CREDIT CARD FRAUD DETECTION MODEL

USING

- LOGISTIC REGRESSION
- XGBOOST
- NEURAL NETWORK

Model Description & Usage

Purpose: Estimate the probability of Fraudulent transaction for a customer of the bank when a credit/ debit card swiped

Uses : Can be used to block the card when the fraudulent transaction is detected

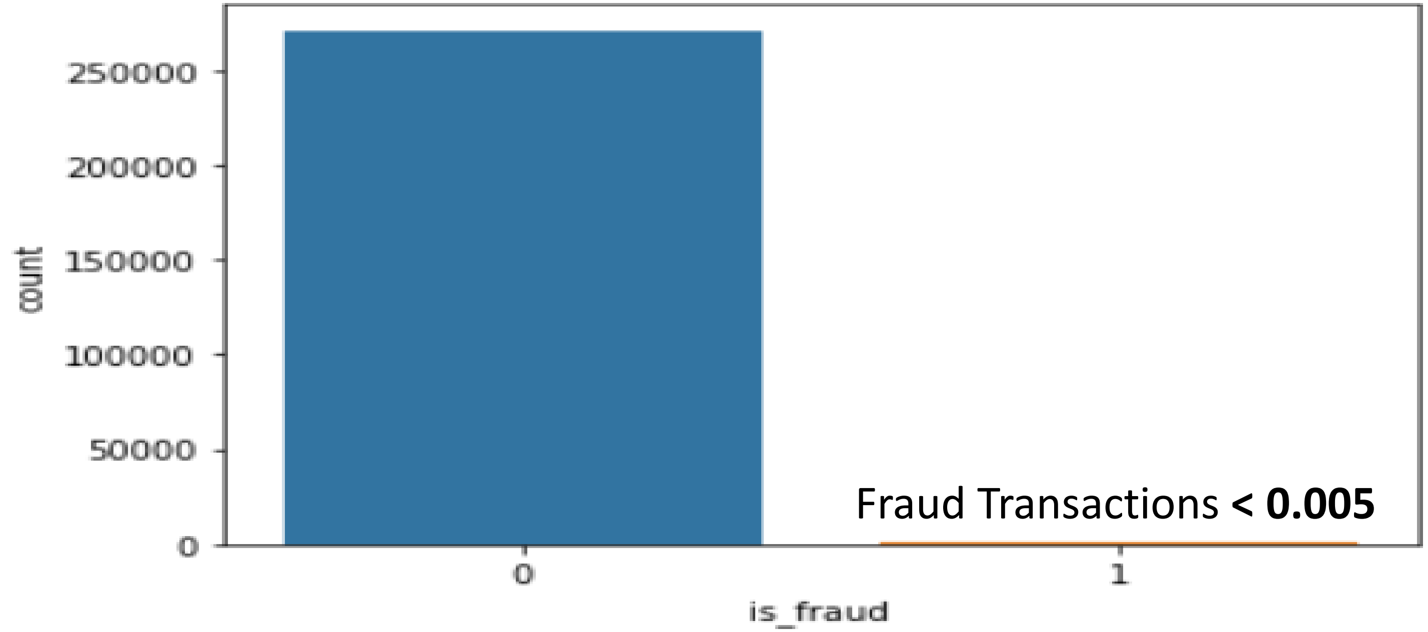
Strategy: P(.90) =block, P(0.7-0.9) = send them a message and block temp for few minutes until we get a reply, P(<.70) = no action required

Agenda

1. **Variables Explained**
   - Y – variable
   - X - Variable
2. **Train-Test Split**
3. **Feature Engineering**
   - Transformations
   - Target Encoding
   - Aggregate Encoding – Unique selling point
   - How Feature Importance changes before and after aggregate encoding
4. **Models & Hyper parameter tuning**
   - Logistic Regression
   - XGBoost
   - Neural Network
5. **K-fold validation score comparison**
6. **Rank ordering analysis**

Y Variable Explained

Non-Fraud Transactions: > **99.5%**



X Variable (1296675 Observations)

Attributes	Type	Unique Values
Trans_date_trans_num	Date – Time	Time Ranges from 01-01-2019 to 01-12-2019
Cc_num	int	
Merchant	Categorical	169
Category	Categorical	14
Gender	Categorical	2
Name	Object	
Job	Categorical	28

