Credit Card Fraud Detection: Capstone Project (BA)

PUJA PRAMOD PATHAK
BATCH - OCT 2020 DS C25
IIITB ROLL NO- DDS20110190

Business Understanding

- **Finex** - leading financial service provider based out of Florida, US. offers a wide range of products and business services through different channels.
- **Finex** has been facing a huge revenue and profitability crisis due to significantly large number of unauthorised transactions.
- **Finex** is not equipped with the latest financial technologies, to track these data breaches.
- The Branch Manager wants to identify the possible root causes and action areas to come up with a long-term solution to prevent losses and maximise profit.

Credit/Debit Card Fraud

Credit/Debit card fraud is any dishonest act or behaviour to obtain information without the proper authorisation of the account holder for financial gain.

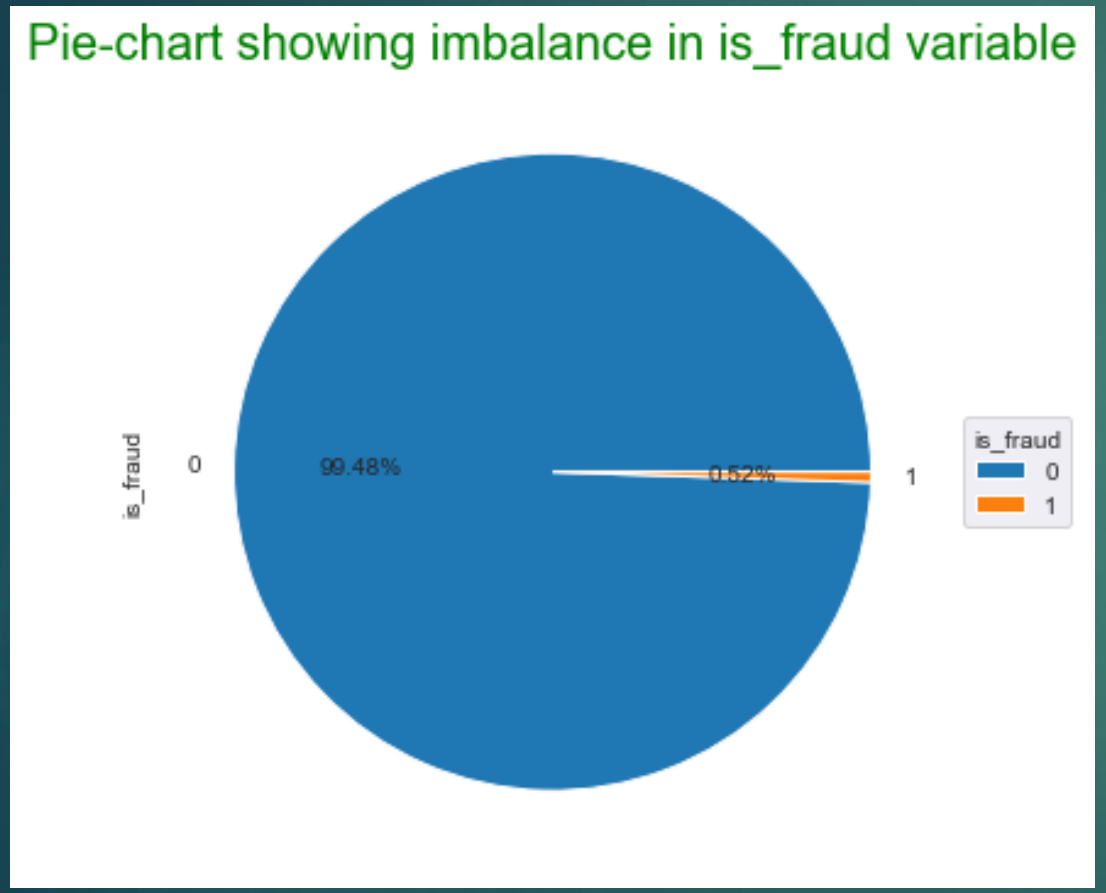
Various ways of committing this fraud are:

- Skimming
- Manipulation or alteration of genuine cards
- Creation of counterfeit cards
- Stolen or lost credit cards
- Fraudulent telemarketing

