Summary

Identity fraud, also known as identity theft, is a type of fraud in which a suspect deliberately uses another person's personal identifying information such as name, Social Security number, driver's license or other ID, credit card or bank account information without their permission or knowledge for financial gain or other illegal purposes. This type of fraud usually happens in the financial sector, mainly in the banking industry.

The overall goal of the project is to develop a supervised machine-learning model that can be used to detect and predict fraud in identification applications. It aims to develop an efficient and effective statistical analysis model that can be applied in practice to predict fraud and identify fraudulent identity applications.

This report aims to document the process of creating and completing a supervised machine-learning algorithm that can detect and prevent fraud in real-time.

- **Data cleaning:** examine the dataset and remove any inaccuracies
- **Feature Engineering:** generate new and insightful features from the existing data
- **Feature selection:** select the most relevant features from the dataset.
- **Modeling:** build and test multiple machine learning models, to determine the most efficient and effective model for detecting and predicting fraud in real-time.
- **Conclusion:** Analyze the results of the machine learning models and draw conclusions.

After testing different algorithms to fit the data, such as Logistic Regression, Random Forest, XGBoost, Light GBM, and Neural Network experimented with various hyperparameters combinations for each model & compared their performance based on FDR at a 3% rejection rate.

The LightGBM Classifier with parameters: max_depth=6, n_estimators=1000, num_leaves=10, and learning_rate=0.05, was the best and final model. The FDR scores for this model on training, testing, and OOT data were 53.00%, 52.80%, and 50.07%, respectively. (Taking number of variables =20)

Final model shows that we are not overfitting and getting good performance and we can catch 53.43 % of all the Fraud by rejecting the top 3% of the application.

1. **Data Description**

The data is synthetic application data containing Personal Identifying Information in fields such as Name, SSN, address, DOB, and Phone Number. The data contains **10 fields** and **1000000 records** indicating **the PII of individuals** from US applications such as applications for opening credit cards or cell phone accounts from the date **2017-01-01** to **2017-12-31**.

Summary Tables

A. Numerical Table

Field Name	% Populated	Min	Max	Mean	Stdev	# uniqueValues	% Zero
dob	100.00	1900-01-01	2016-10-31	/	/	42,673	0.00
date	100.00	2017-01-01	2017-12-31	/	/	365	0.00

B. Categorical Table

Field Name	% Populated	# Unique Values	Most Common field Value
record	100.00	100,0000	NA
ssn	100.00	83,5819	999,999,999
firstname	100.00	78136	EAMSTRMT
lastname	100.00	177001	ERJSAXA
address	100.00	82,8774	123 MAIN ST
zip5	100.00	26370	68138
homephone	100.00	28244	999,999,9999
Fraud_label	100.00	2	0

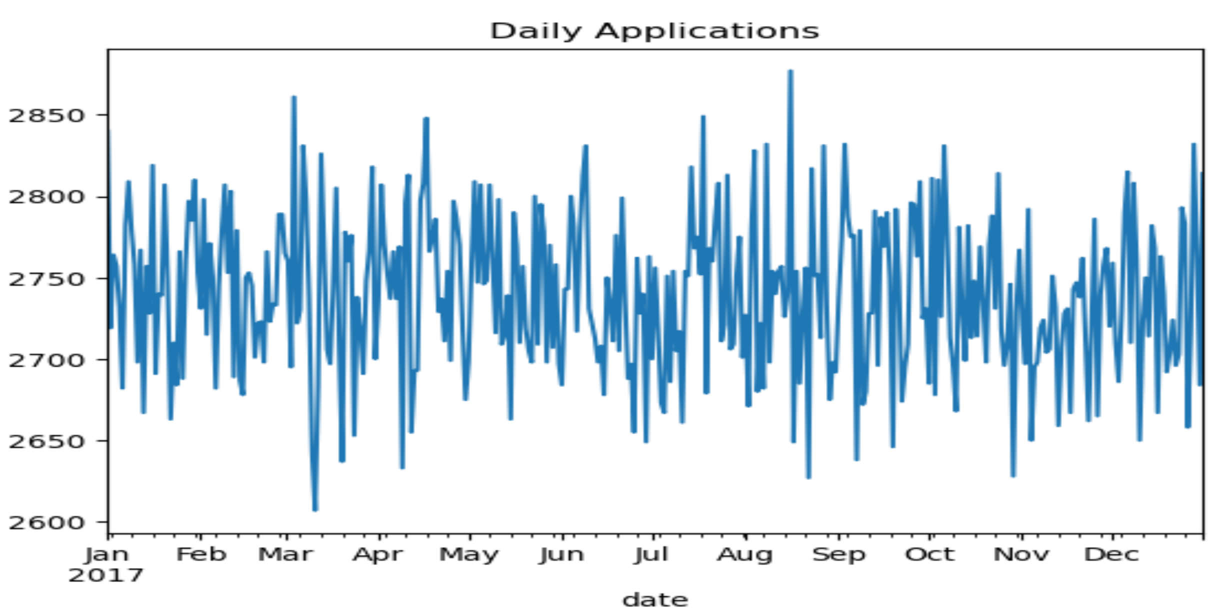
2. Visualization of Each Field

1. **Field Name:** record
   Description: A categorical field containing unique positive integers for each record from 1 to 10,000,00.

