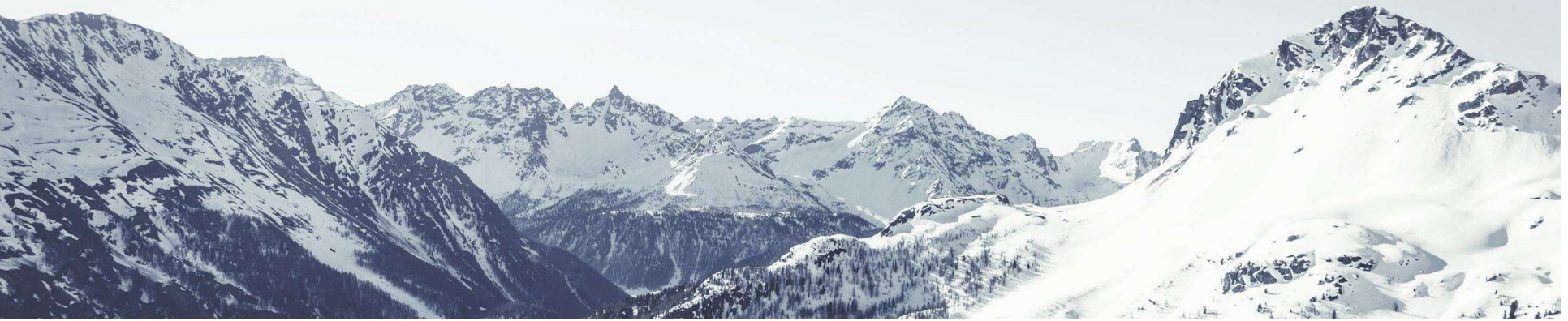
Credit Card Fraud Detection



Problem Statement

1.Create a classifier to predict whether a credit card transaction is fraudulent or not.

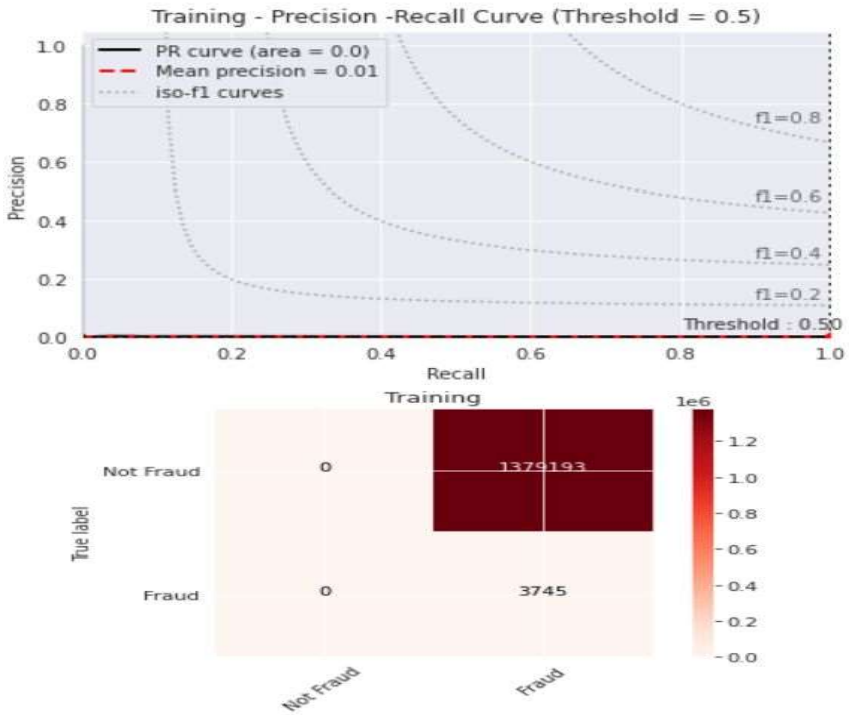
2. Model has to minimize overall cost.

