Credit Card Fraud Detection using Machine Learning

Aahad
Abubaker

03/18/24

Introduction to Dataset

Impact of Fraud:

- U.S. losses from credit card fraud will total $165.1 billion over the next 10 years

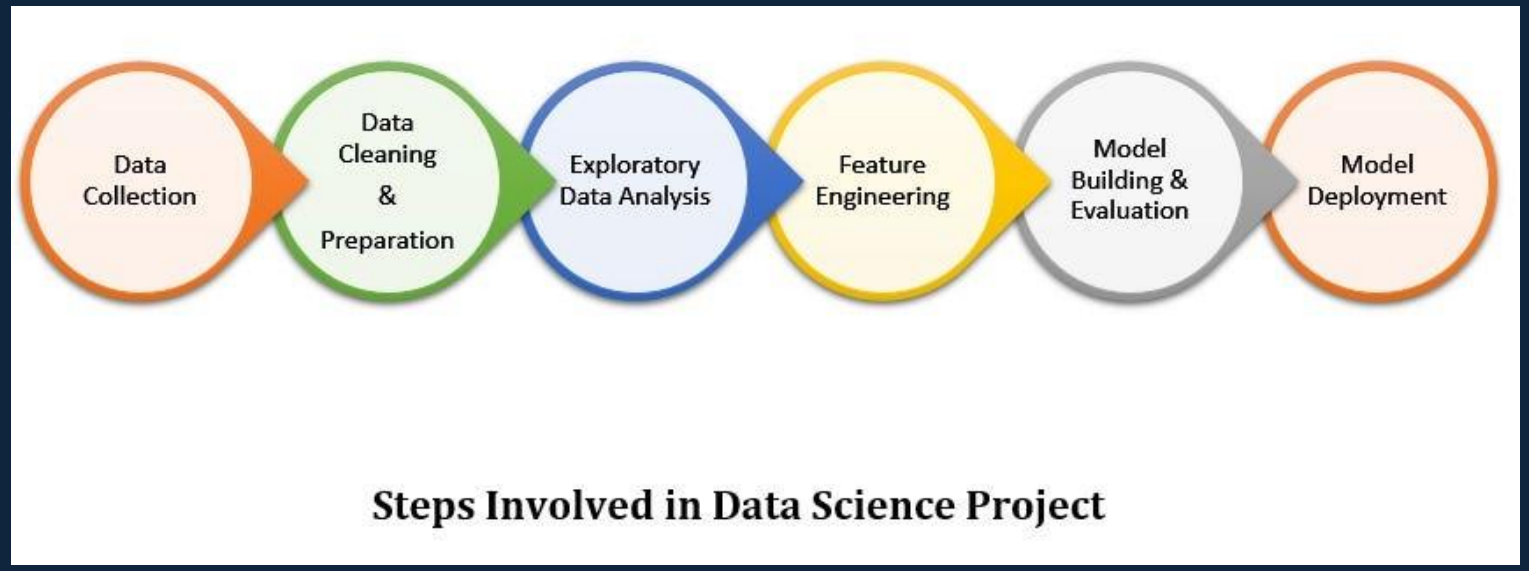
Introduction to Dataset

Why a Simulated Dataset?

- Confidentiality and Explainability of the Model

Methodology

- Preprocessing / EDA
- Feature Selection/Engineering
- Categorical Variable Encoding
- Class Imbalance
- Model Training and Evaluation



Data Size and Description

- 1.3m rows
- 22 Features (Categorical mostly)