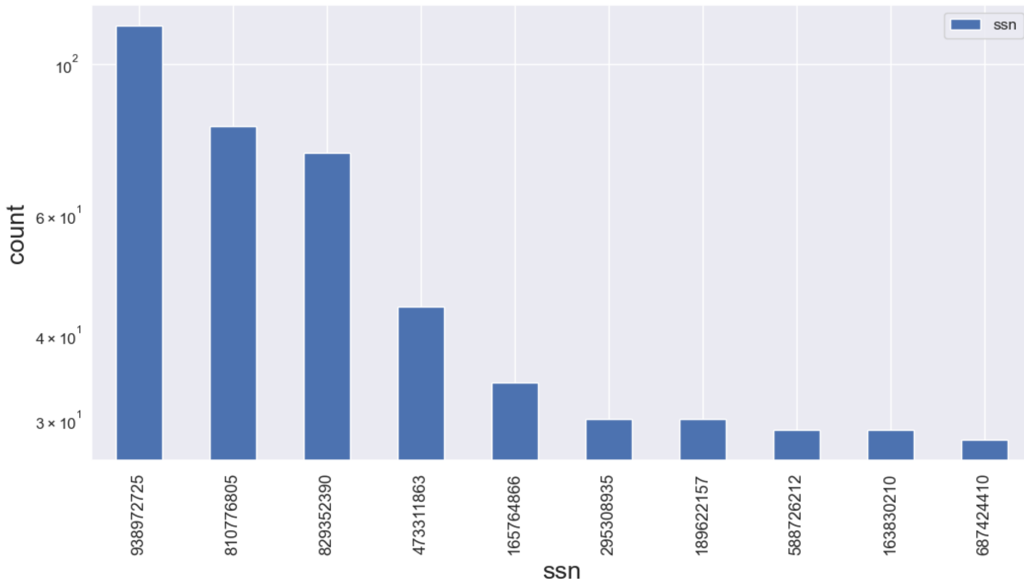
2. **Field Name:** date
   Description: Containing date of application when it was filled. The most common date is 2017-08-16, the count for which is 2877 .

Date of Application	Number of Applications
2017-08-16	2877
2017-03-04	2861
2017-07-18	2849
2017-04-17	2848
2017-01-01	2840
2017-09-03	2832
2017-08-08	2832
2017-12-28	2832
2017-08-27	2831
2017-10-06	2831

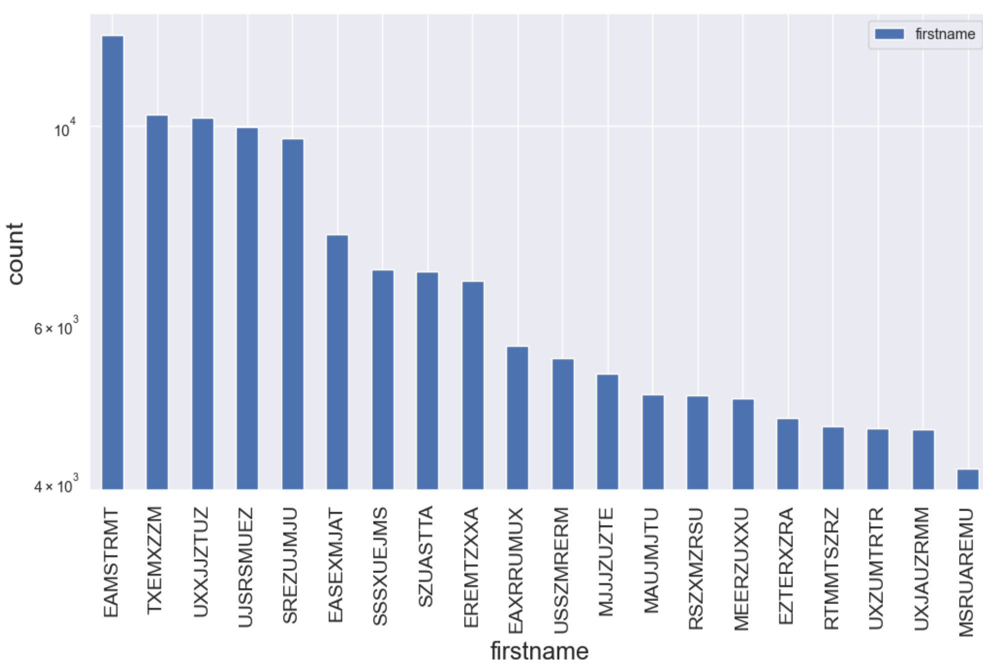
Table containing top 10 dates on which most applications were filled.



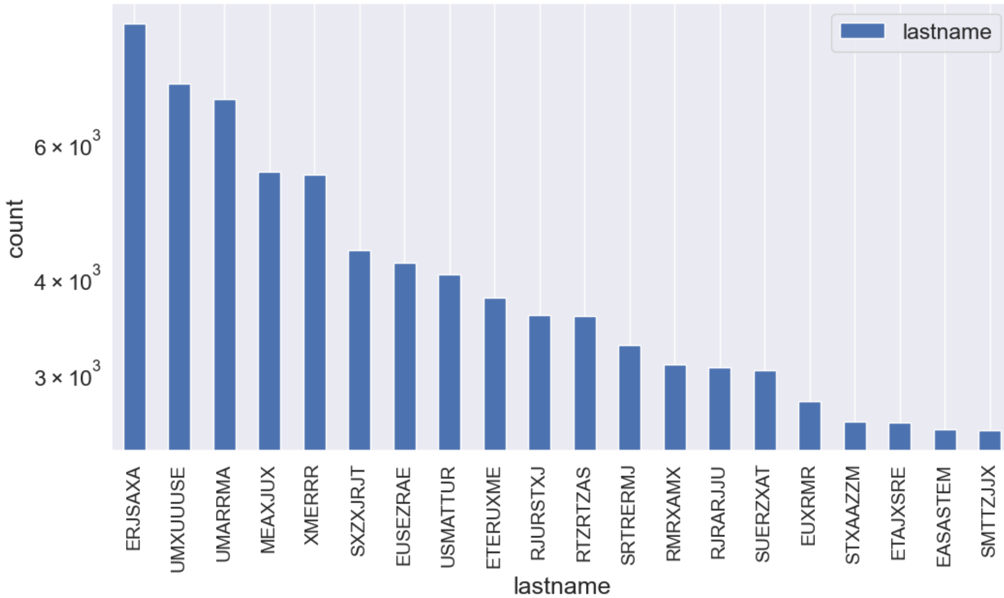
3. **Field Name:** SSN
Description: This field contains Social Security Numbers of applicants. The frequency distribution below shows top 10 field values of the most frequent SSN. The most common value of SSN is 999999999 counts for which is 16,935.