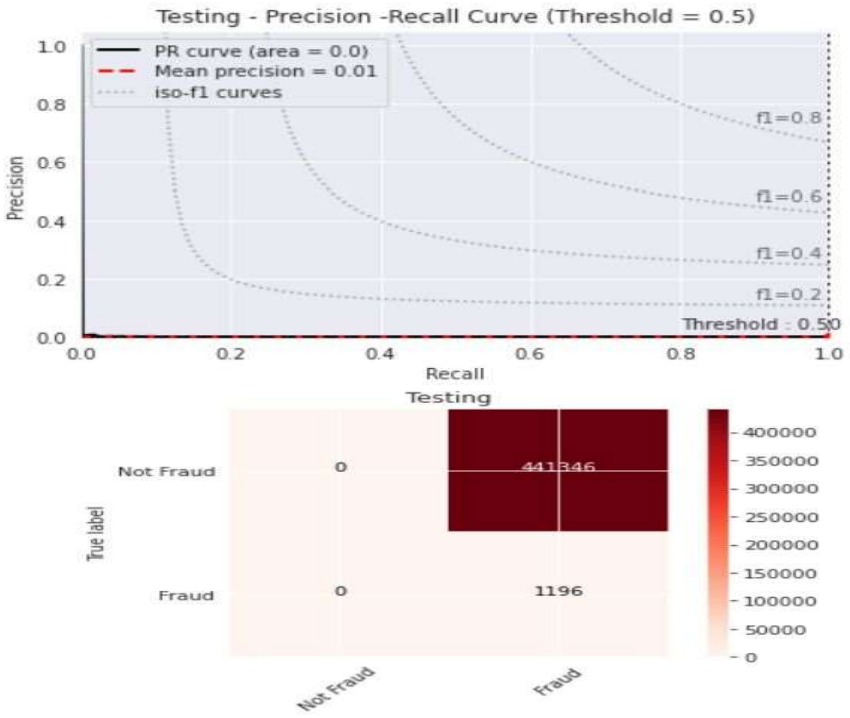
Data Source:

1.Sparkov Data Generation

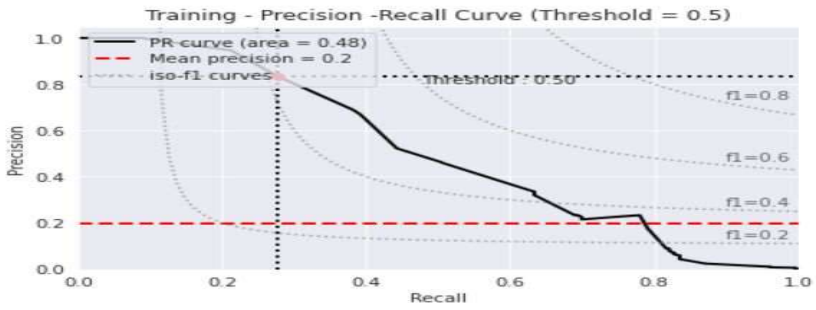
2. Data Imbalance - 0.23% Fraudulent

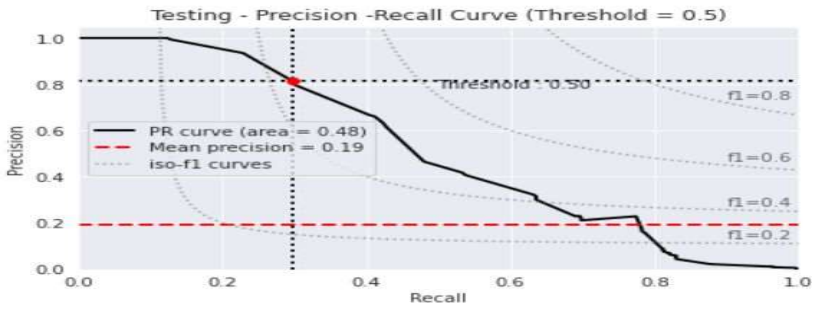
Baseline Model: Logistic Regression

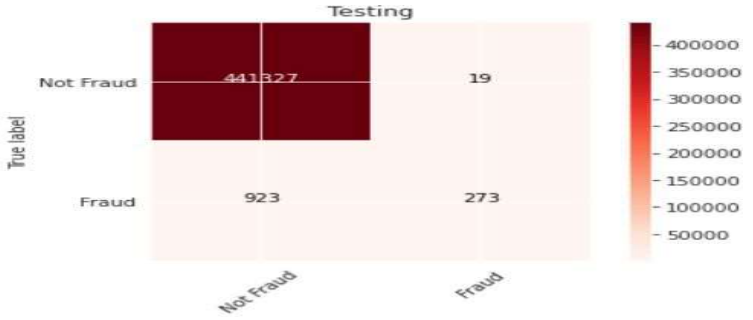




Baseline Model: XGBoost