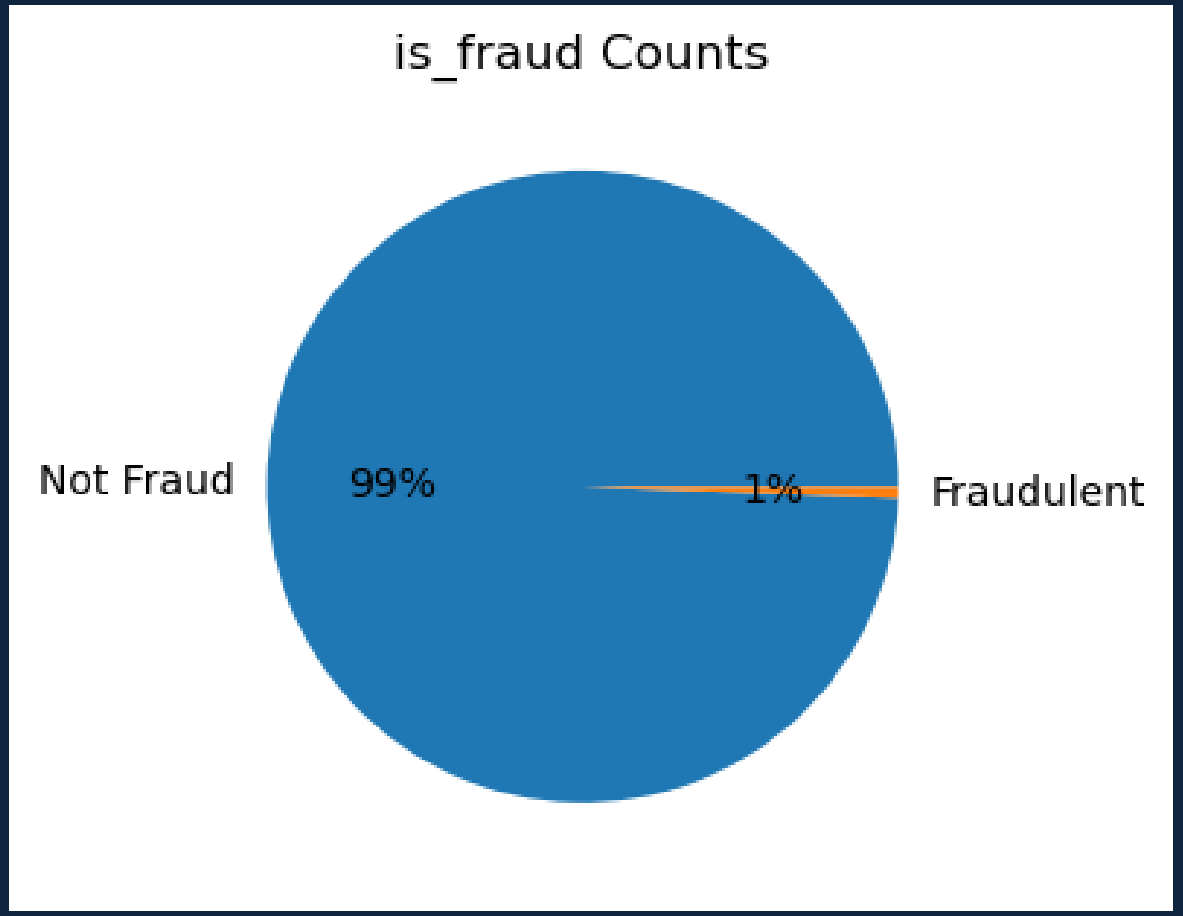
	cc_num	amt	zip	lat	long	city_pop	unix_time	merch_lat	merch_long	is_fraud
count	1.296675e+06	1296675.0	1296675.0	1296675.0	1296675.0	1296675.0	1.296675e+06	1296675.0	1296675.0	1296675.0
mean	4.171920e+17	70.0	48801.0	39.0	-90.0	88824.0	1.349244e+09	39.0	-90.0	0.0
std	1.308806e+18	160.0	26893.0	5.0	14.0	301956.0	1.284128e+07	5.0	14.0	0.0
min	6.041621e+10	1.0	1257.0	20.0	-166.0	23.0	1.325376e+09	19.0	-167.0	0.0
25%	1.800429e+14	10.0	26237.0	35.0	-97.0	743.0	1.338751e+09	35.0	-97.0	0.0
50%	3.521417e+15	48.0	48174.0	39.0	-87.0	2456.0	1.349250e+09	39.0	-87.0	0.0
75%	4.642255e+15	83.0	72042.0	42.0	-80.0	20328.0	1.359385e+09	42.0	-80.0	0.0
max	4.992346e+18	28949.0	99783.0	67.0	-68.0	2906700.0	1.371817e+09	68.0	-67.0	1.0

```
Index: 1296675 entries, 0 to 1296674
Data columns (total 22 columns):
 #   Column                  Non-Null Count
---  ------                  --------------
 0   trans_date_trans_time   1296675 non-null
 1   cc_num                  1296675 non-null
 2   merchant                1296675 non-null
 3   category                1296675 non-null
 4   amt                     1296675 non-null
 5   first                   1296675 non-null
 6   last                    1296675 non-null
 7   gender                  1296675 non-null
 8   street                  1296675 non-null
 9   city                    1296675 non-null
 10  state                   1296675 non-null
 11  zip                     1296675 non-null
 12  lat                     1296675 non-null
 13  long                    1296675 non-null
 14  city_pop                1296675 non-null
 15  job                     1296675 non-null
 16  dob                     1296675 non-null
 17  trans_num               1296675 non-null
 18  unix_time               1296675 non-null
 19  merch_lat               1296675 non-null
 20  merch_long              1296675 non-null
 21  is_fraud                1296675 non-null
```