Attributes	Type	Unique Values
Amt	Continuous	
City	Object	
State	Object	
Zip	Object	946
City_population	continuous	
Date_of_birth	Date	

Attributes	Type	Unique Values
Latitude, Longitude	Float	
Merch- Lat, Long	Float	

Test – Train Split

My train and split will be based on time

Data set	Time Range
Train Set	01 – 01 -2019 to 01 – 10 -2019
Test Set	01 – 11 – 2019 to 01 – 12 - 2019

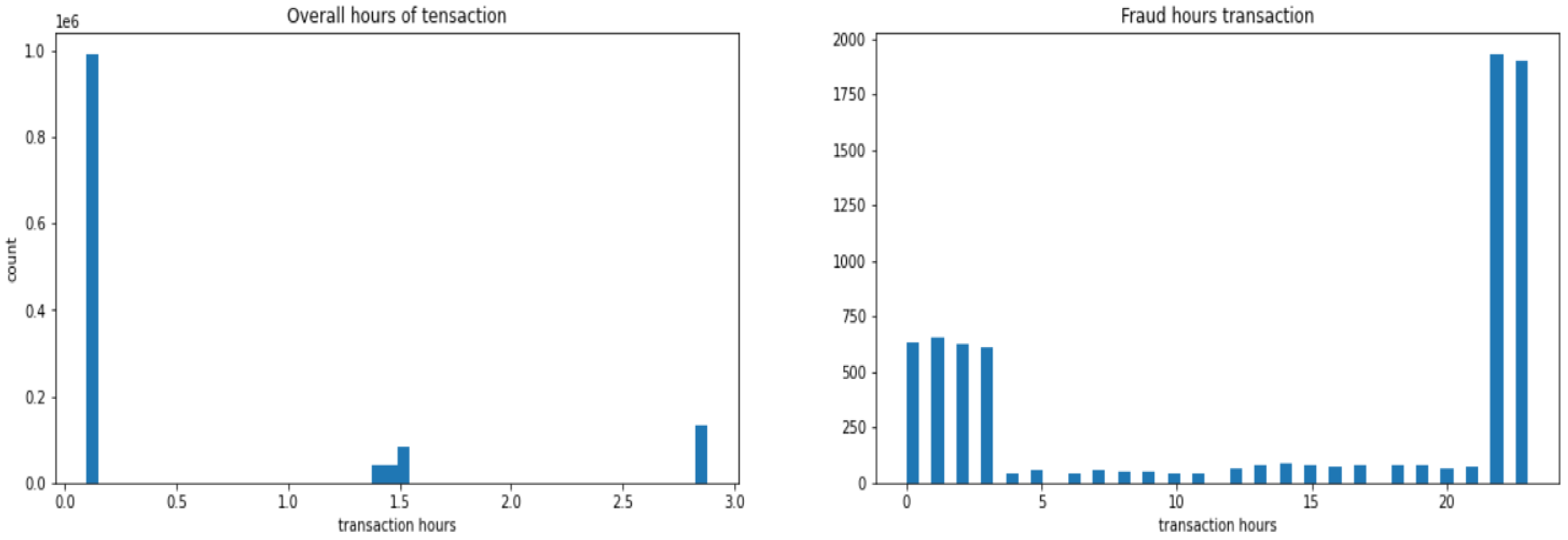
Train Data	
Non-Fraud	907671
Fraud	4178 (0.46%)

Test Data	
Non-Fraud	389003
Fraud	5121 (1.31%)