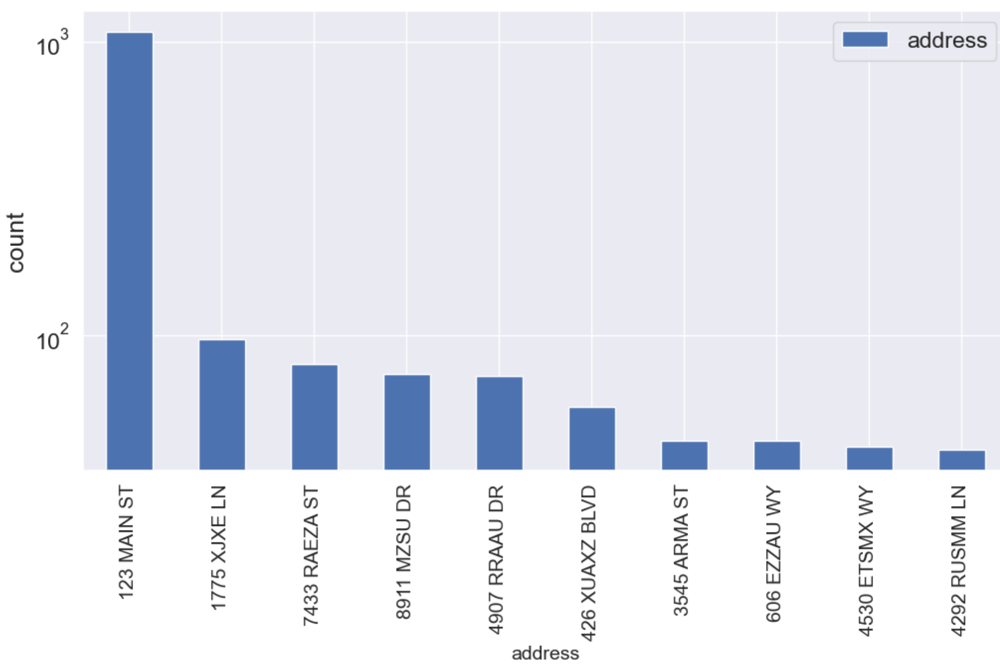
4. **Field Name:** firstname
Description: This field Contains the first name of the applicant and the distribution below shows the top 20 most common first names used. The most common first name used is EAMSTRMT, count for which is 12,658.



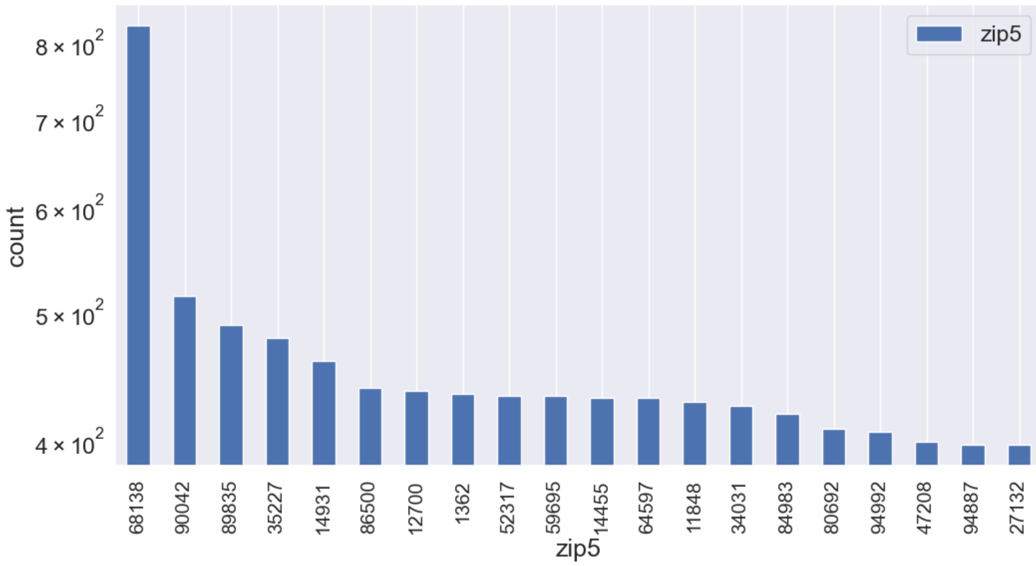
5. **Field Name:** lastname
   Description: This field contains last names of applicants. The distribution below shows the top 20 field values of last name, The most common last name used is ERJSAXA, count for which is 8,580.



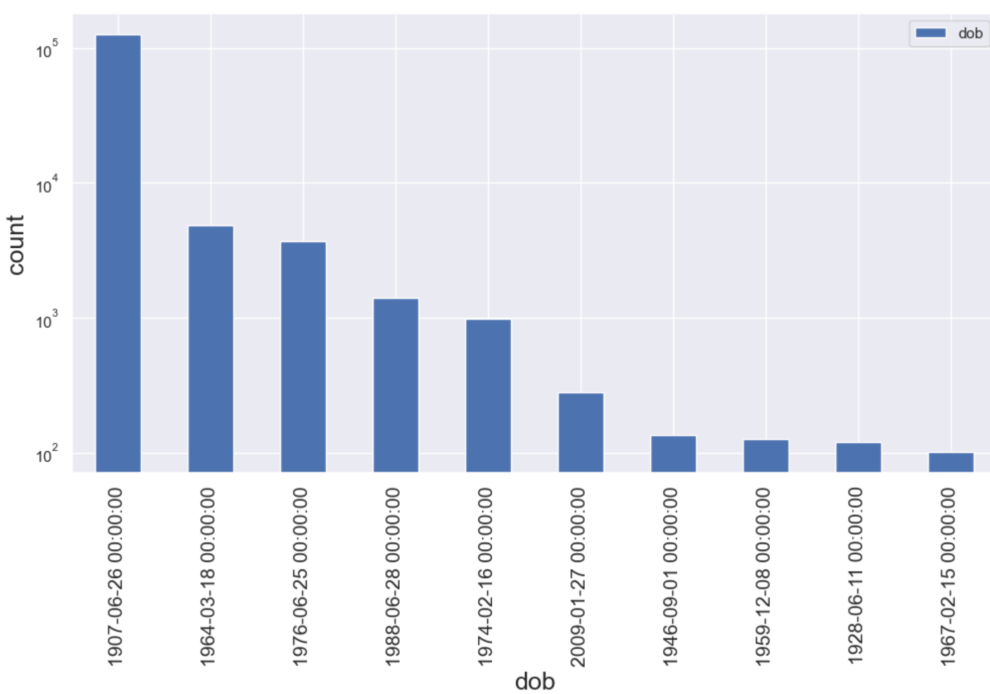
6. **Field Name:** address
   Description: This Field contains the address details of each applicant. The distribution below shows the top 10 most frequent addresses used in the application. Most common address used is 123 MAIN ST, the count for which is 1,097.