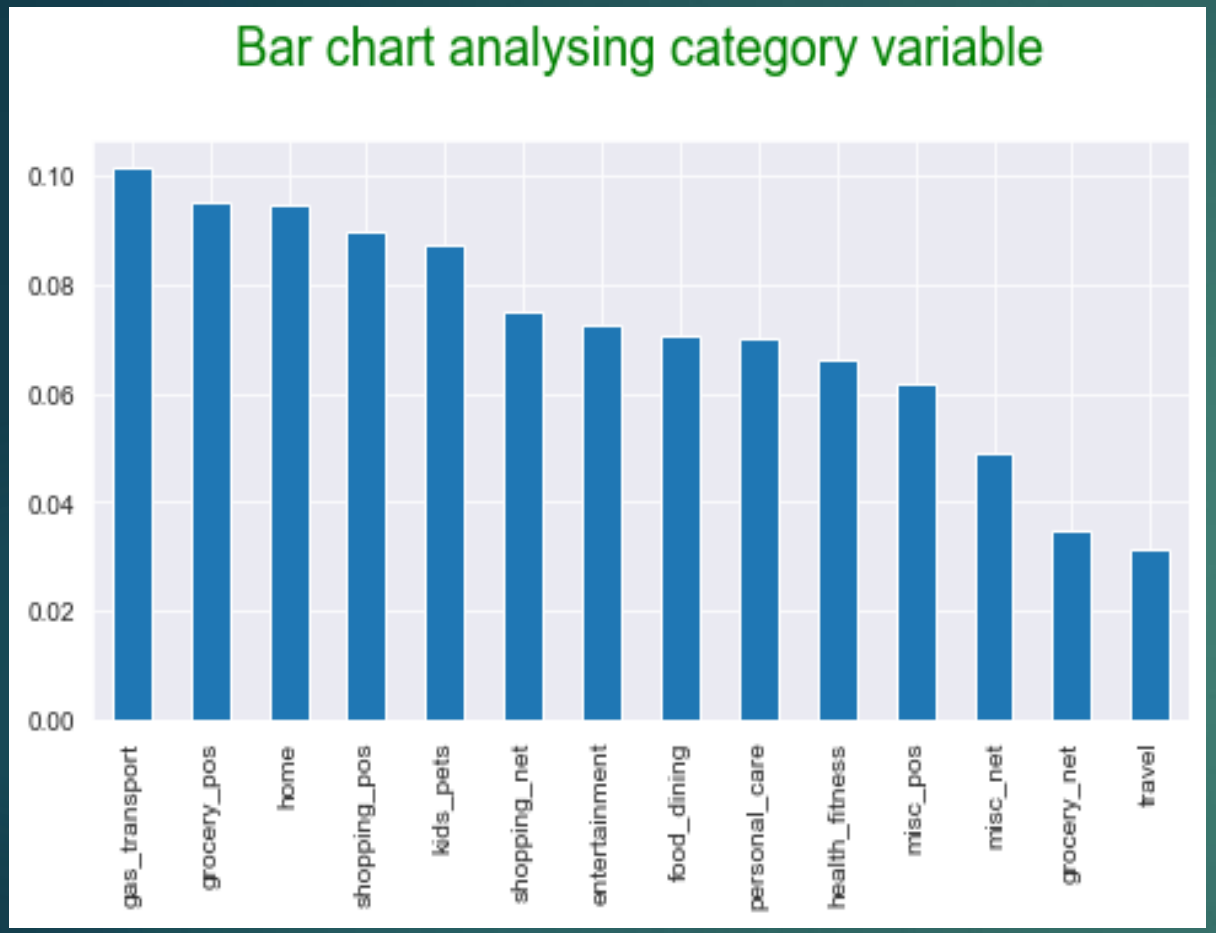
Aim

To develop a Fraud Detection System using a machine learning model to detect fraudulent transactions based on the historical transactional data of customers with a pool of merchants.