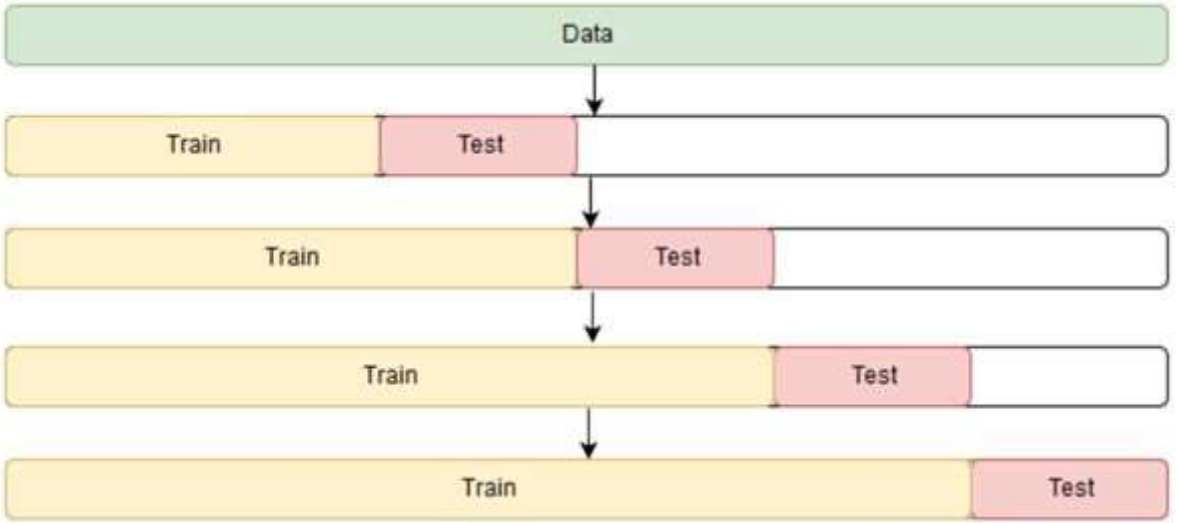




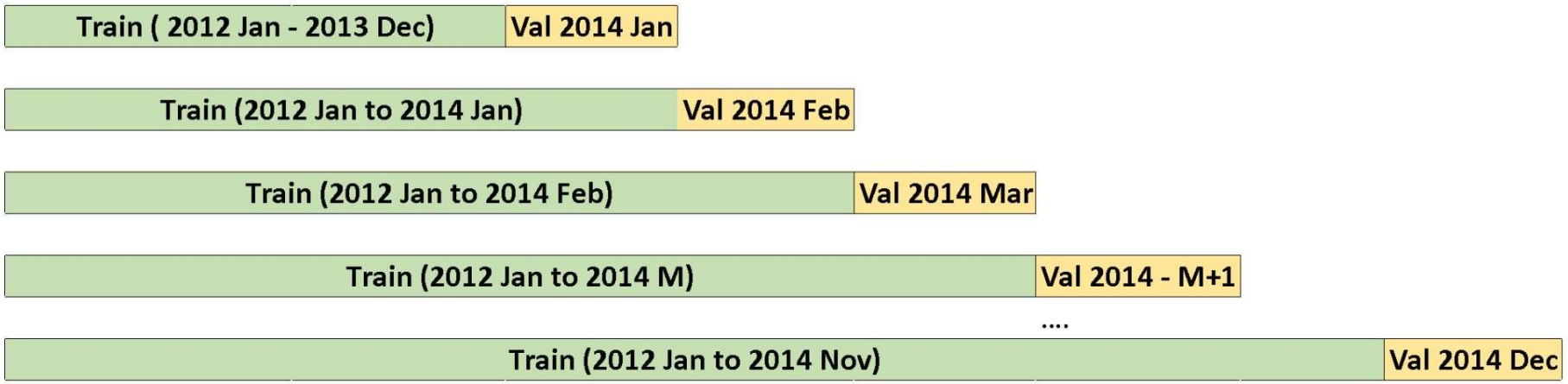
Train-test-split Methodology

Expanding training window

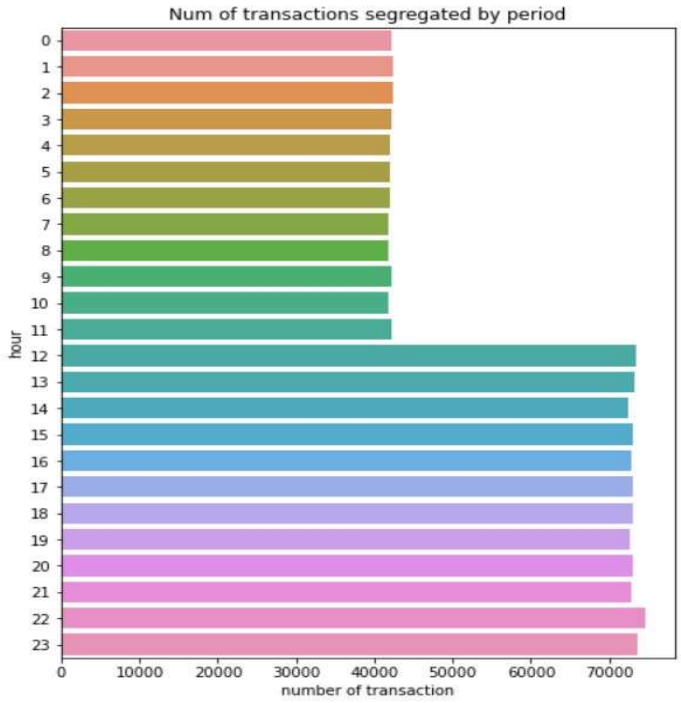
Train: 2012 to 2013
Validation: 2014
Test: 2015