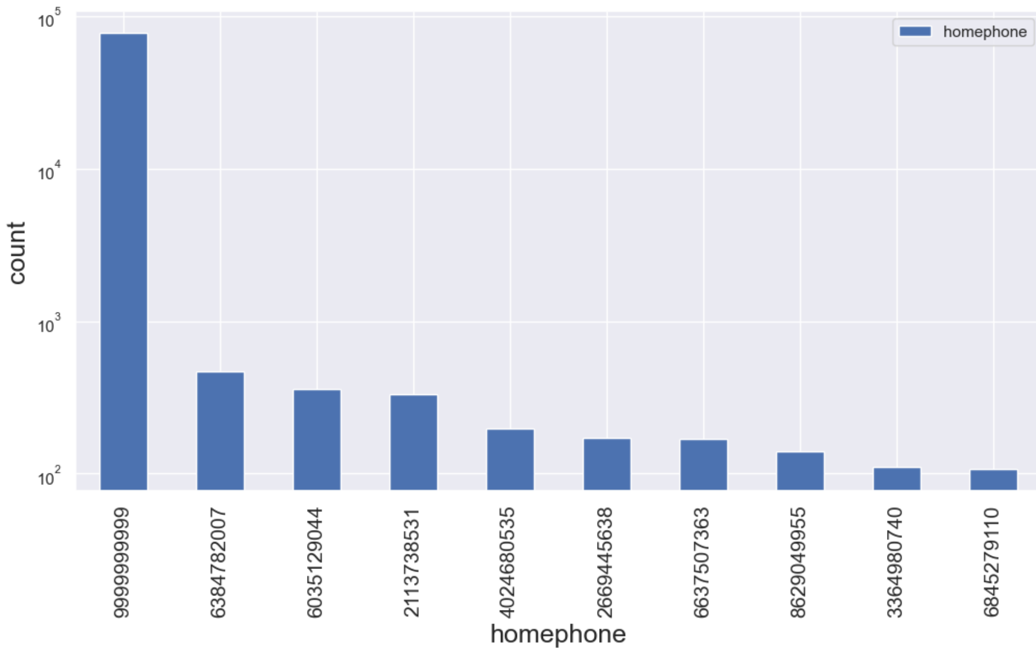
7. **Field Name:** zip5
   Description: Zip code used in applications. The distribution below shows top 20 field values of zip code. Most common zip code is 68138, count for which is 823.



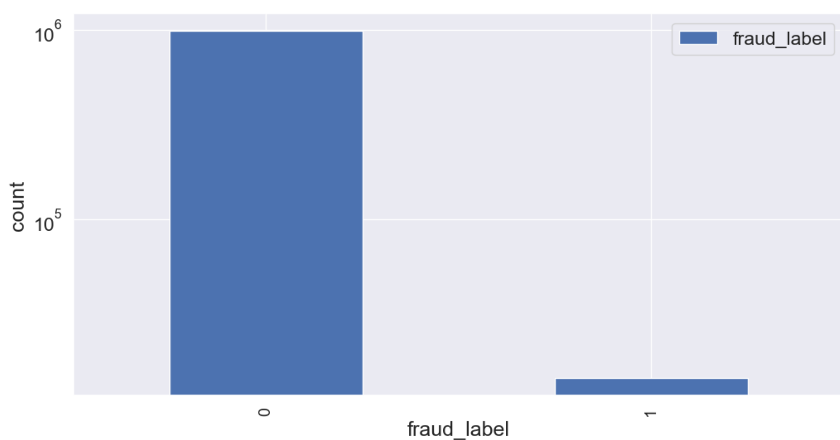
8. **Field Name:** dob
   Description: This field contains date of birth of each applicant. The distribution below shows the top 10 date of birth used in the application. The most common date of birth is 1907-06-26, count for which is 26,568.



9. **Field Name:** homephone
Description: This field contains the home phone number of applicants. The distribution below shows top 10 values of the phone number used. The most common home phone number used is 9999999999, count for which is 78,512.



10. **Field Name:** fraud_label:
Description: This field assigned fraud_label = 0 for applications that are not fraud and 1 to those which are potential fraud. The count for fraud_label=0 is 98,5607 and count for fraud_label =1 is 14,393



3. Data Cleaning

The process of cleaning the data involves addressing the issue of identity fraud detection by identifying applications that use identical or similar identity information. However, upon analyzing the data, I came across four fields containing insignificant values that serve as placeholders for fields that were not collected. As a result, these values could create false links between unrelated applications and adversely affect the efficacy of the model. To rectify this, we need to substitute these frivolous values with non-linking values, for which we have used negative record numbers. Table below illustrates the replacement of these values.