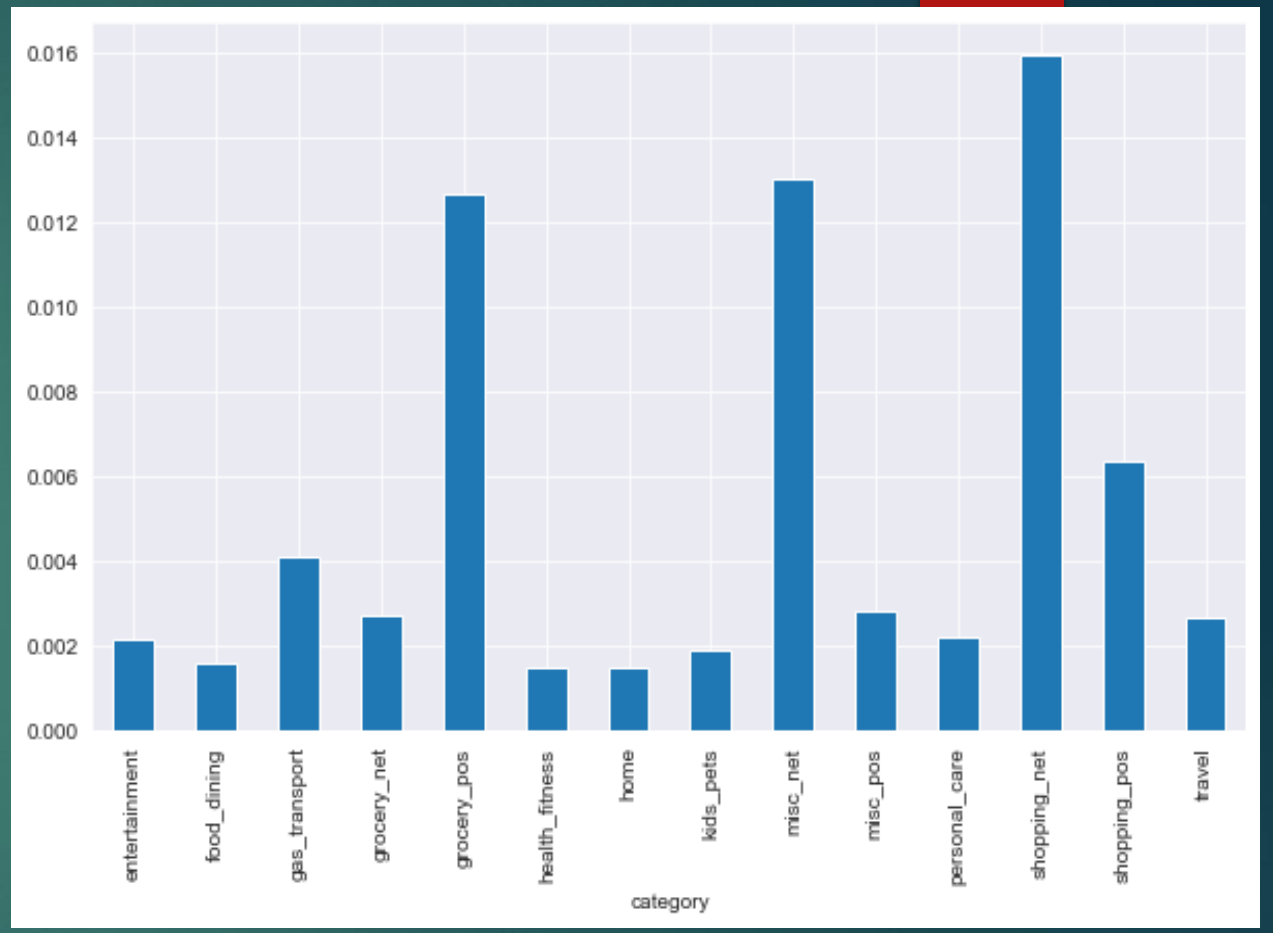
Exploratory Data Analysis



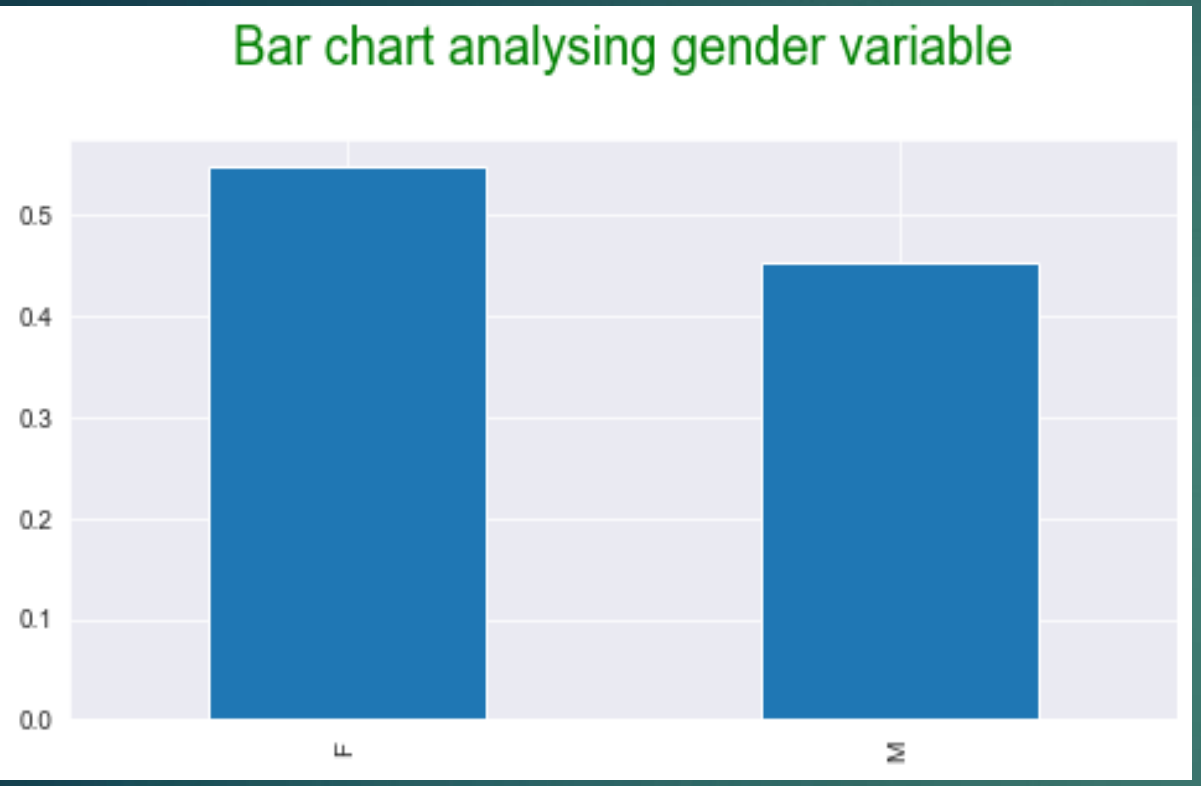
- The historical data is highly imbalanced.
- 99.48% transactions are non-fraudulent (represented by 0)