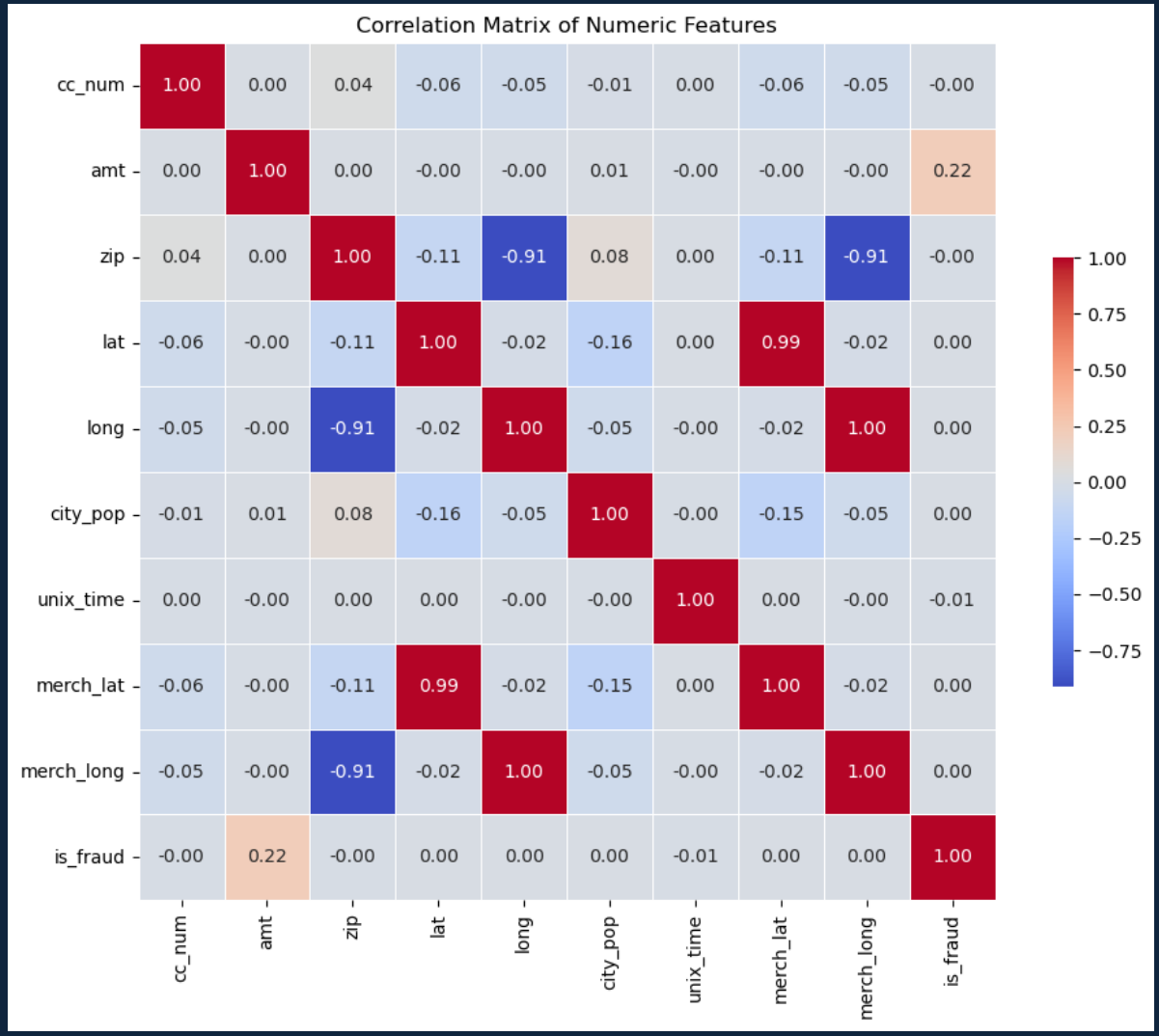
Data Imbalance

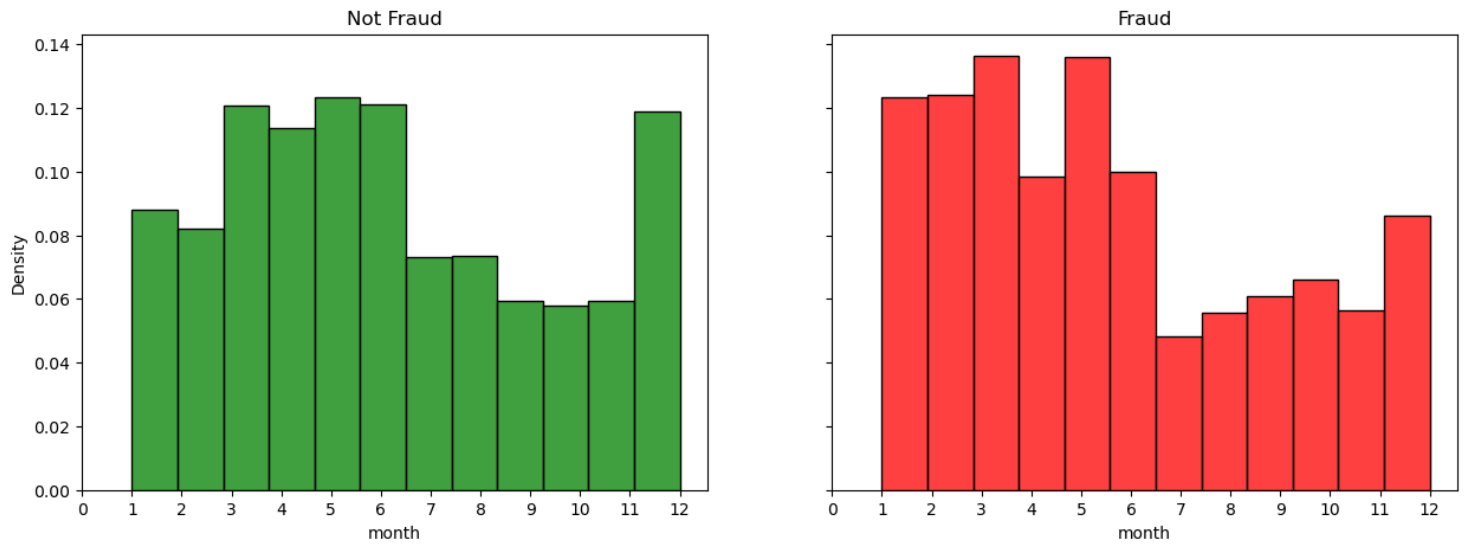
Fraud: ~7600

Not Fraud: ~1.3m



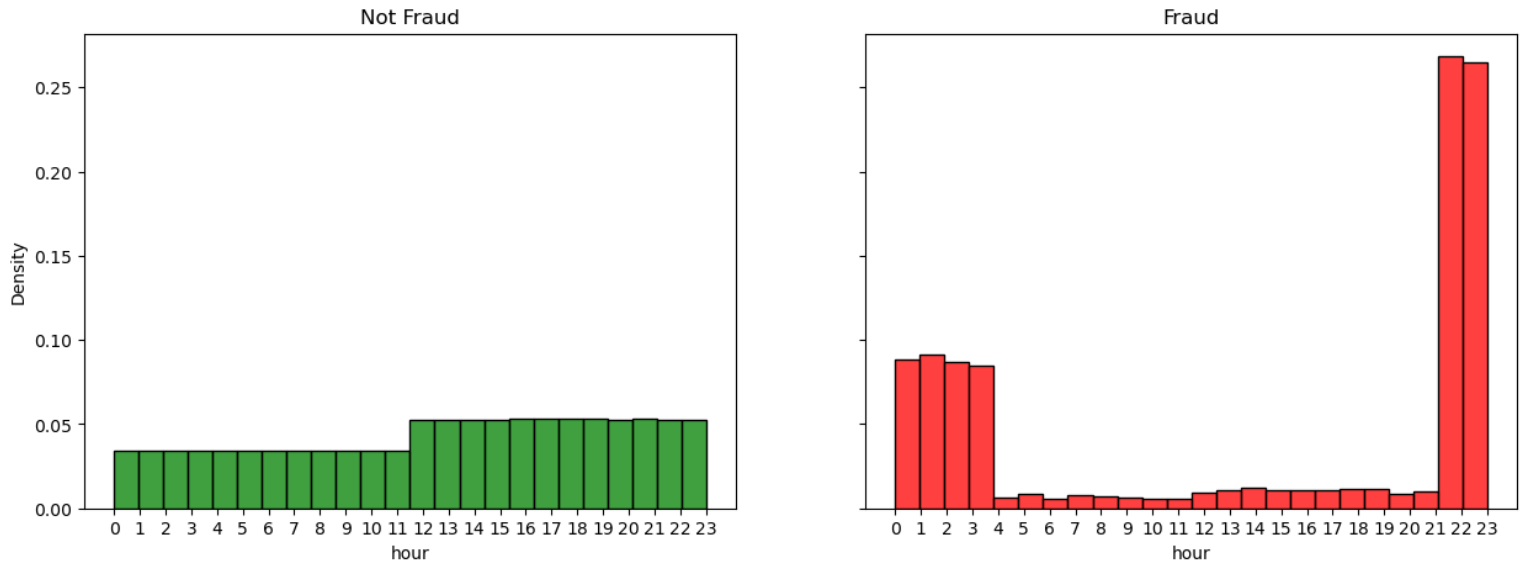
Correlation Matrices





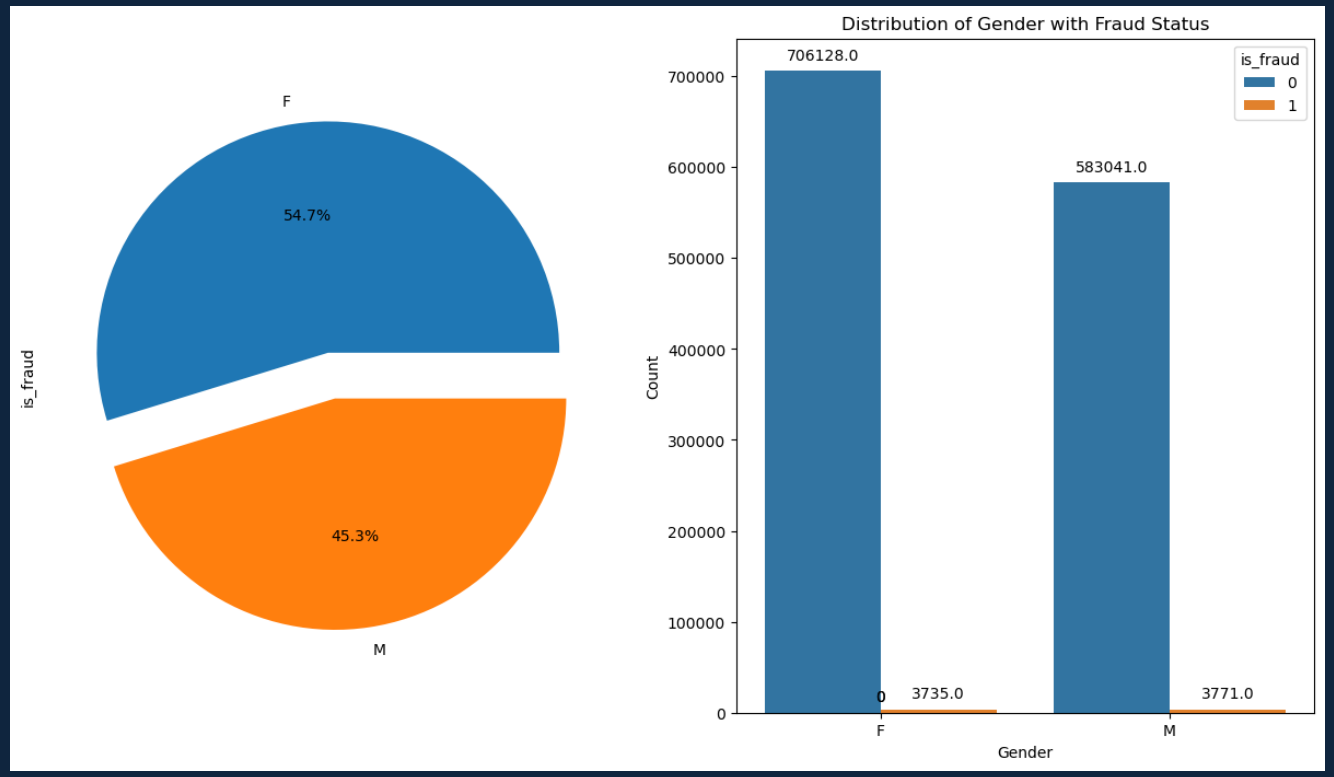
