234,800
3	1,665	1,598	67	95.98	4.02	4,995	3,570	1,425	2.18	59.72	57.55	2.51	5,700,000	357,000	5,343,000
4	1,665	1,653	12	99.28	0.72	6,660	5,223	1,437	3.18	60.23	57.04	3.63	5,748,000	522,300	5,225,700
5	1,665	1,656	9	99.46	0.54	8,325	6,879	1,446	4.19	60.60	56.41	4.76	5,784,000	687,900	5,096,100
6	1,665	1,650	15	99.10	0.90	9,990	8,529	1,461	5.20	61.23	56.03	5.84	5,844,000	852,900	4,991,100
7	1,664	1,656	8	99.52	0.48	11,654	10,185	1,469	6.21	61.57	55.36	6.93	5,876,000	1,018,500	4,857,500
8	1,665	1,656	9	99.46	0.54	13,319	11,841	1,478	7.22	61.94	54.73	8.01	5,912,000	1,184,100	4,727,900
9	1,665	1,653	12	99.28	0.72	14,984	13,494	1,490	8.22	62.45	54.22	9.06	5,960,000	1,349,400	4,610,600
10	1,665	1,657	8	99.52	0.48	16,649	15,151	1,498	9.23	62.78	53.55	10.11	5,992,000	1,515,100	4,476,900
11	1,665	1,655	10	99.40	0.60	18,314	16,806	1,508	10.24	63.20	52.96	11.14	6,032,000	1,680,600	4,351,400
12	1,665	1,657	8	99.52	0.48	19,979	18,463	1,516	11.25	63.54	52.29	12.18	6,064,000	1,846,300	4,217,700
13	1,665	1,654	11	99.34	0.66	21,644	20,117	1,527	12.26	64.00	51.74	13.17	6,108,000	2,011,700	4,096,300
14	1,665	1,647	18	98.92	1.08	23,309	21,764	1,545	13.26	64.75	51.49	14.09	6,180,000	2,176,400	4,003,600
15	1,665	1,653	12	99.28	0.72	24,974	23,417	1,557	14.27	65.26	50.99	15.04	6,228,000	2,341,700	3,886,300
16	1,665	1,656	9	99.46	0.54	26,639	25,073	1,566	15.28	65.63	50.35	16.01	6,264,000	2,507,300	3,756,700
17	1,665	1,648	17	98.98	1.02	28,304	26,721	1,583	16.28	66.35	50.06	16.88	6,332,000	2,672,100	3,659,900
18	1,665	1,653	12	99.28	0.72	29,969	28,374	1,595	17.29	66.85	49.56	17.79	6,380,000	2,837,400	3,542,600
19	1,664	1,654	10	99.40	0.60	31,633	30,028	1,605	18.30	67.27	48.97	18.71	6,420,000	3,002,800	3,417,200
20	1,665	1,653	12	99.28	0.72	33,298	31,681	1,617	19.31	67.77	48.47	19.59	6,468,000	3,168,100	3,299,900

Table 8. Model Results for OOT Dataset (Top 20 Percent Score Cutoff)

Financial Curve and Recommended Cutoff

Based on our results table above, more specifically for the OOT dataset, we can obtain this plot of financial outcome curves under different cutoff values, as shown in Figure 5 below.