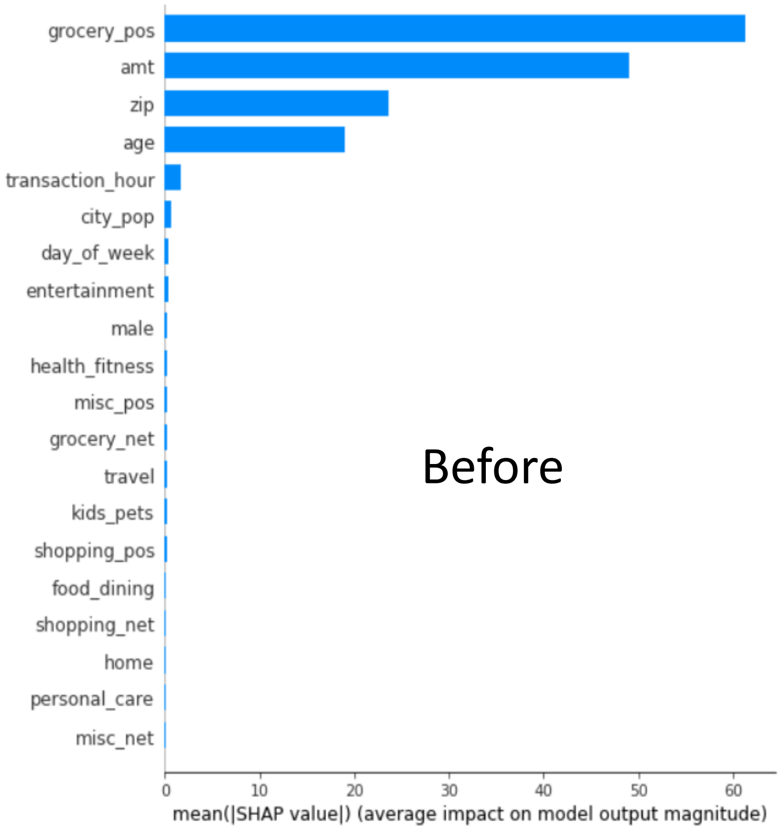
Feature Engineering

Attributes	Transformed as
Trans_date_trans_num	• Transaction hour • Fraud rate
Trans_date_trans_num	• Transaction week • One hot encoding
Category	One hot encoding
Gender	Male = 1
Job	• High_risk, low_risk • One hot encoding
Latitude, Longitude	Distance
Merch- Lat, Long	

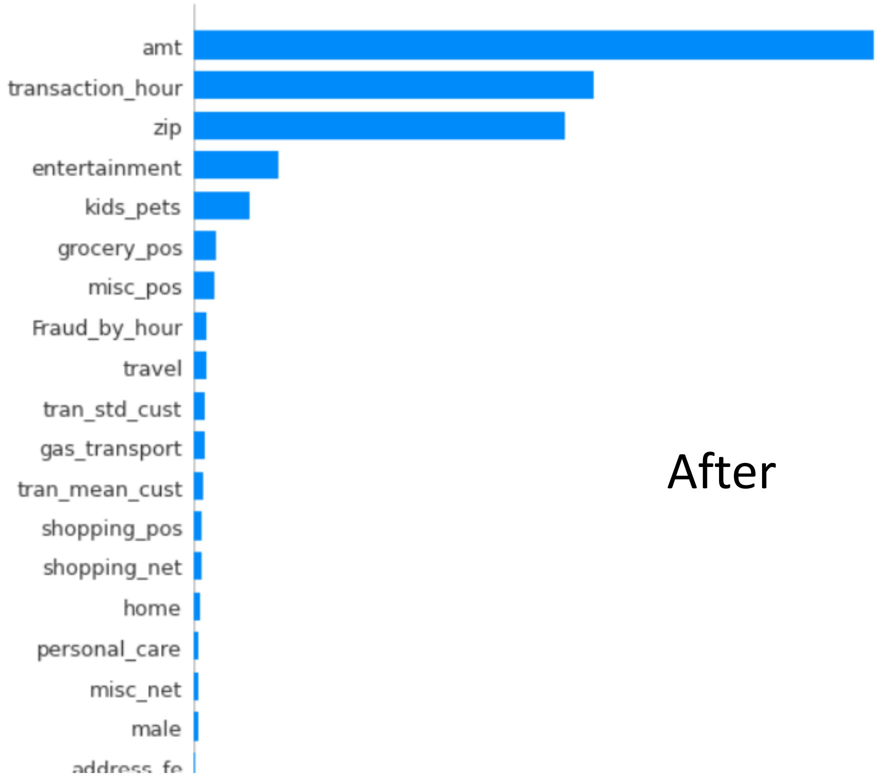
Attributes	Transformed as
Amt	Normalised
City	Fraud rate at the Zip code Target encoded – (street+city+state)
State	
Zip	
City_population	Continuous
Date_of_birth	• Age Calculated • Normalised
Name, cc_num	Agregate encoded- Calc avg amt, std amt