Feature Engineering

Time Variable split into Hours, Month, Day

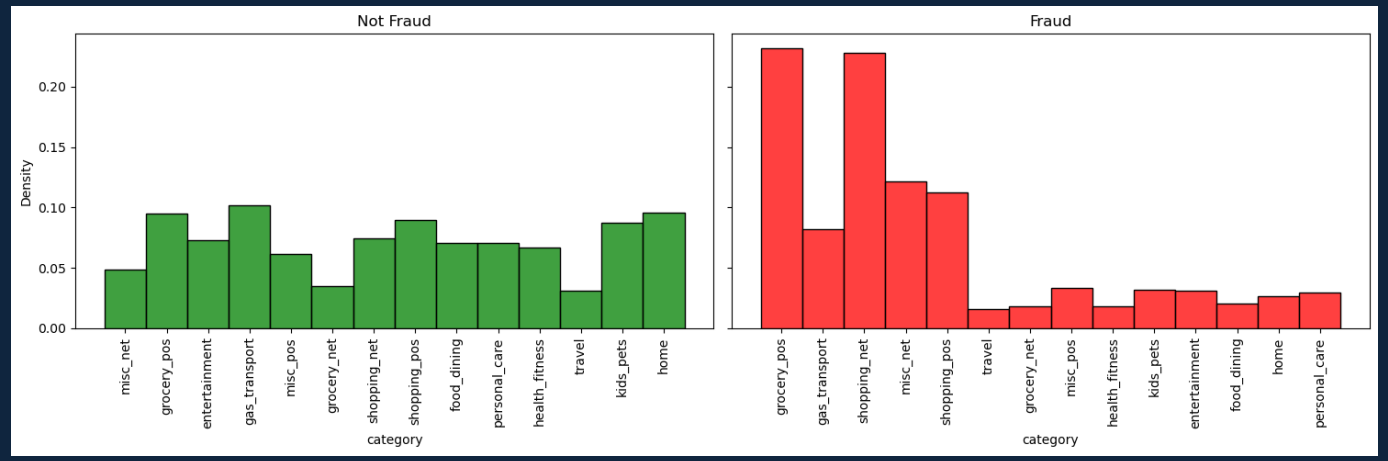


Exploring the Data

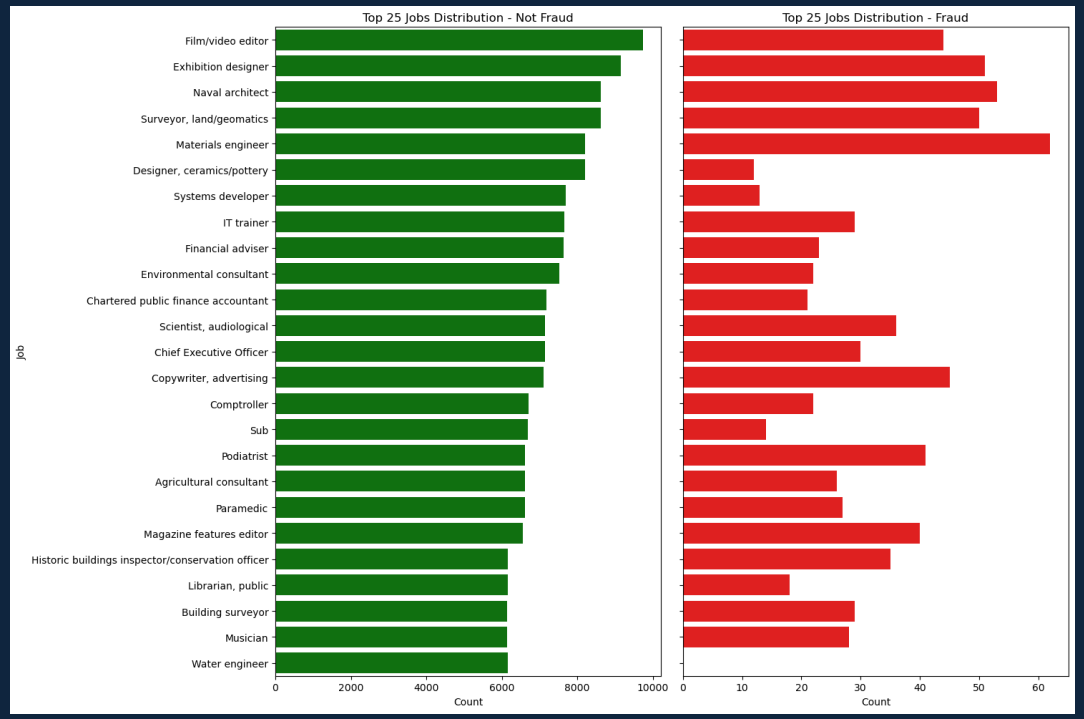
- More Females in Transaction data but about the same amount of fraud transactions between gender