- Only 0.52% transactions are fraudulent (represented by 1)

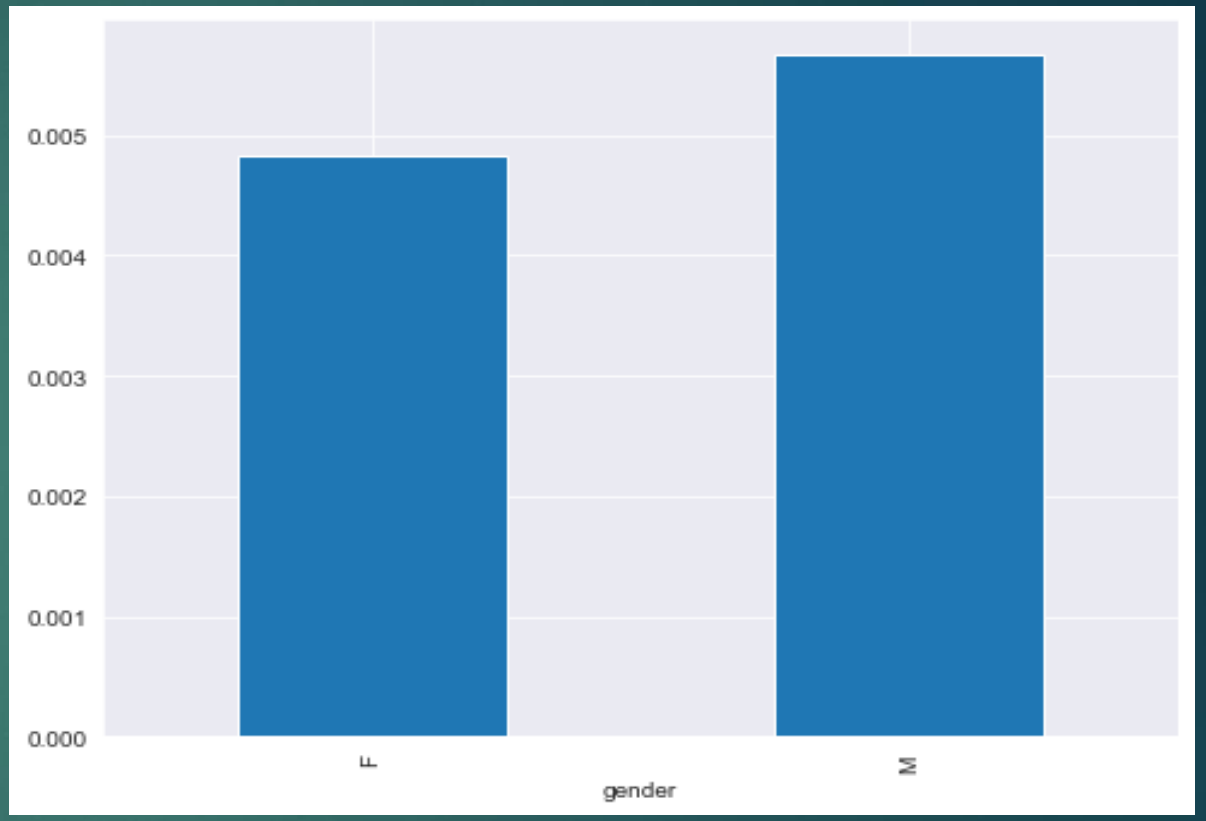


Maximum transactions are performed in the category gas_transport.