Hour of the transaction – EDA



Before



After

How my aggregation encoding helped in feature selection

Model 1 – Logistic Regression

```
lr_model = LogisticRegression(solver='saga', max_iter=500)
```

Parameters:

By default -> solver = lbfgs & max_iter = 100

Changed to Solver – ‘saga & Max_iter – 500

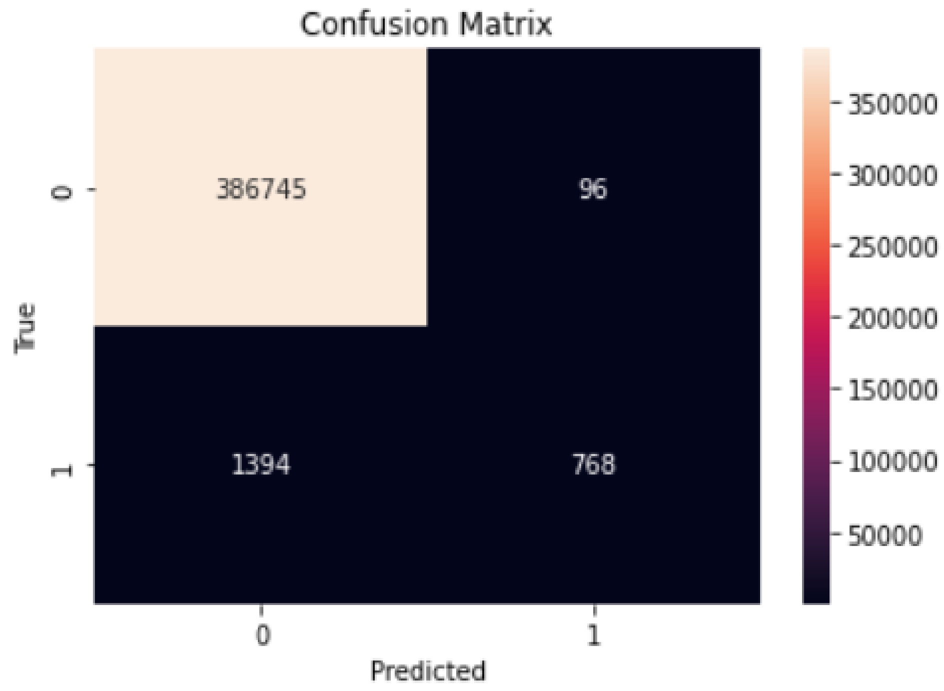
Problem:

Convergence problem

Max iteration reached

Performance

	precision	recall	f1-score	support
0	1.00	1.00	1.00	386841
1	0.89	0.36	0.51	2162
accuracy			1.00	389003
macro avg	0.94	0.68	0.75	389003
weighted avg	1.00	1.00	1.00	389003



Model 2 – XGBoost

Initial parameters:

'objective':'	binary:logistic',
'max_depth':	6,
'learning_rate':	1.0,
'n_estimators':	20