Exploring the Data



Categories that were found in Fraudulent Transactions



Jobs of Credit Card Owner that were found in Fraudulent Transactions

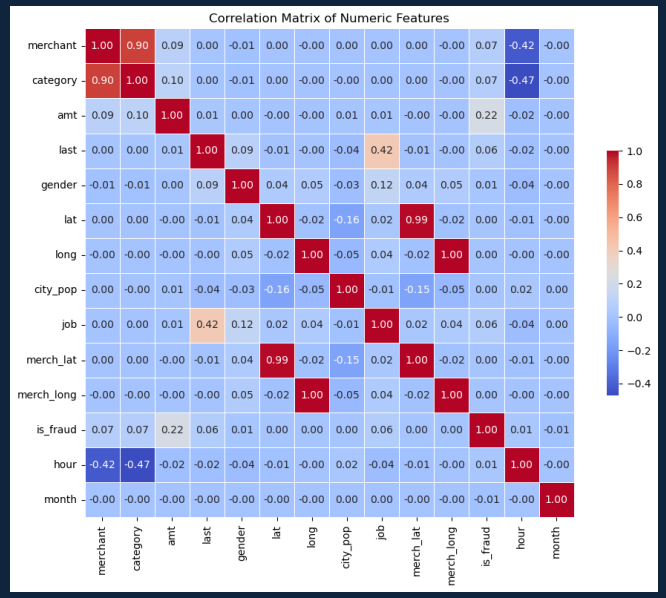
Which Categorical Features to Remove or Encode?