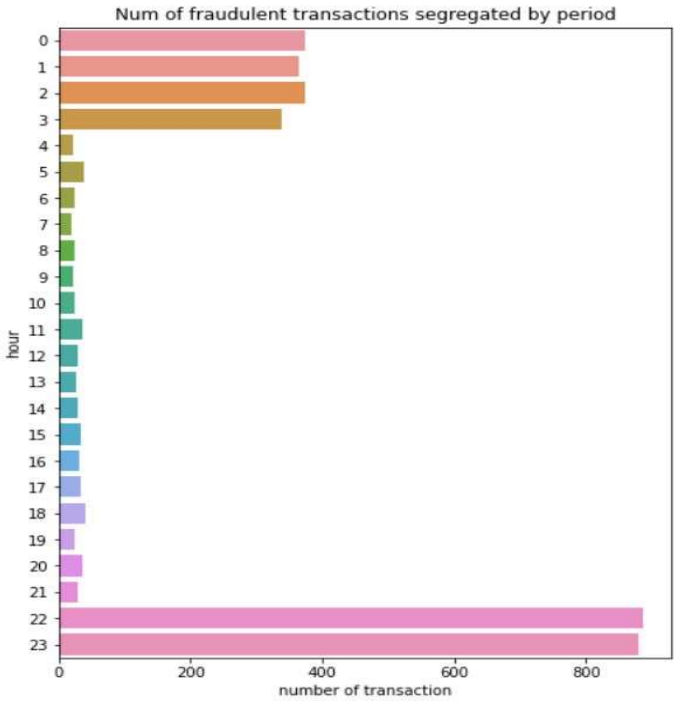


Train-test-split Methodology

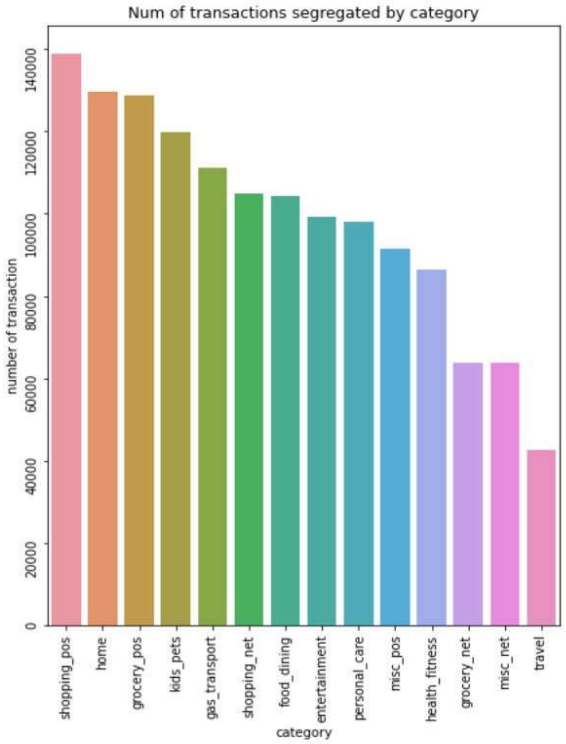


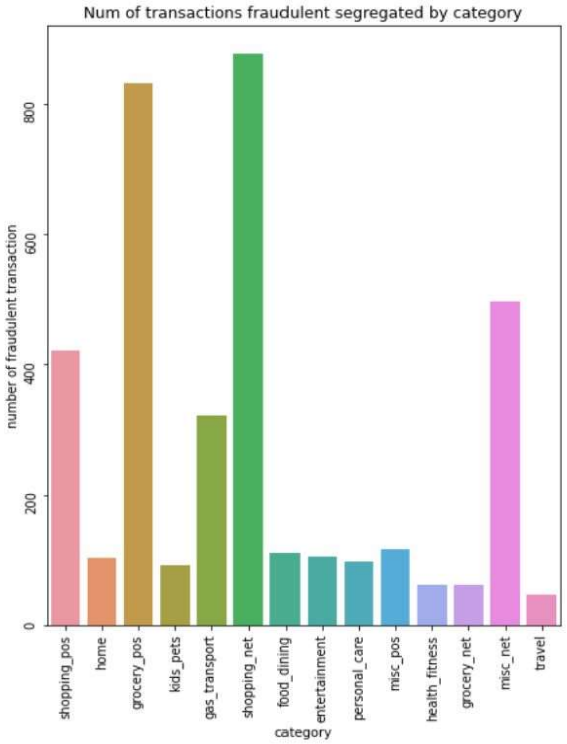
EDA :


Num of transactions segregated by period
hour
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
0
10000
20000
30000
40000
50000
60000
70000
number of transaction

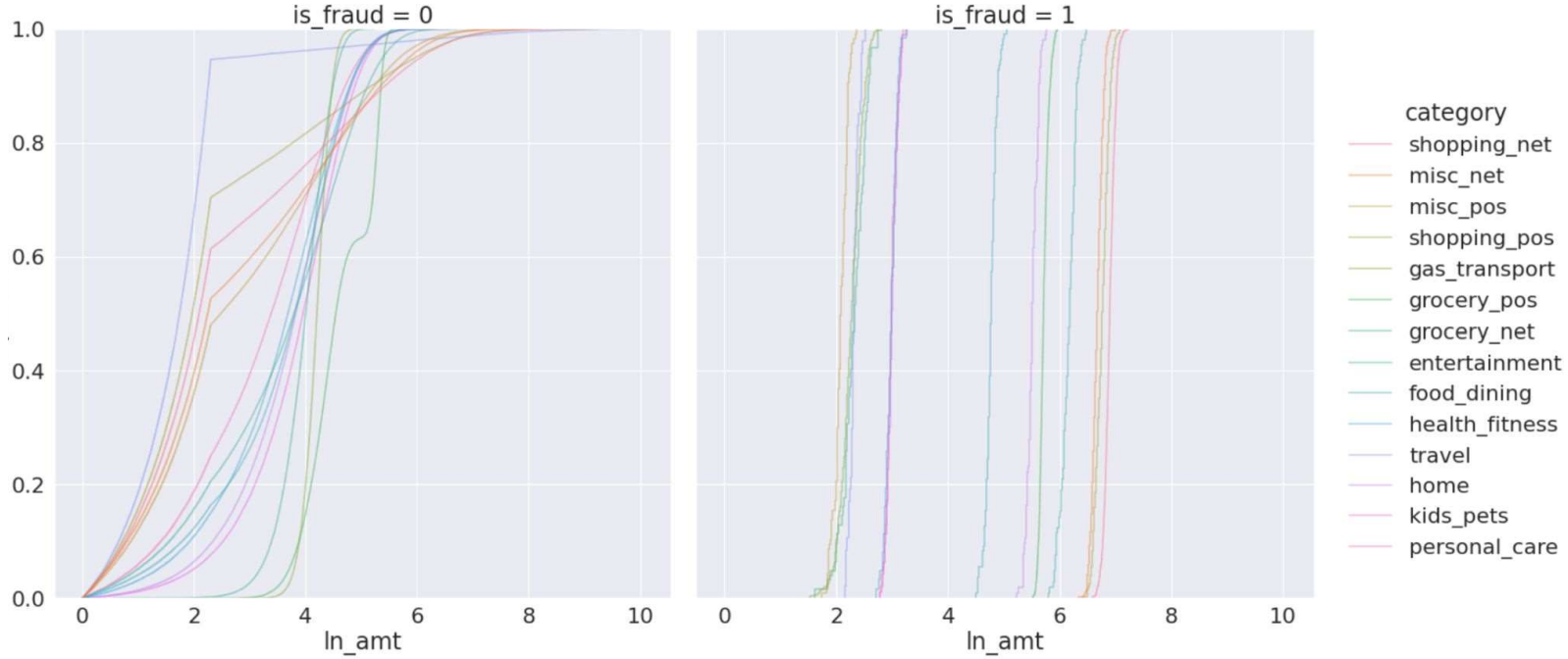

Num of fraudulent transactions segregated by period
hour
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
0
200
400
600
800
number of transaction