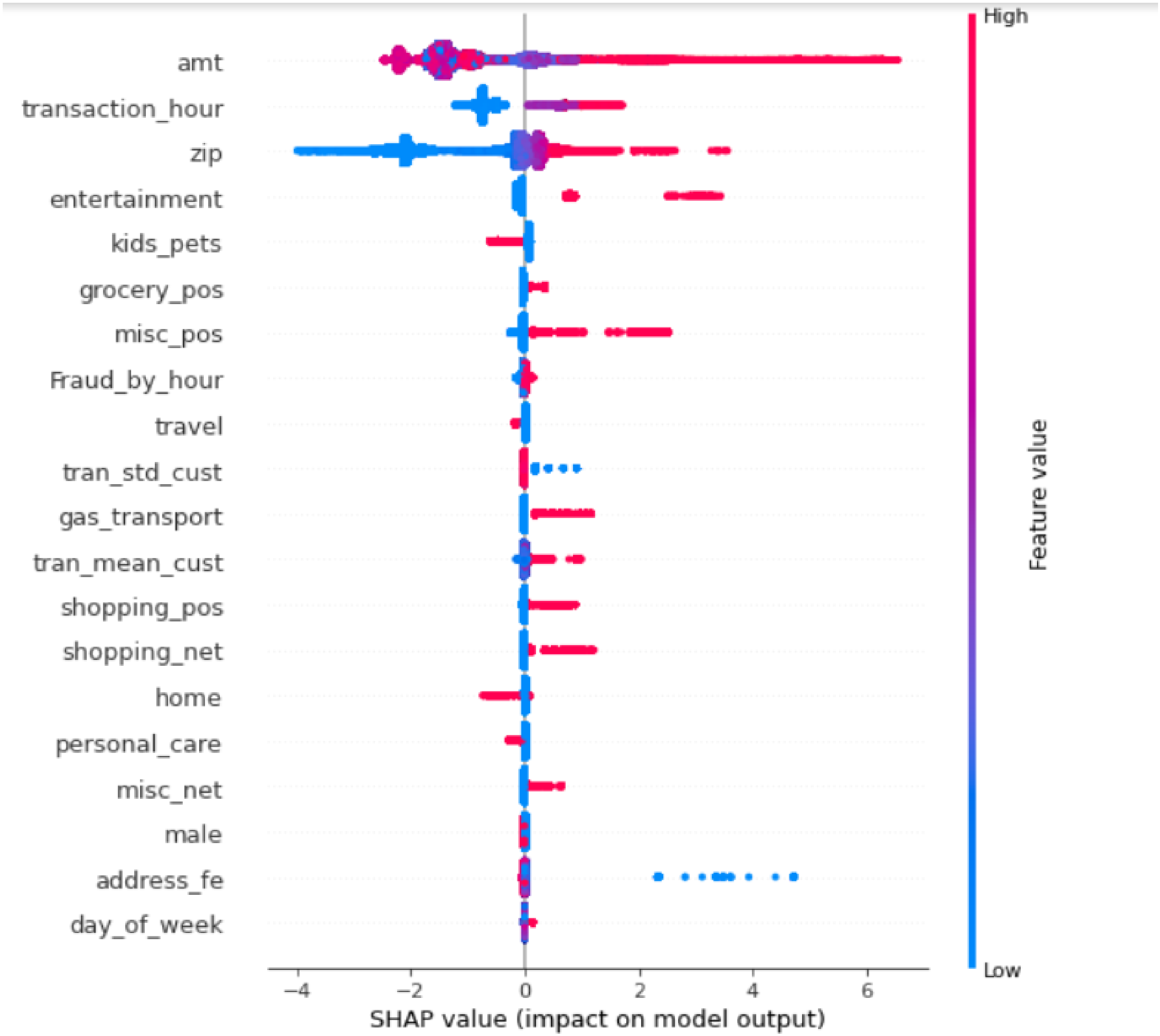
Accuracy
0.897

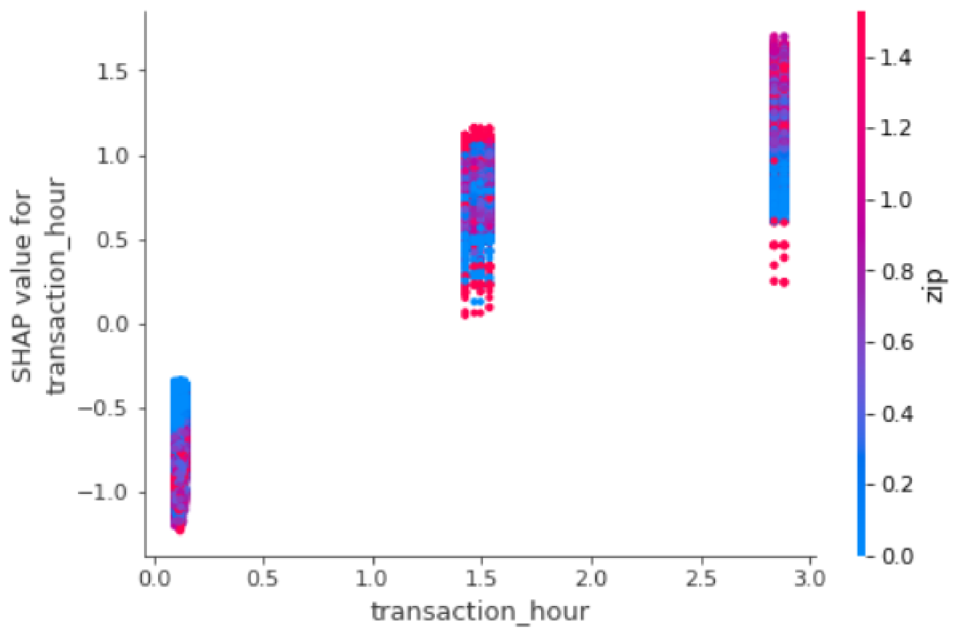
Hyper-parameter Tuning

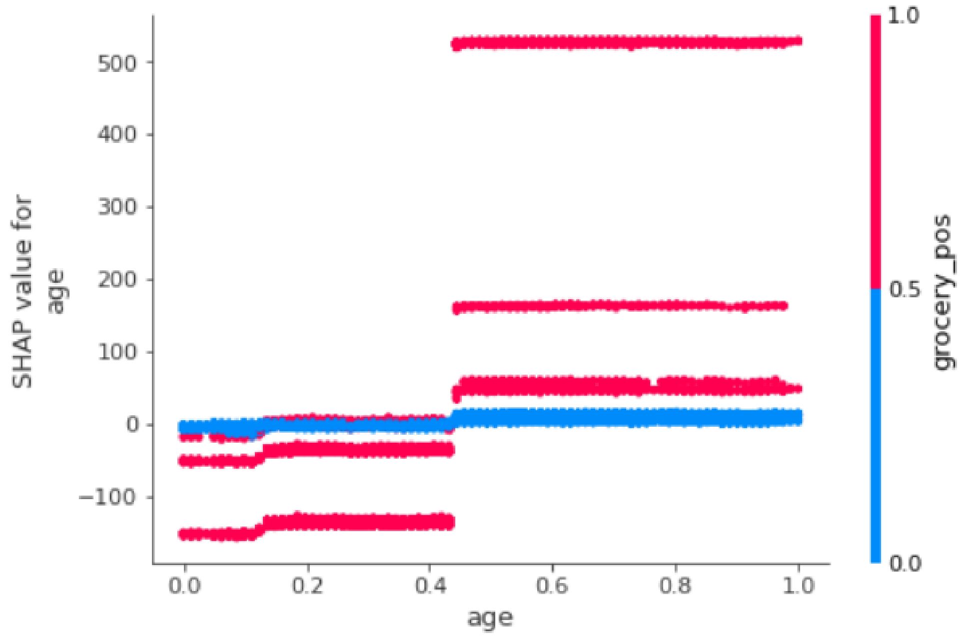
Parameter	Values		
n_estimators	10	50	100
Learning_rate	0.1	0.2	0.3
Max_depth	3	4	5

	# Trees		Depth	AUC Train	AUC Test	Learning rate
15	100	NaN	3	0.999045	0.99863	0.3
16	100	NaN	3	0.999814	0.99896	0.3
17	100	NaN	3	0.999976	0.999008	0.3