- **Remove:**
  - ‘first’, 'unix_time', 'dob', 'cc_num', 'zip', 'city','street', 'state', 'trans_num', 'trans_date_trans_time’
    - Multicollinearity and Irrelevant
- **Keep:**

	merchant	category	last	gender	job
count	1296675	1296675	1296675	1296675	1296675
unique	693	14	481	2	494
top	Kilback LLC	gas_transport	Smith	F	Film/video editor
freq	4403	131659	28794	709863	9779

Categorical Variables Encoding

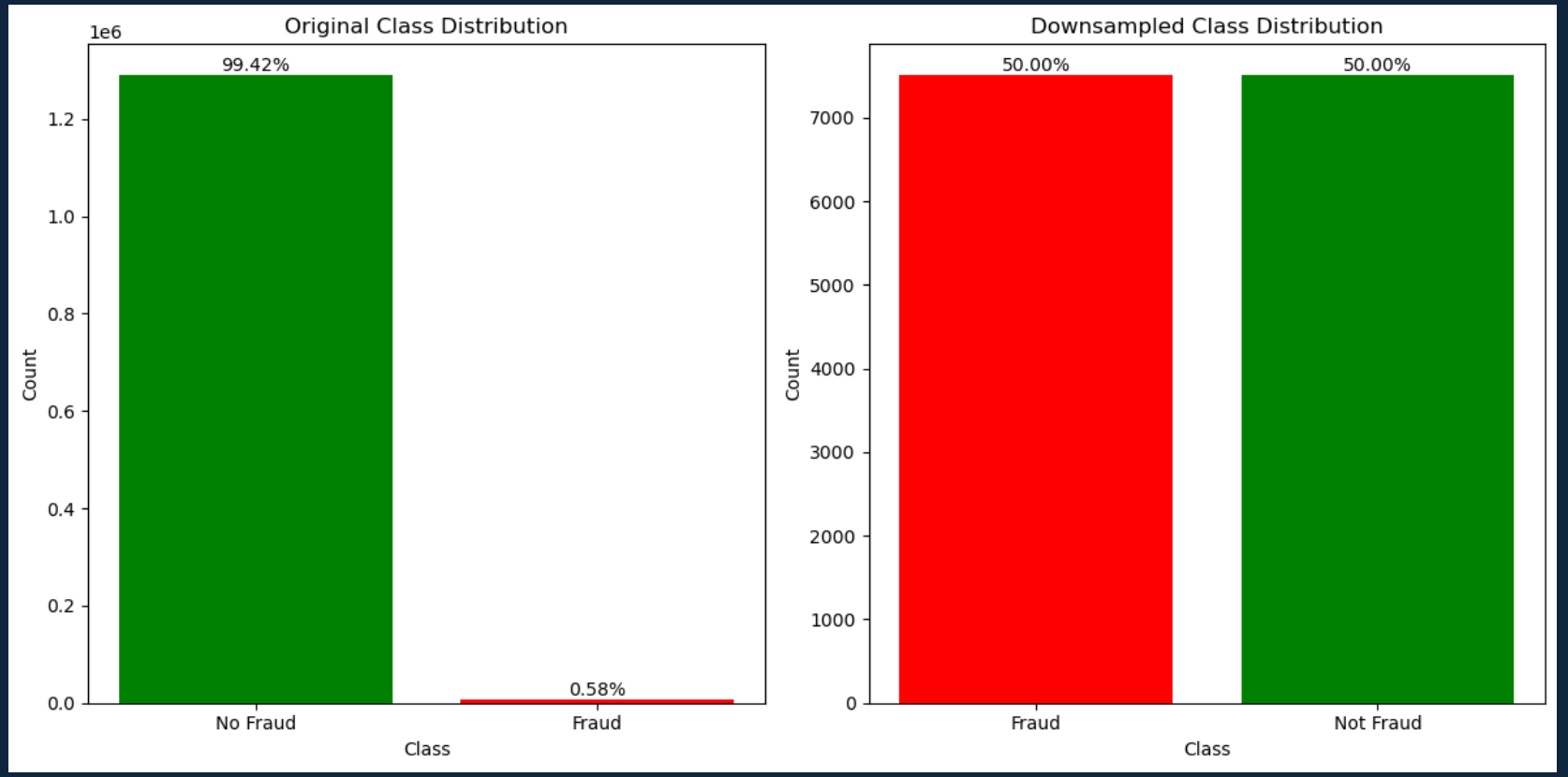
- A lot of unique categorical variables
  - One Hot encoding for gender
  - Weight of Evidence encoding from Literature
  - Label Encoder



	merchant	category	amt	last	gender	lat	long	city_pop	job	merch_lat	merch_long	is_fraud	hour	month
0	0.959326	0.924914	4.97	-2.469513	0	36.0788	-81.1781	3495	-1.080186	36.011293	-82.048315	0	0	1
1	0.663187	0.898799	107.23	-0.673638	0	48.8878	-118.2105	149	-0.904144	49.159047	-118.186462	0	0	1