EDA:

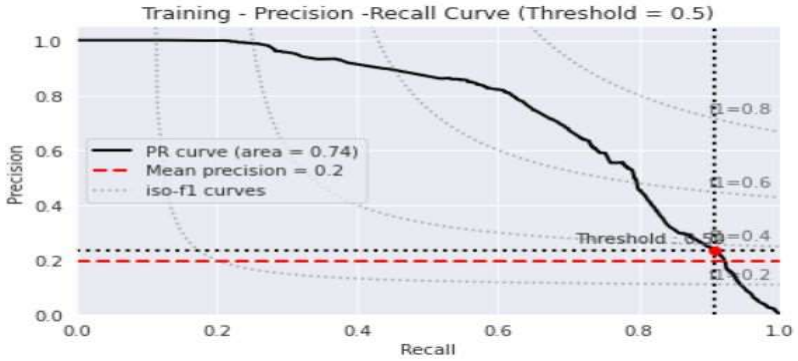

Num of transactions segregated by category
number of transaction
0
20000
40000
60000
80000
100000
120000
140000
shopping_pos
home
grocery_pos
kids_pets
gas_transport
shopping_net
food_dining
entertainment
personal_care
misc_pos
health_fitness
grocery_net
misc_net
travel
category

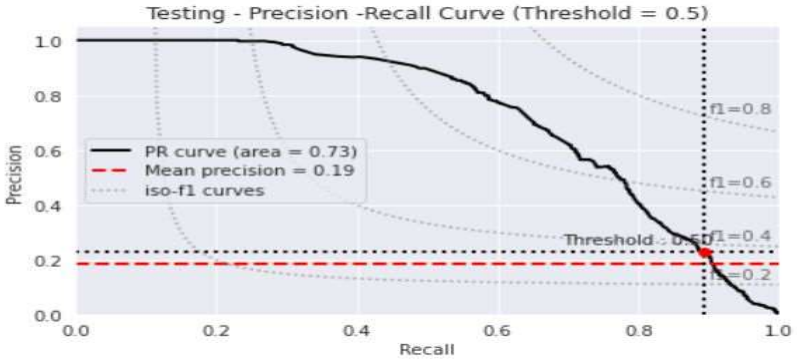

Num of transactions fraudulent segregated by category
number of fraudulent transaction
0
200
400
600
800
shopping_pos
home
grocery_pos
kids_pets
gas_transport
shopping_net
food_dining
entertainment
personal_care
misc_pos
health_fitness
grocery_net
misc_net
travel
category

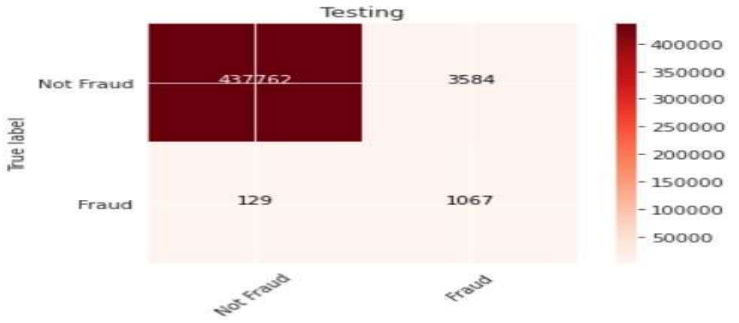
EDA :


is_fraud = 0
is_fraud = 1
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
ln_amt
category
shopping_net
misc_net
misc_pos
shopping_pos
gas_transport
grocery_pos
grocery_net
entertainment
food_dining
health_fitness
travel
home
kids_pets
personal_care