Field	Frivolous Value	Replacement
address	123 MAIN ST	Replaced with Record number + “RECORD”
SSN	999999999	Negative record number with right alignment 0
DOB	19070626	Negative record number
homephone	9999999999	With negative record number

3.Feature Engineering

It is the process of selecting, transforming, and creating variables from raw data to improve the performance of machine learning models. It involves steps like data cleaning, transformation, and selection of relevant features**.**

Initially we created 4052 variables then after deduping a set of 2242 potential variable were generated by utilizing the existing PII field and combinations of PII fields, aimed at depicting the underlying structures within the data, leading to more accurate measurement of fraudulent activities. The categories and corresponding number of variables generated for each are outlined in the table below

Note: We are not taking <u>max indicator</u> variables because they are looking into future.

Description of Variables	Number of Variables Created
Original fields from dataset excluding 'record' and 'fruad_label'	**8(existing)**
Amount Variables: This set of variables include min,max,mean, total count acorss other field ,ssn,address,zip5,dob,name, homephone etc over past [0,1,3,7,14,30] days	**140**
Velocity: # how many times records with same entity have been seen over past [0,1,3,7,14,30] days	**138**
Relative Velocity Variables: These set of variables include ratio of applications with attributes see in recent past(0,1) to applications with same attribute seen in past[3,4,30] days (long term averaged velocity)	**184**
Number of unique values of an entity for a specific value of another entity over a window of {0,1,3,7,14,30} days	**1753**
Days since an application with that entity has been seen over past {0,1,3,7,14,30} days	**23**
Date of week (dow_risk) average fruad percentage of day of week (likelyhood of fraud any day of week	**1**
New entities combining/concatening different original fields	**9**
fraud_label (taget variable)	1
age_when_apply: age of applicant at the time of application	**1**
Total Variable	**2242**

Modes of Identity Fraud:

- **Identity theft –** fraudster uses a real but stolen identity (different from their own). Look for the SSN, DOB, Name to be associated with multiple contact points (address, phone, email) ,velocity around all PII elements.
- **Identity manipulation –** fraudster slightly changes their own identity. Small changes to SSN, DOB, name. Look for many slight, systematic variations in PII elements.
- **Synthetic identity –** fraudster makes up a completely fabricated identity. Look for all the PII elements to be associated with multiple different identities.

Some of the examples of variables built to catch identity fraud are:

1. **Velocity variables:** Velocity can be used as a measure of how often a particular PII or combination of PII appears within a specific time frame. This is represented by the number of records that contain the same PII or combination of PII that have been seen within a given window of time. A higher value for velocity indicates that the PII has been used more frequently, which may increase the risk of fraud. The time frames used for velocity calculations include 0, 1, 3, 7, 14, and 30 days, with 0 representing the same day as the last application.

2. **Day_of_ week _Risk** : To consider the possibility that the probability of a person committing fraud may differ on different days of the week, a categorical variable was generated to represent the weekday. Target Encoding was used to encode this categorical variable.

3. **Relative Velocity :** The reasoning behind relative velocity is that if the same number of occurrences of the same PII happens in a set time period, a situation where they occur on the same day is more likely to be fraudulent than if they were spread out over different dates.

   → Relative Velocity = $(\#\ apps\ with\ that\ group\ seen\ in\ the\ past\ 1\ day) / (\#\ apps\ with\ that\ same\ group\ seen\ in\ the\ past\ [3,7,14,30]\ days)$

4. **Days Since an Application:** The frequency with which a particular PII or combination of PIIs appears is a crucial indicator of fraudulent activity. One such indicator is the "day since last seen" variable, which measures the number of days since the previous occurrence of the same PII or PII combination. A smaller value for this variable indicates a higher likelihood of fraud

4. Feature Selection:

Feature selection is a process of selecting a subset of relevant features or variables from a larger set of features that are available in a dataset. It is done to improve the performance of a machine-learning model by reducing the complexity of the data and preventing overfitting. The main objective of feature selection is to remove irrelevant, redundant, or noisy features that do not contribute significantly to the predictive accuracy of the model while retaining the most important features that capture the underlying patterns in the data.

Feature selection can be motivated by several factors, including:

1. Overfitting:
2. Computational Efficiency
3. Model Interpretability

The process of feature selection typically involves the following steps:

Feature Ranking: This involves ranking the features in order of their importance or relevance to the problem being solved. We used Filters(**depending upon KS Score** and Wrappers(**by multivariate importance)** to rank the top 25 features.

$$KS = \max_{x} \int_{x_{min}}^{x} \left[P_{\text{goods}} - P_{\text{bads}}\right] dx$$

In our model we have used FDR@ 3% as filter metric

Fraud detection Rate (FDR) @ 3% : The FDR (Fraud Detection Rate) indicates the percentage of all detected frauds at a specific threshold level. For instance, if the FDR is 55% at a 3% threshold, it means that the model can detect 55% of all frauds within 3% of the total population. To compute FDR, the number of true fraud cases identified by the model is divided by the total number of actual fraud cases in the dataset. Essentially, FDR represents the model's ability to capture fraud cases using a fixed number of predicted positives. FDR@3% is calculated by dividing the number of frauds detected at a 3% rejection rate by the total number of frauds present in the dataset.