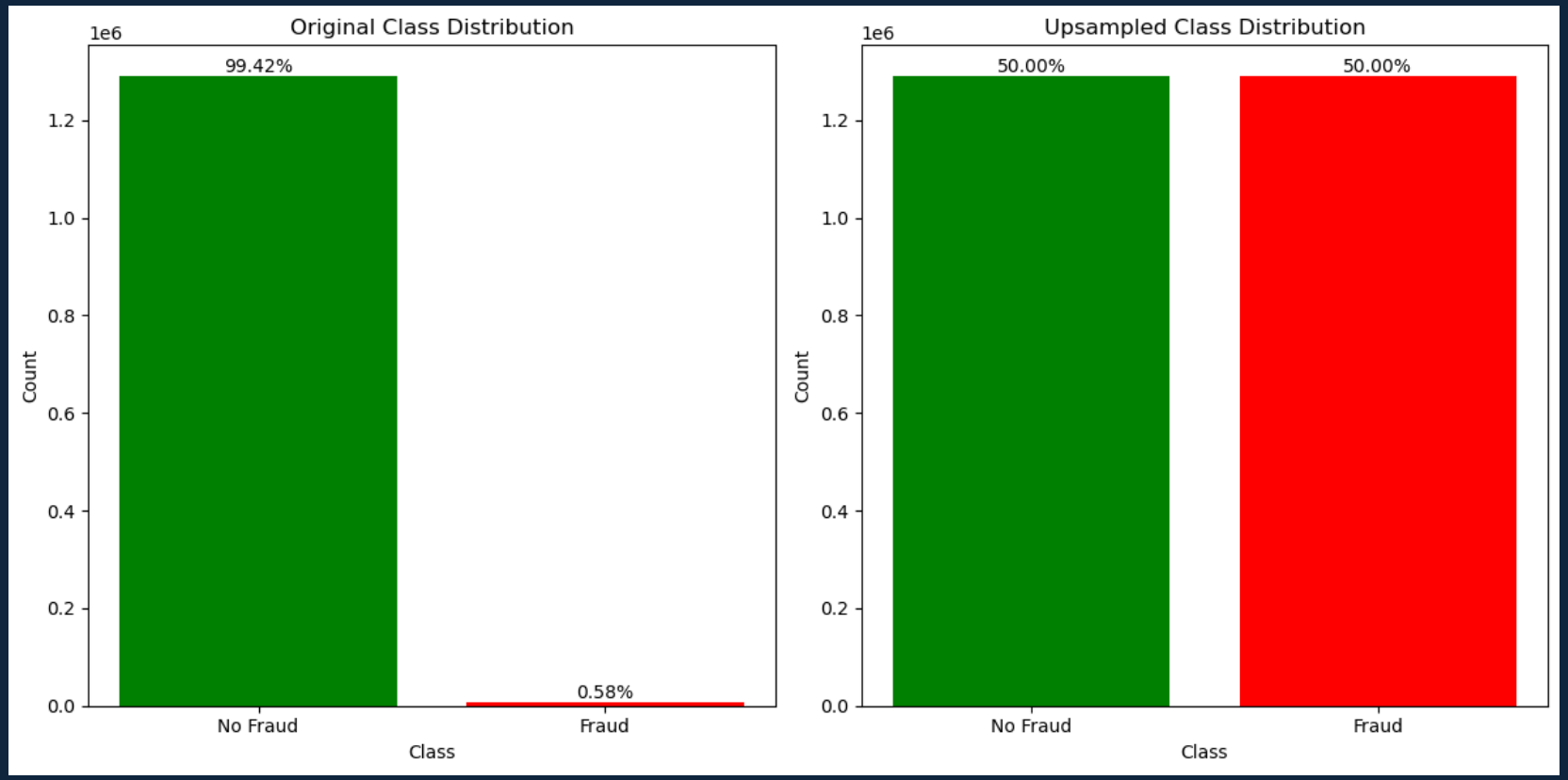
Dealing with Class Imbalance (Undersampling)

- From 1.3m samples to about 14k. Loss of a lot of data



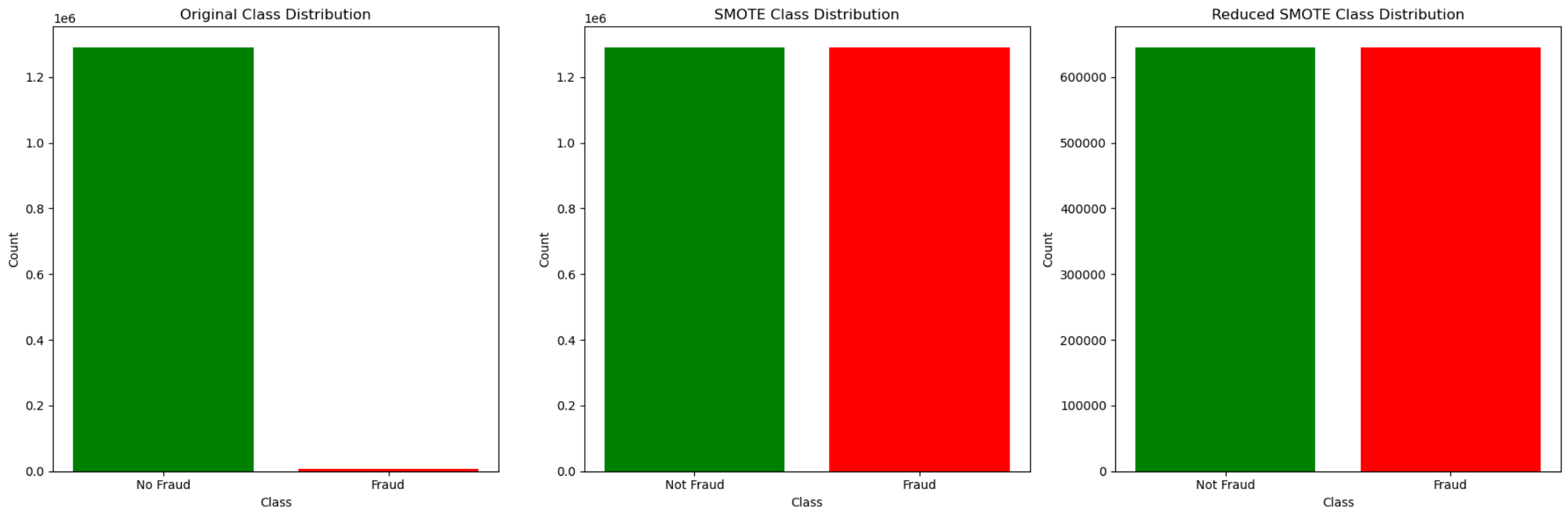
Dealing with Class Imbalance (Oversampling)

- From 1.3m samples to about 2.6m. Takes much longer to train!