Feature Importance



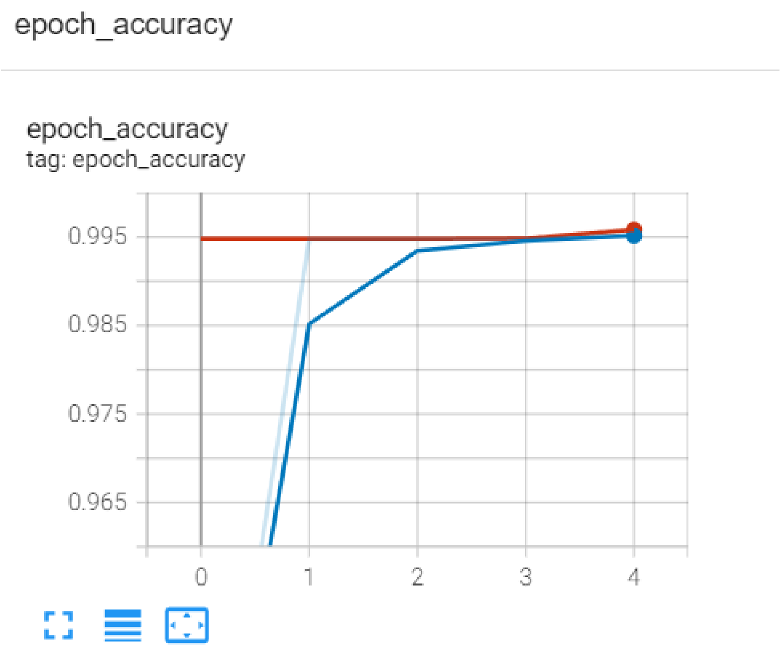


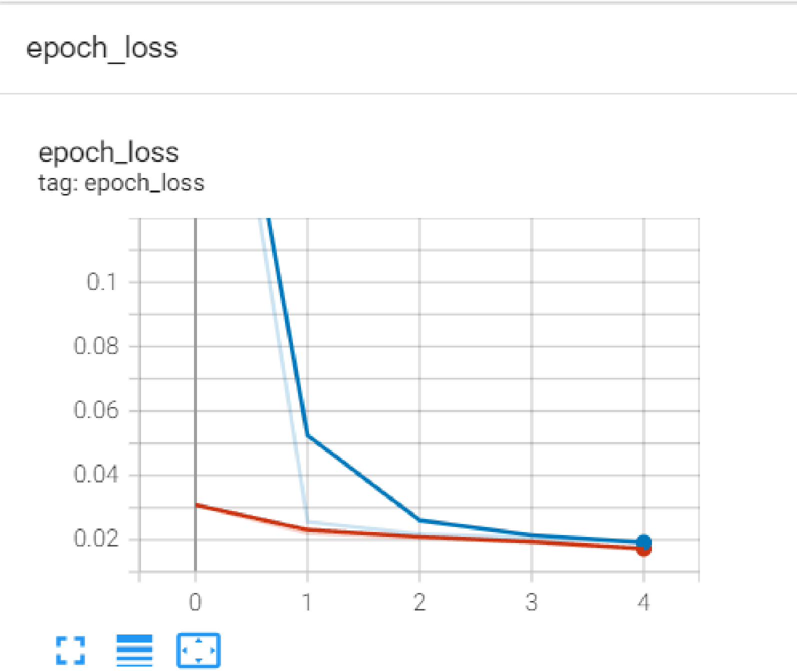


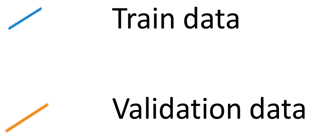
SHAP dependency plot

Model 3 -Neural Network
(2 hidden layer and 1 output)