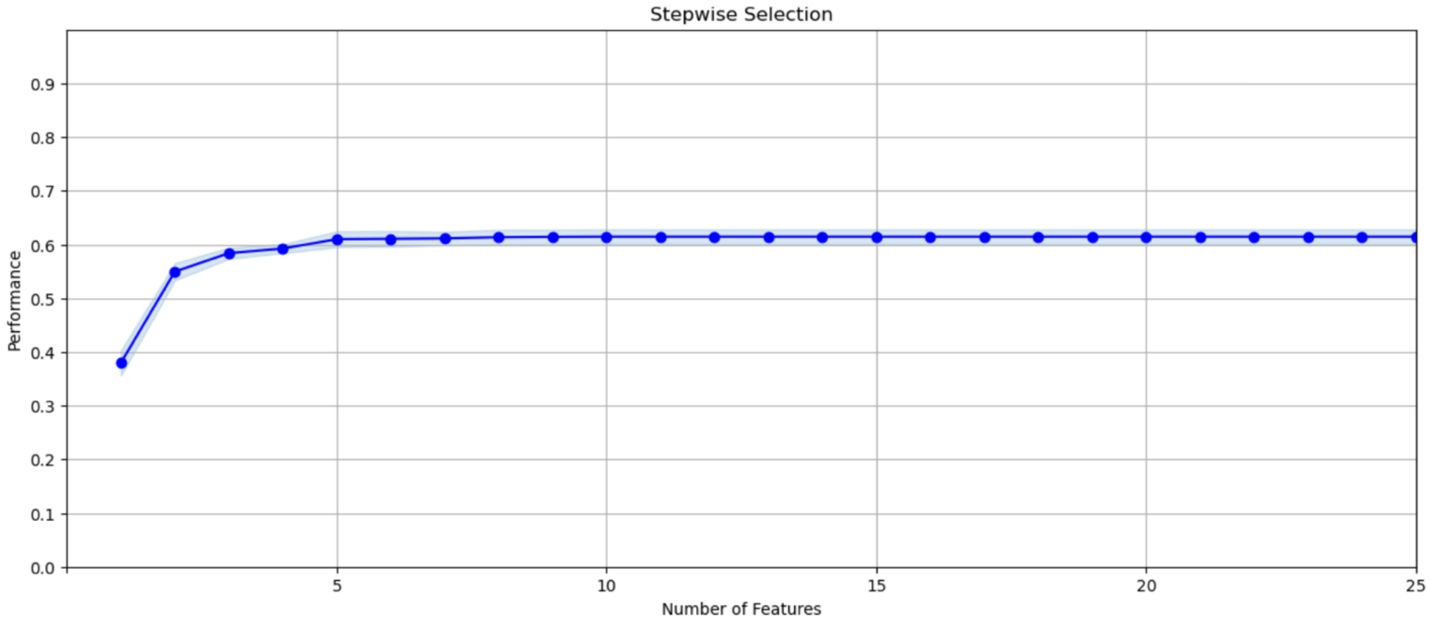
Feature Subset Selection: This involves selecting a subset of the top-ranked features based on some criterion, such as a fixed number of features, or a threshold value for feature importance.

Evaluation: The selected subset of features is evaluated by training a machine learning model on the data using only the selected features, and then evaluating its performance on a held-out test set. The performance of the model is compared to that of a model trained on the full set of features to determine whether feature selection has improved the accuracy and performance of the model.

Note: We are not considering max_indicator variables are they are causing target leak

25 most important list of variables sorted by multivariate importance and with corresponding filter score and avg_cv score

wrapper order	variable name	avg_cv_score	filter score	Description of variables
1	max_count_by_address_30	0.379646557	0.3592154645	number of application that use particulat address in the past 30 days
2	max_count_by_ssn_dob_7	0.5494907286	0.2284008373	number of application that use particular ssn & dob combination in the past 7 days
3	max_count_by_homephone_3	0.5838774267	0.2247574356	number of application that use particular homephone in the past 3 days
4	max_count_by_fulladdress_30	0.5924958649	0.3599139689	number of application that use particular full address in the past 30 days
5	zip5_count_3	0.6099068512	0.2247055806	# appearance with the given zip code seen in the past 3 days
6	max_count_by_ssn_dob_30	0.6106032907	0.2408356905	number of application that use particular ssn & dob combination seen in the past 30 days
7	max_count_by_homephone_7	0.6113867851	0.2322352909	number of application that use particular homephone seen in the past 7 days
8	fulladdress_count_0_by_30	0.6136502133	0.2907221306	# appearance with the given full address seen today divided by number of times this fulladdress was seen in past 30 days
9	max_count_by_fulladdress_homephone_	0.6142595978	0.2497237493	# appearance with the given full address & homephone combination seen in the past 30 days
10	ssn_dob_day_since	0.6146078175	0.2286263261	# appearance with the given ssn & dob combination seen since past number of days
11	max_count_by_address_7	0.6146078175	0.3433354317	# appearance with the given address seen in the past 7 days
12	address_day_since	0.6146078175	0.3341399436	# of appearance with the given address seen in the past days
13	fulladdress_day_since	0.6146078175	0.3332685356	particular full address seen since past n days
14	max_count_by_fulladdress_3	0.6146078175	0.329537708	number of application that use particular full address seen over past 3 day
15	max_count_by_address_3	0.6146078175	0.3294447055	number of application that use particular address over past 3 days
16	address_count_14	0.6146078175	0.3224362795	# appearance with the given address seen in the past 14 days
17	fulladdress_count_14	0.6146078175	0.3219529251	# appearance with the given fulladdress seen in the past 14 days
18	max_count_by_address_1	0.6146078175	0.3153319064	number of application that use particular address seen in the 1 days
19	max_count_by_fulladdress_1	0.6146078175	0.3152526097	number of application that use particular address seen in the 1 days
20	address_count_7	0.6146078175	0.3017352771	# appearance with the given address seen in the 7 days
21	fulladdress_count_7	0.6146078175	0.3016662468	# appearance with the given fulladdress seen in the past 7days
22	address_unique_count_for_name_homep	0.6146078175	0.2924379574	# appearance ofunique combination of address+name+homphone seen in past 60 days
23	address_count_0_by_30	0.6146078175	0.291922189	# appearance with the given address seen today divided by number of times this address was seen in past 30 days
24	address_unique_count_for_homephone_	0.6146078175	0.2914097875	# appearance of unique combination of address+homphone+ name +dob seen in past 60 days
25	fulladdress_unique_count_for_ssn_home	0.6146078175	0.2899906218	# appearance ofunique combination of fulladdress+ ssn +homphone seen in past 60 days



5 Modeling:

Tested various algorithms with different parameter combinations, such as Logistic Regression, Random Forest, XGBoost, LightGBM, and Neural Network. To prevent overfitting, we employed 5-fold cross-validation during the selection process , and calculated the average FDR at a 3% rejection rate to reduce the effect of randomness. The average FDR score was then used to compare the performance of different models. After comparing the results, LightGBM exhibited the highest FDR on the testing data and was selected as the final model.