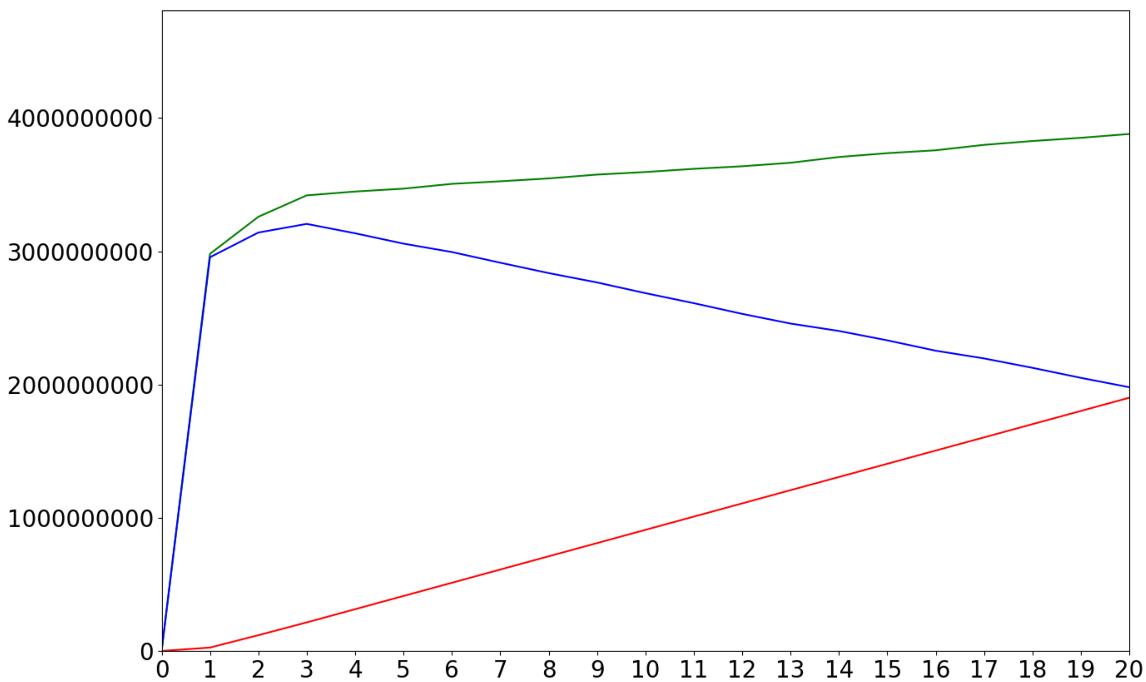


Figure 5. Financial Outcomes at Varying Cutoffs

In Figure 5, the green curve represents the monetary value of detected fraud. As the threshold for identifying fraud increases, this value rises quickly at first but eventually plateaus as fewer additional fraudulent applications are captured. The red curve shows lost revenue, which steadily increases as higher thresholds result in more non-fraudulent applications being mistakenly classified as fraud. The blue curve reflects the combined financial outcome, representing overall savings by balancing fraud detection with revenue loss.

Upon reviewing the blue curve and the results table from the previous section, it is evident that the green curve (fraud detection value) rises sharply initially, and then flattens around a threshold of 3. Meanwhile, the red curve (lost revenue) continues to increase steadily, and the blue curve (overall savings) peaks around the same threshold before starting to decline.

Therefore, we recommend a score cutoff of 3%. This threshold strikes a balance between maximizing fraud detection and minimizing lost revenue, as it corresponds to the peak of overall savings. Beyond this point, the financial benefit diminishes due to the growing cost of misclassifying legitimate applications as fraud.

Summary

In this project, we developed a machine learning-based fraud detection model specifically aimed at identifying fraudulent product applications. Using a LightGBM classifier, we aimed to optimize fraud detection while minimizing the impact on legitimate applications. We started with a dataset containing various fields, including dates, categorical values, and the target variable, "fraud_label." Initial data exploration revealed substantial class imbalance, with the vast majority of applications labeled as non-fraudulent. This imbalance informed our approach, particularly regarding model selection and threshold tuning.

Data cleaning and transformation were essential steps in preparing the dataset for modeling. We identified and treated placeholder values—such as "123 MAIN ST" in the address field and "999999999" in the ssn and homephone fields—by imputing these entries with unique identifiers, helping to reduce noise and improve data integrity. Dates of birth and application dates, stored as integers, were converted to datetime formats. Through feature engineering, we created a robust set of 651 variables aimed at capturing temporal and behavioral patterns indicative of fraud. These included "day-since" features, which track the recency of applications, and velocity features, which capture changes in application frequency over time. We also developed "unique counts" variables to identify links between entities and "maximum indicators" to record peak values in application counts. Redundant variables were removed, leaving a final, comprehensive set of attributes tailored for fraud detection.