Using SMOTE

- Reduced due to large data size

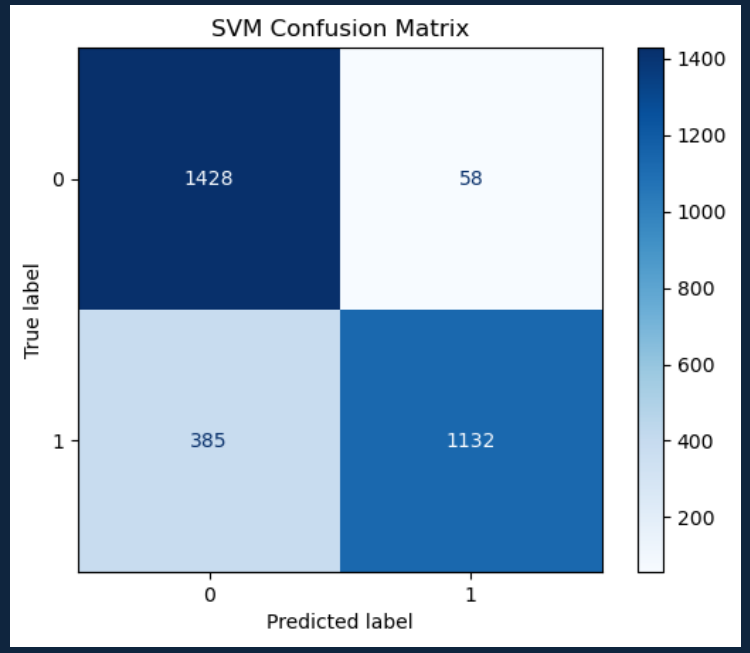


SVM Model and Features

- Kernal = ‘linear’
- Gamma=‘scale’

Model	ACC	AUC	Runtime
SVM	0.86	0.89	61.16s

- All features

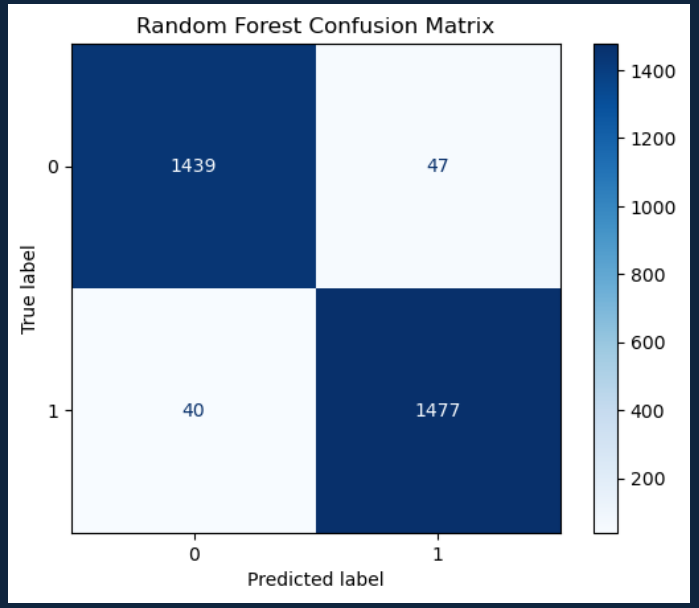


Random Forest Model and Features

- Criterion='entropy'
- Min_samples_split=3

Model	ACC	AUC	Runtime
RF	0.96	0.99	20.88s

- All features

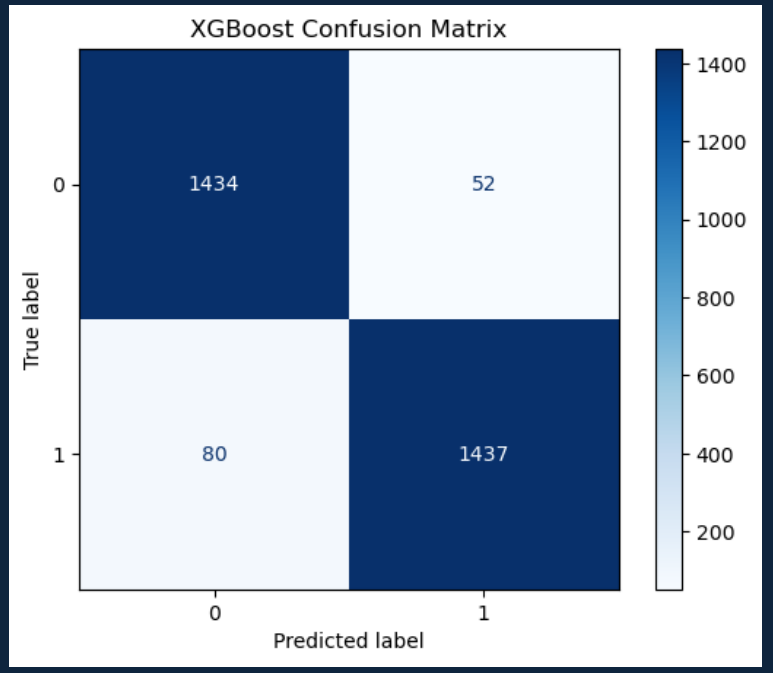


XGBoost Model and Features

- Learning_rate=0.1
- Max_depth=3

Model	ACC	AUC	Runtime
XGBoost	0.96	0.99	0.61s

- All features



Model Training with Wrapper Based FS

Model	ACC	AUC	Runtime	# of Features
Logistic Regression	0.85	0.90	0.26s	5
SVM	0.86	0.87	63.8s	3
Random Forest	0.97	1.00	19.15s	4
XGBoost	0.96	0.99	0.71s	3

Selected Features:

```
LR: ['merchant', 'category', 'amt', 'last', 'job’]
SVM:['merchant', 'amt', 'job’]
RF:['merchant', 'category', 'amt', 'hour’]
XGBoost:['category', 'amt', 'hour']
```

Conclusions and Best Model

- Using rebalancing
  - Undersampling: Overfit and losing a lot of data (from 1.3m to ~7000)
  - SMOTE: More information is retained but takes much longer to train and validate

- Most Important Features are `['merchant', 'category', 'amt', 'hour']`
- Best Model for runtime and model performance was XGBoost