The LightGBM Classifier with parameters: max_depth=6, n_estimators=1000, num_leaves=10, and learning_rate=0.05, was the best and final model. The FDR scores for this model on training, testing, and OOT data were 53.00%, 52.80%, and 50.07%, respectively. (Taking number of variables =20)

The model performance table provides additional details on the various models.

Model Scores										
Model		Parameters					Average FDR at 3%			
Logistic Regression	# of Variables	solver	penalty	C			Train	Test	OOT	
	5	lbfgs	l2	1			0.479	0.472	0.463	
	10	liblinear	l2	10			0.487	0.496	0.475	
	15	lbfgs	l2	1			0.489	0.485	0.475	
	20	liblinear	l2	1			0.489	0.49	0.474	
	20	lbfgs	l2	0.1			0.478	0.482	0.465	
Decision Tree	# of Variables	max_depth	min_sample_leaf	min_samples_split			Train	Test	OOT	
	10	5	30	50			0.512	0.512	0.489	
	10	10	20	30			0.526	0.529	0.504	
	20	20	15	25			0.540	0.515	0.500	OverFitting
	20	10	20	30			0.528	0.524	0.504	
Random Forest	# of Variables	n_estimators	max_depth	in_samples_sp	min_samples_leaf	max_features	Train	Test	OOT	
	10	20	4	50	30	5	0.522	0.532	0.498	Underfitting
	10	40	10	25	20	7	0.527	0.53	0.504	
	20	60	15	25	15	5	0.534	0.527	0.504	
	20	100	25	20	10	8	0.541	0.523	0.502	Overfitting
LGBM	# variables	n_estimators	max_depth	learning_rate	num_leaves		Train	Test	OOT	
	10	5	2	0.1	2		0.512	0.512	0.489	
	15	600	6	0.06	9		0.533	0.518	0.506	Overfitting
	20	700	5	0.05	6		0.527	0.528	0.506	
	20	1000	6	0.05	10		0.530	0.528	0.507	Final model score
GBC	# variables	n_estimators	max_depth	Learning_rate			Train	Test	OOT	
	10	100	2	0.1			0.524	0.518	0.500	
	10	200	4	0.05			0.529	0.528	0.506	
	200	600	5	0.05			0.537	0.521	0.506	
	20	1000	5	0.05			0.539	0.525	0.504	Overfitting
XGB	# variables	n_estimators	max_depth	Learning_rate			Train	Test	OOT	
	10	50	2	0.1			0.511	0.512	0.489	
	10	600	5	0.01			0.508	0.509	0.486	
	20	700	5	0.3			0.541	0.522	0.500	
	20	1000	5	0.3			0.545	0.515	0.501	Overfitting
Neural Network	# Variables	activation	solver	learning_rate	learning_rate_init	alpha	Train	Test	OOT	
	10	relu	adam	constant	0.01	0.1	0.511	0.512	0.491	
	10	logistic	adam	constant	0.001	0.01	0.514	0.515	0.492	
	20	relu	lbfgs	adaptive	0.001	0.01	0.526	0.527	0.506	
	20	relu	lbfgs	adaptive	0.0001	0.0001	0.526	0.528	0.506	

6 . Conclusion

The overall goal of the project is to develop a supervised machine-learning model that can be used to detect and predict fraud in identification applications. It aims to develop an efficient and effective statistical analysis model that can be applied in practice to predict fraud and identify fraudulent identity applications.

After testing different algorithms to fit the data, such as Logistic Regression, Random Forest, XGBoost, Light GBM, and Neural Network experimented with various hyperparameters combinations for each model & compared their performance based on FDR at a 3% rejection rate. After training, we sorted each dataset's observations by their predicted probability of being fraud and split them into 100 bins. We then calculated the key statistics of the top 20 bins for each dataset: training, testing, and OOT.

The LightGBM Classifier with parameters: max_depth=6, n_estimators=1000, num_leaves=10, and learning_rate=0.05, was the best and final model. The FDR scores for this model on training, testing, and OOT data were 53.04%, 52.99%, and 50.80%, respectively. (Taking number of variables =20)

Final model shows that we are not overfitting and getting good performance and that we can catch 53.43 % of all the Fraud by rejecting the top 3% of the application.

Details of Top 20 Bins for Training Data

Training	#records	# Goods	# Bads		Fraud Rate							
	583454	575090	8364		0.0143353							
		Bin Statisctics						Cummulative Statistics				
Population Bin %	#records	# Goods	# Bads	% Goods	% Bads	total # Records	Cumulative goods	Cumulative Bads	% Cumulative goods	% Bads (FDR)	KS	FPR
1	5835	1566	4269	26.84	73.16	5835	1566	4269	0.27	51.04	50.77	0.37
2	5834	5694	140	97.60	2.40	11669	7260	4409	1.26	52.71	51.45	1.65
3	5835	5775	60	98.97	1.03	17504	13035	4469	2.27	53.43	51.16	2.92
4	5834	5793	41	99.30	0.70	23338	18828	4510	3.27	53.92	50.65	4.17
5	5835	5792	43	99.26	0.74	29173	24620	4553	4.28	54.44	50.15	5.41
6	5834	5789	45	99.23	0.77	35007	30409	4598	5.29	54.97	49.69	6.61
7	5835	5786	49	99.16	0.84	40842	36195	4647	6.29	55.56	49.27	7.79
8	5834	5802	32	99.45	0.55	46676	41997	4679	7.30	55.94	48.64	8.98
9	5835	5789	46	99.21	0.79	52511	47786	4725	8.31	56.49	48.18	10.11
10	5834	5793	41	99.30	0.70	58345	53579	4766	9.32	56.98	47.67	11.24
11	5835	5794	41	99.30	0.70	64180	59373	4807	10.32	57.47	47.15	12.35
12	5834	5780	54	99.07	0.93	70014	65153	4861	11.33	58.12	46.79	13.40
13	5835	5798	37	99.37	0.63	75849	70951	4898	12.34	58.56	46.22	14.49
14	5835	5791	44	99.25	0.75	81684	76742	4942	13.34	59.09	45.74	15.53
15	5834	5792	42	99.28	0.72	87518	82534	4984	14.35	59.59	45.24	16.56
16	5835	5785	50	99.14	0.86	93353	88319	5034	15.36	60.19	44.83	17.54
17	5834	5795	39	99.33	0.67	99187	94114	5073	16.37	60.65	44.29	18.55
18	5835	5798	37	99.37	0.63	105022	99912	5110	17.37	61.10	43.72	19.55
19	5834	5791	43	99.26	0.74	110856	105703	5153	18.38	61.61	43.23	20.51
20	5835	5797	38	99.35	0.65	116691	111500	5191	19.39	62.06	42.68	21.48