Feature selection was performed using forward selection with LightGBM as the classifier, choosing the top 20 features based on their predictive power. These features primarily consisted of application counts and entity linkage metrics, ensuring that the model focused on the variables most likely to differentiate fraudulent from legitimate applications. Following feature selection, we tested multiple models, including Decision Tree, Random Forest, and Neural Network, but ultimately selected LightGBM due to its superior generalization and performance across training, testing, and out-of-time (OOT) datasets.

The model was tuned with optimal hyperparameters, including 31 leaves, no limit on depth, a learning rate of 0.01, and 200 trees, to achieve a Fraud Detection Rate (FDR) of 59.72% at a 3% cutoff on the OOT dataset. This cutoff was determined by analyzing financial outcome curves, which illustrated the trade-offs between fraud detection value, lost revenue, and overall savings. At this threshold, we estimate the model can save approximately $3,205,800,000 by efficiently identifying fraud while minimizing revenue loss from mistakenly flagged legitimate applications.

Future steps to enhance the model could include implementing a more sophisticated cost-sensitive approach to better balance fraud detection with revenue preservation or incorporating additional data sources to capture broader patterns of fraudulent behavior. Further testing with alternative algorithms like XGBoost or deep learning models could also be valuable in optimizing performance across different data distributions or further refining the threshold settings.

Appendix

DQR on the Applications Data

This document is the data quality report for the product application data. It contains two descriptive tables for both numerical fields and categorical fields, as well as distributions for each field visualized. The ‘applications’ dataset has 10,000 rows and 10 columns.

Numerical Fields Table

Field Name	Field Type	# Records Have Values	% Populated	# Zeros	Min	Max	Mean	Standard Deviation	Most Common
date	numeric	1,000,000.00	100.0%	0	20,170,101.00	20,171,231.00	20,170,667.78	344.99	20,170,816.00
dob	numeric	1,000,000.00	100.0%	0	19,000,101.00	20,161,031.00	19,517,248.66	356,887.02	19,070,626.00

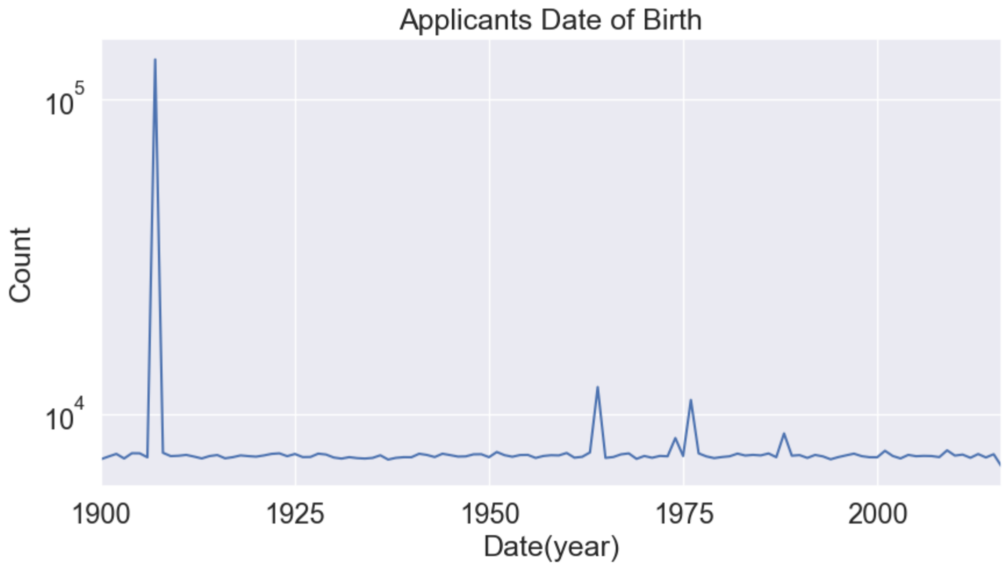
*Strictly speaking, these two fields contain dates instead of numerical values. Later in the data cleaning process, we will transform them into datetime variables.