Final Model : XGBoost

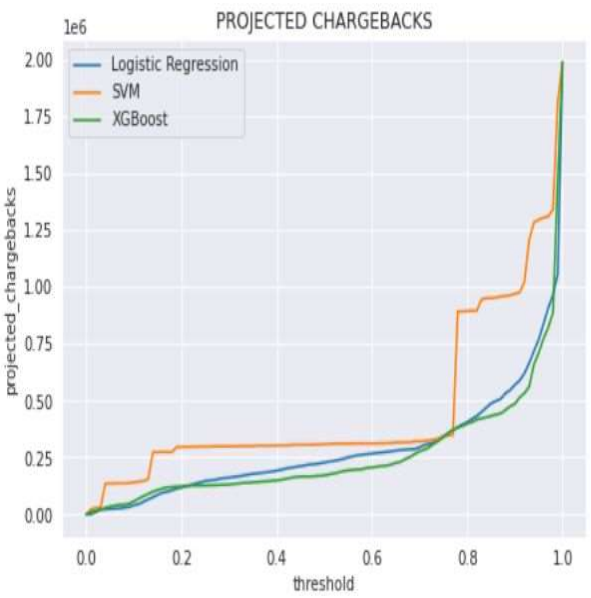


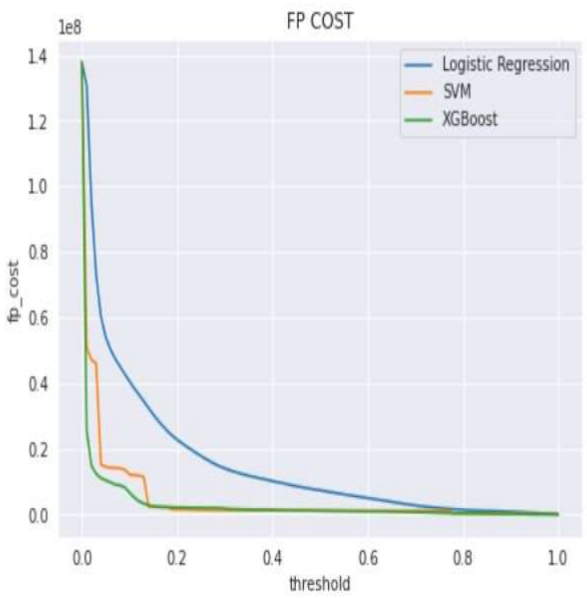


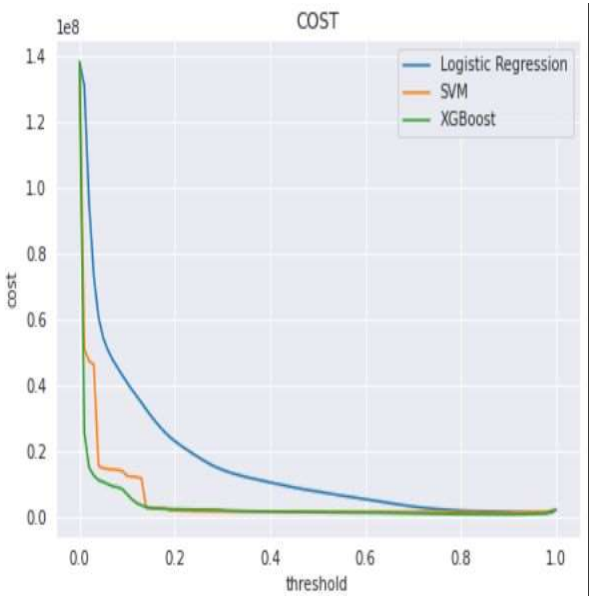


Recommendations : Threshold selection

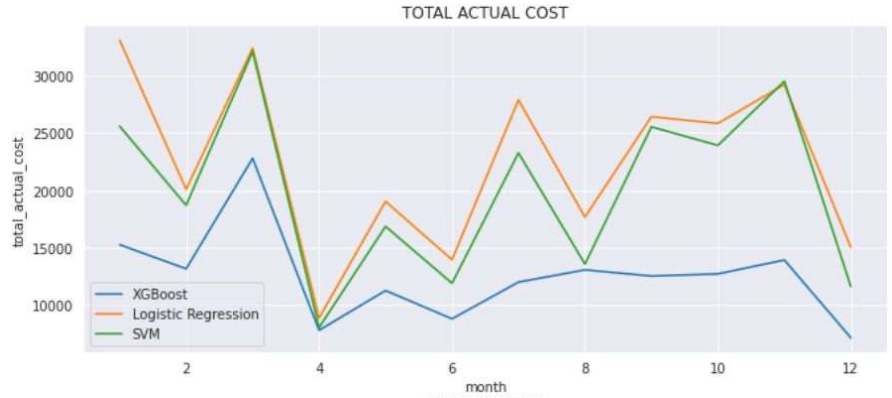
Assumptions - $100 marginal cost per False positive.
- $531 marginal cost per False negative