Parameter	Value
Batch size	1000
Epoch	5

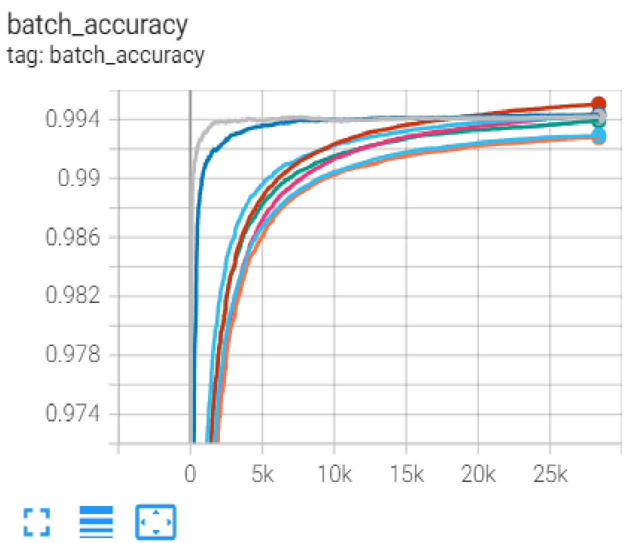


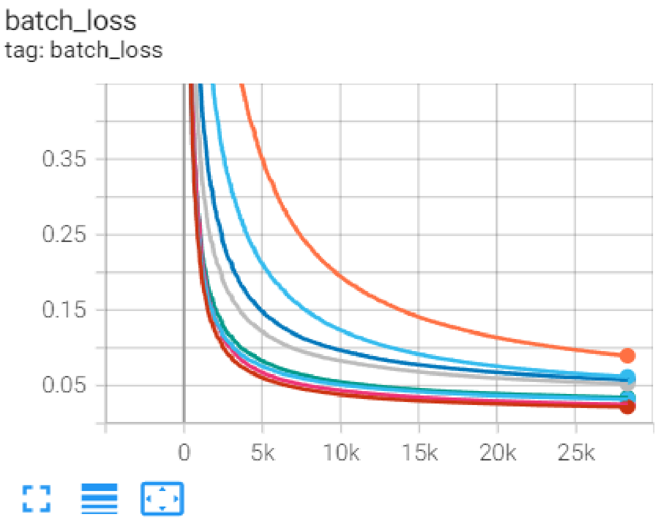




Hyperparameter Tuning

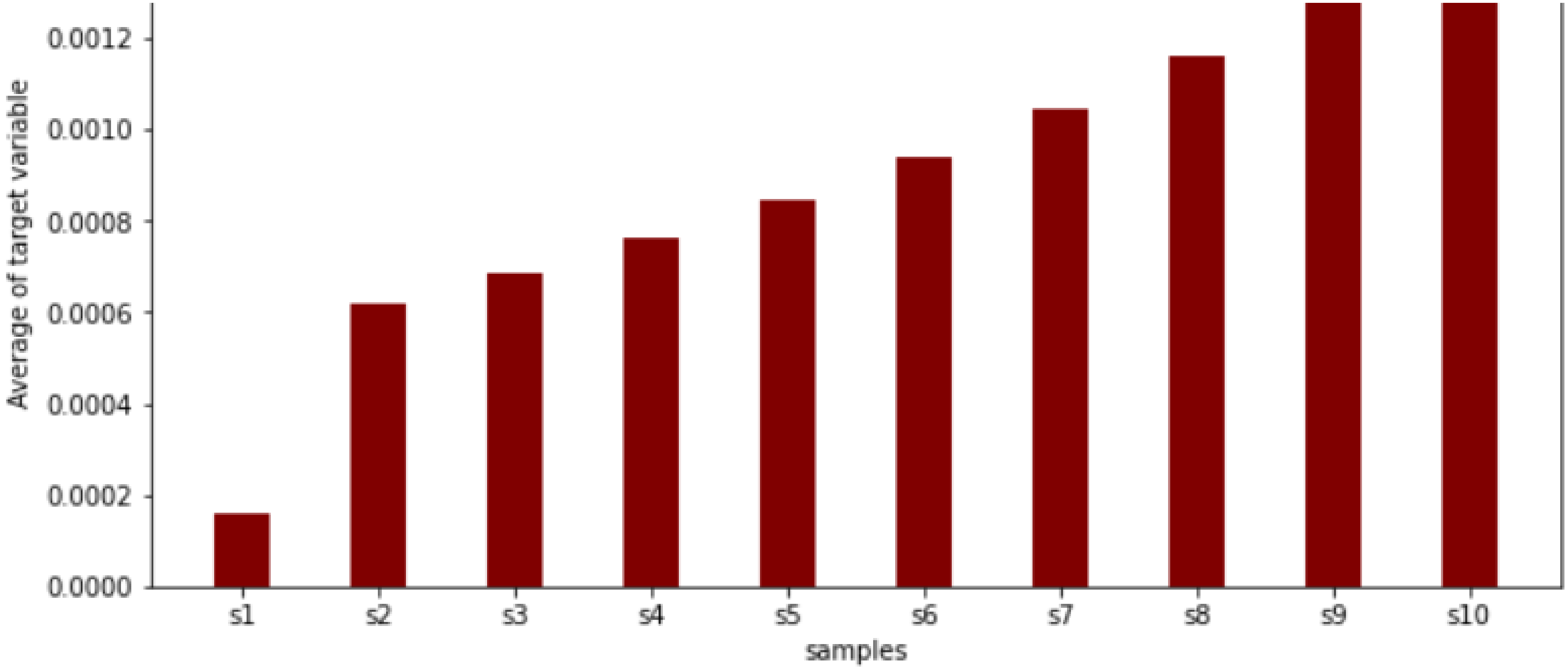
Parameter	Values	
Optimizer	Adam	sgd
Drop out	0.1	0.2
Units	5	6





Model Analysis by K-Fold Validation in test set

Model	Accuracy	Loss
XGBoost	**99.97**	-
Neural Network	99.58	0.018
Logistic regression	92.70	-



Model Analysis – Rank order Analysis with XGBoost prediction