Categorical Fields Table

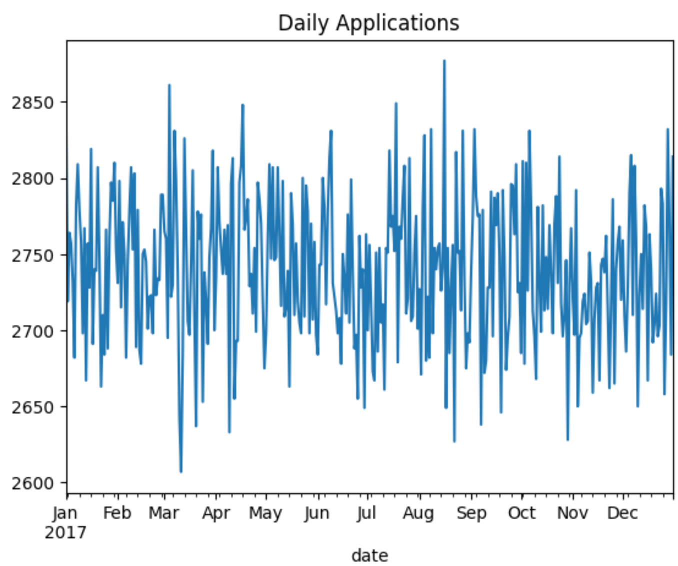
Field Name	Field Type	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
firstname	categorical	1,000,000.00	100.0%	0.00	78,136.00	EAMSTRMT
lastname	categorical	1,000,000.00	100.0%	0.00	177,001.00	ERJSAXA
address	categorical	1,000,000.00	100.0%	0.00	828,774.00	123 MAIN ST
record	categorical	1,000,000.00	100.0%	0.00	1,000,000.00	1
fraud_label	categorical	1,000,000.00	100.0%	985,607.00	2.00	0
ssn	categorical	1,000,000.00	100.0%	0.00	835,819.00	999999999
zip5	categorical	1,000,000.00	100.0%	0.00	26,370.00	68138
homephone	categorical	1,000,000.00	100.0%	0.00	28,244.00	9999999999

Distribution Plots

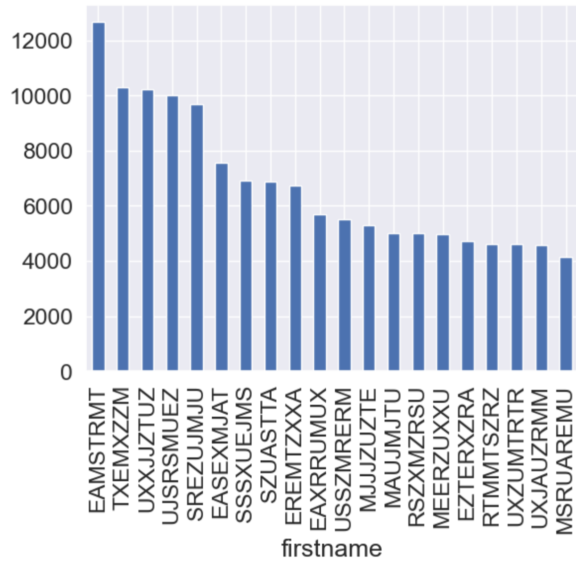
Distribution of the numerical field ‘dob’ (grouped by year)



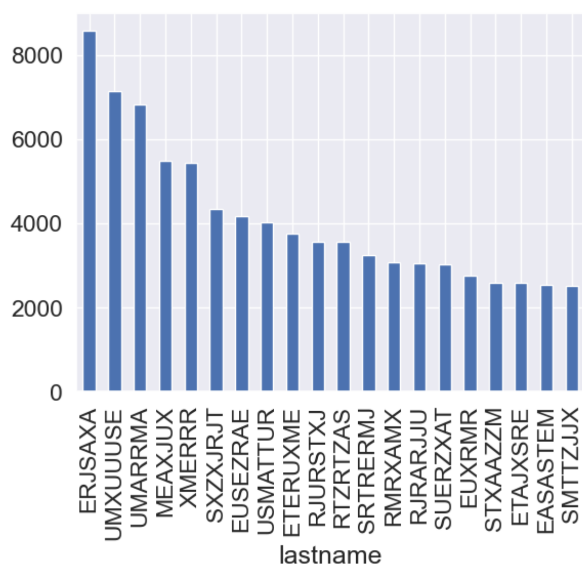
Distribution of the numerical field ‘date’