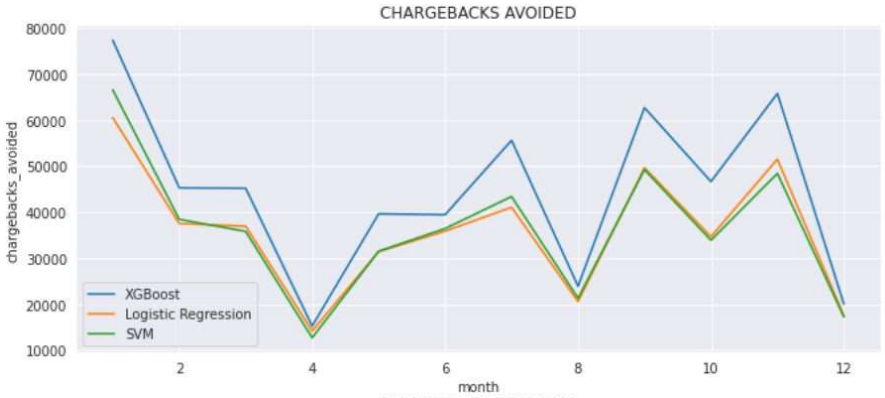


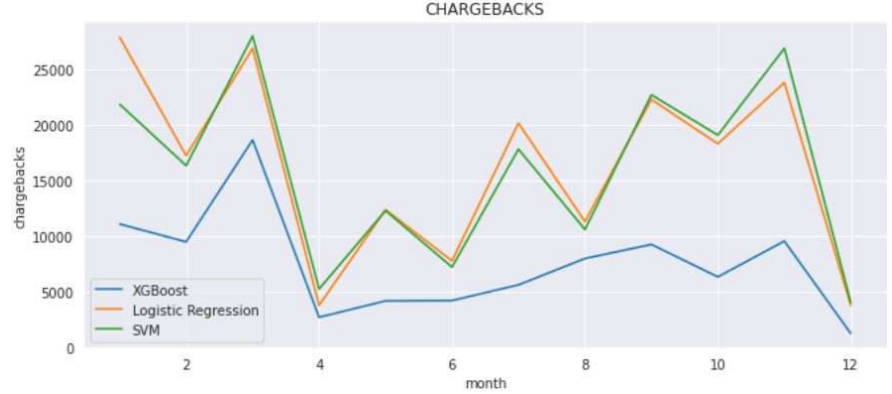


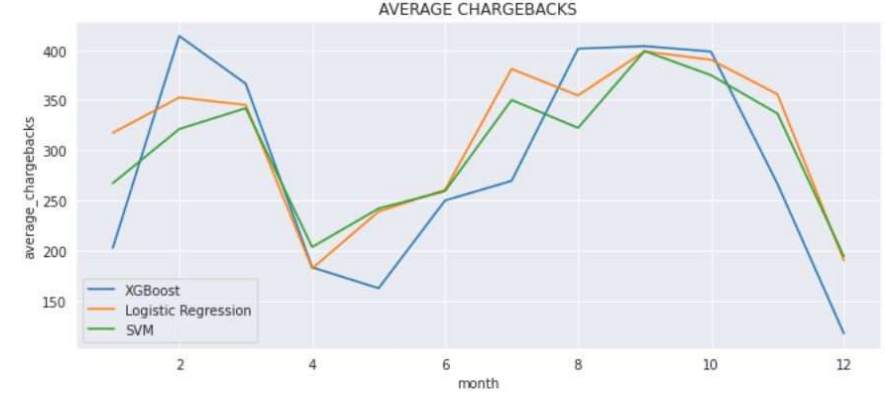


Model performance :









Deliverables

False Positive

0.13%

Out of almost 450,000 transactions

Correctly identified

74%

Of all fraudulent transactions that occurred in 2015

Total Savings

$536,324

From correctly identifying fraudulent transactions

Future work:

1. Implementation on scala
2. Explore using SMOTENC
3. Explore using other anomaly detection techniques

Thank you.