Details of Top 20 Bins for Testing Data

	#records	# Goods	# Bads		Fraud Rate							
	250053	246410	3643		0.0145689							
Testing		Bin Statisctics						Cummulative Statistics				
Population Bin %	#records	# Goods	# Bads	% Goods	% Bads	total # Records	Cumulativ e goods	Cumulative Bads	% Cumulative goods	% Bads (FDR)	KS	FPR
1	2501	666	1835	26.63	73.37	2501	666	1835	0.27	50.37	50.10	0.36
2	2500	2460	40	98.40	1.60	5001	3126	1875	1.27	51.47	50.20	1.67
3	2501	2482	19	99.24	0.76	7502	5608	1894	2.28	51.99	49.71	2.96
4	2500	2478	22	99.12	0.88	10002	8086	1916	3.28	52.59	49.31	4.22
5	2501	2486	15	99.40	0.60	12503	10572	1931	4.29	53.01	48.72	5.47
6	2500	2479	21	99.16	0.84	15003	13051	1952	5.30	53.58	48.29	6.69
7	2501	2477	24	99.04	0.96	17504	15528	1976	6.30	54.24	47.94	7.86
8	2500	2474	26	98.96	1.04	20004	18002	2002	7.31	54.95	47.65	8.99
9	2501	2475	26	98.96	1.04	22505	20477	2028	8.31	55.67	47.36	10.10
10	2500	2492	8	99.68	0.32	25005	22969	2036	9.32	55.89	46.57	11.28
11	2501	2479	22	99.12	0.88	27506	25448	2058	10.33	56.49	46.16	12.37
12	2500	2478	22	99.12	0.88	30006	27926	2080	11.33	57.10	45.76	13.43
13	2501	2487	14	99.44	0.56	32507	30413	2094	12.34	57.48	45.14	14.52
14	2500	2487	13	99.48	0.52	35007	32900	2107	13.35	57.84	44.49	15.61
15	2501	2482	19	99.24	0.76	37508	35382	2126	14.36	58.36	44.00	16.64
16	2500	2478	22	99.12	0.88	40008	37860	2148	15.36	58.96	43.60	17.63
17	2501	2479	22	99.12	0.88	42509	40339	2170	16.37	59.57	43.20	18.59
18	2501	2487	14	99.44	0.56	45010	42826	2184	17.38	59.95	42.57	19.61
19	2500	2481	19	99.24	0.76	47510	45307	2203	18.39	60.47	42.09	20.57
20	2501	2491	10	99.60	0.40	50011	47798	2213	19.40	60.75	41.35	21.60

Details of Top 20 Bins for OOT Data

	#records	# Goods	# Bads		Fraud Rate							
	166493	164107	2386		0.0143309							
OOT		Bin Statisctics						Cummulative Statistics				
Population Bin %	#records	# Goods	# Bads	% Goods	% Bads	total # Records	Cumulativ e goods	Cumulative Bads	% Cumulative goods	% Bads (FDR)	KS	FPR
1	1665	502	1163	30.15	69.85	1665	502	1163	0.31	48.74	48.44	0.43
2	1665	1638	27	98.38	1.62	3330	2140	1190	1.30	49.87	48.57	1.80
3	1665	1641	24	98.56	1.44	4995	3781	1214	2.30	50.88	48.58	3.11
4	1665	1658	7	99.58	0.42	6660	5439	1221	3.31	51.17	47.86	4.45
5	1665	1656	9	99.46	0.54	8325	7095	1230	4.32	51.55	47.23	5.77
6	1665	1656	9	99.46	0.54	9990	8751	1239	5.33	51.93	46.60	7.06
7	1665	1654	11	99.34	0.66	11655	10405	1250	6.34	52.39	46.05	8.32
8	1664	1650	14	99.16	0.84	13319	12055	1264	7.35	52.98	45.63	9.54
9	1665	1649	16	99.04	0.96	14984	13704	1280	8.35	53.65	45.30	10.71
10	1665	1657	8	99.52	0.48	16649	15361	1288	9.36	53.98	44.62	11.93
11	1665	1655	10	99.40	0.60	18314	17016	1298	10.37	54.40	44.03	13.11
12	1665	1652	13	99.22	0.78	19979	18668	1311	11.38	54.95	43.57	14.24
13	1665	1650	15	99.10	0.90	21644	20318	1326	12.38	55.57	43.19	15.32
14	1665	1648	17	98.98	1.02	23309	21966	1343	13.39	56.29	42.90	16.36
15	1665	1655	10	99.40	0.60	24974	23621	1353	14.39	56.71	42.31	17.46
16	1665	1650	15	99.10	0.90	26639	25271	1368	15.40	57.33	41.94	18.47
17	1665	1655	10	99.40	0.60	28304	26926	1378	16.41	57.75	41.35	19.54
18	1665	1654	11	99.34	0.66	29969	28580	1389	17.42	58.21	40.80	20.58
19	1665	1655	10	99.40	0.60	31634	30235	1399	18.42	58.63	40.21	21.61
20	1665	1660	5	99.70	0.30	33299	31895	1404	19.44	58.84	39.41	22.72