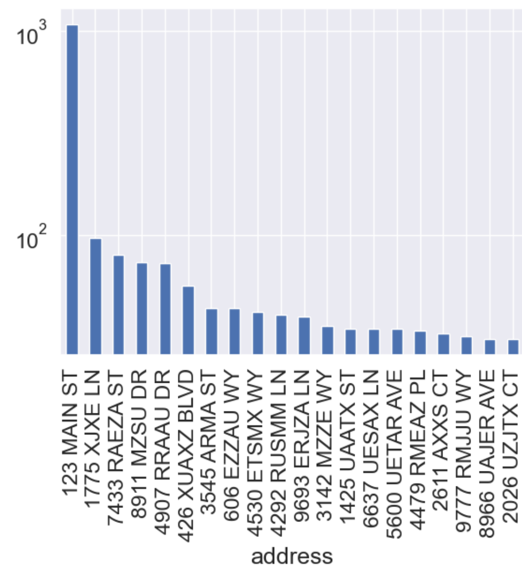
Distribution of the categorical field ‘firstname’ (top 20 most populated)



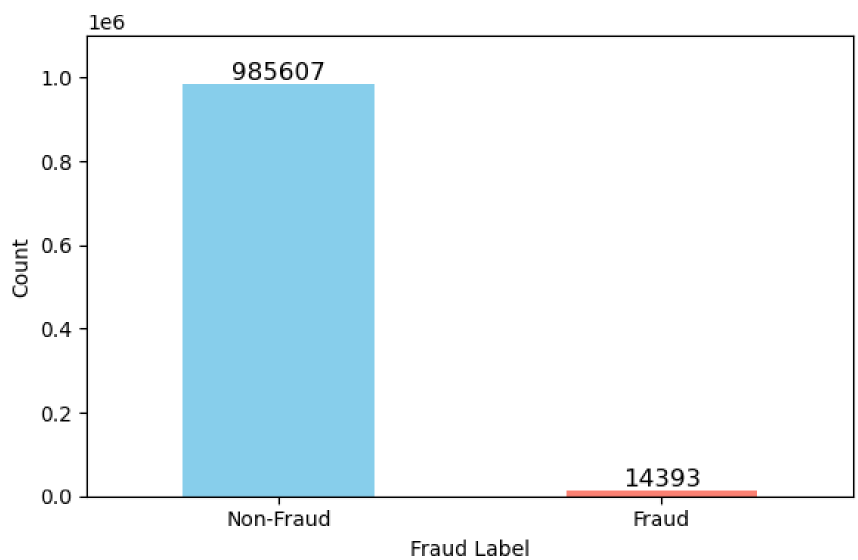
Distribution of the categorical field ‘lastname’ (top 20 most populated)



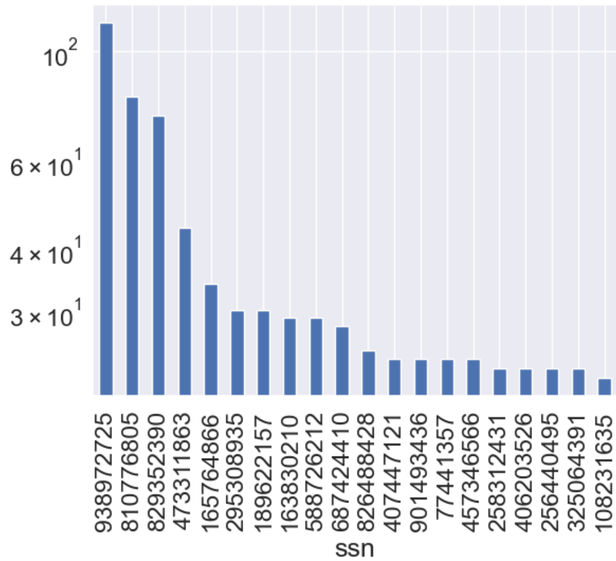
Distribution of the categorical field ‘address’ (top 20 most populated)



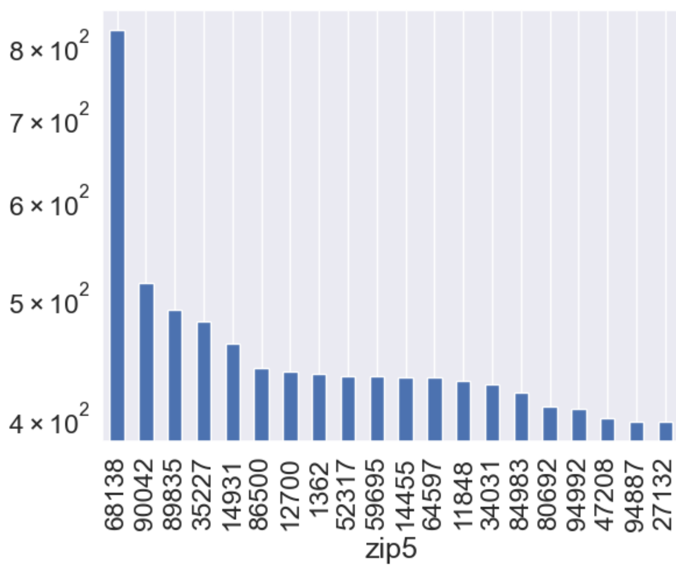
Distribution of the categorical field ‘fraud_label’



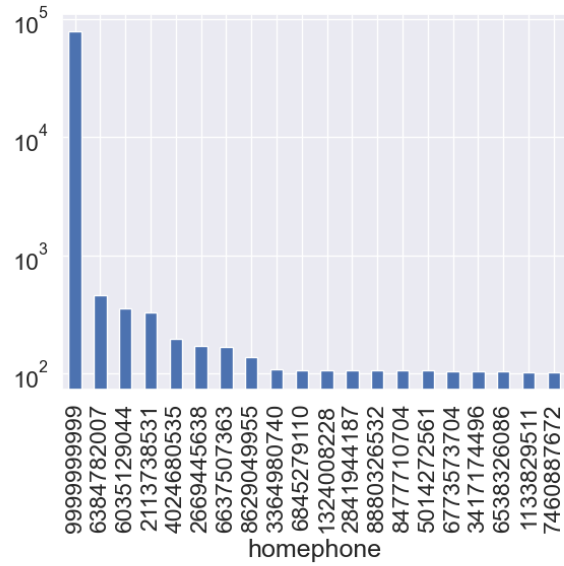
Distribution of the categorical field ‘ssn’ (top 20 most populated)



Distribution of the categorical field ‘zip5’ (top 20 most populated)



Distribution of the categorical field ‘homephone’ (top 20 most populated)



*The 'record' field distribution is not included as it serves as the index for rows and is uniformly distributed

Data Cleaning

Since the dataset we are working with is synthetic, there are no null values present. However, some fields require transformation to the correct data types for proper analysis. Additionally, there are fields with unusual values that need to be addressed.

Data Type Transformation

As noted, the fields 'date' and 'dob' are stored as integers in the dataset. We converted these fields to the correct datetime format to facilitate more accurate handling and analysis.

Unusual Values Treatment

During the data quality review, we identified several fields with unusually frequent placeholder-like values, including 'ssn', 'address', 'dob', and 'homephone'.

'ssn': The 'ssn' field contains 16,935 instances of the value '999999999'. We assume this value is used as a placeholder or default and have replaced it with a unique identifier from the 'record' field to ensure data integrity.

'address': The address '123 MAIN ST' appears 1,079 times, suggesting it may be a placeholder or erroneous entry. To prevent potential bias and improve the accuracy of downstream analysis, we imputed these entries with unique values from the 'record' field.

dob': The date '1907-06-26' occurs 126,568 times, which is highly unusual. This suggests that the value is likely a placeholder rather than an actual birth date. We addressed this by imputing these instances with unique values from the 'record' field.

'homephone': Similar to 'ssn', the value 9999999999 appears in 78,512 rows in the 'homephone' field. We treated this as a placeholder or default and imputed it with a unique value from the 'record' field.