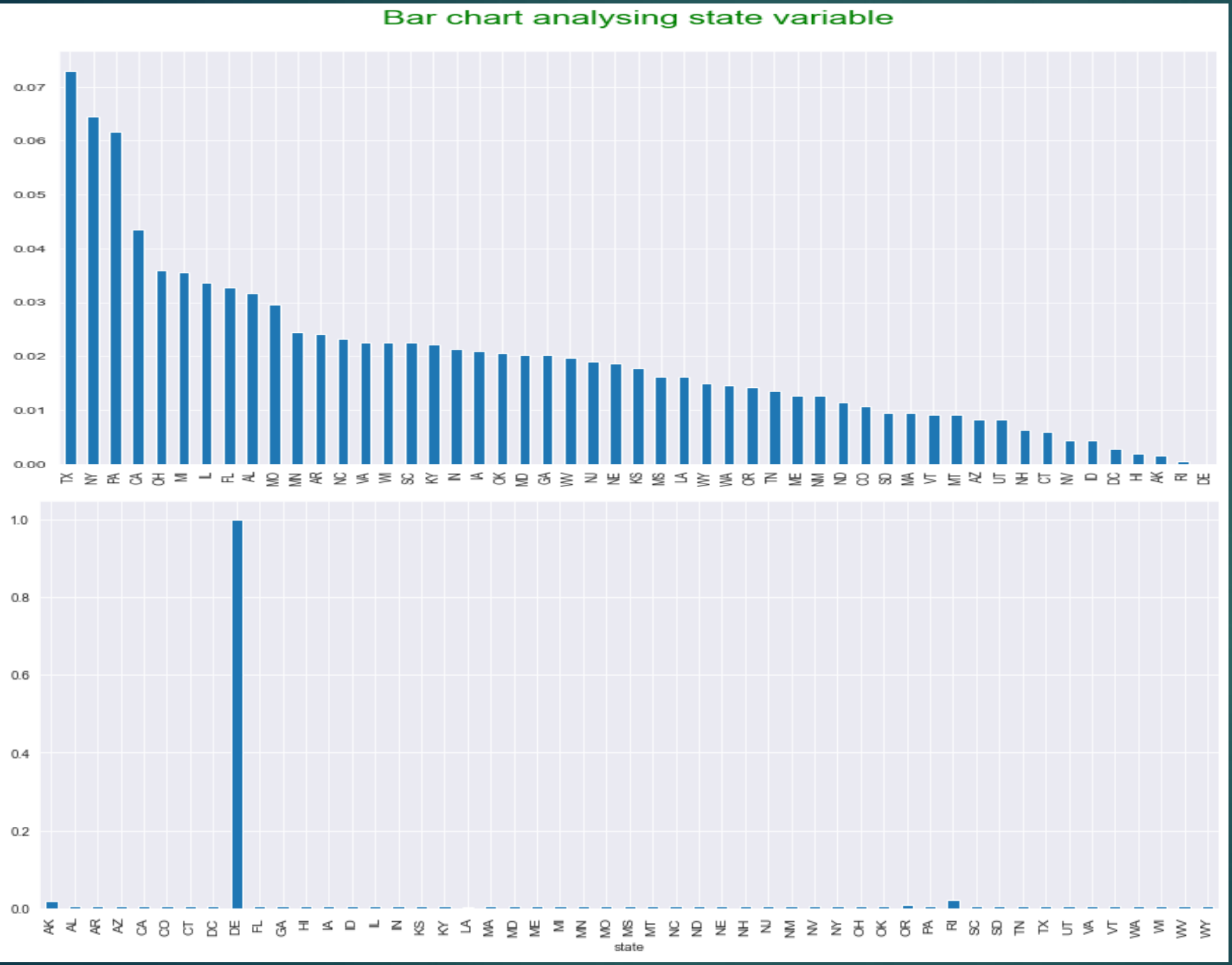
Maximum fraud transactions happened in shopping_net category.



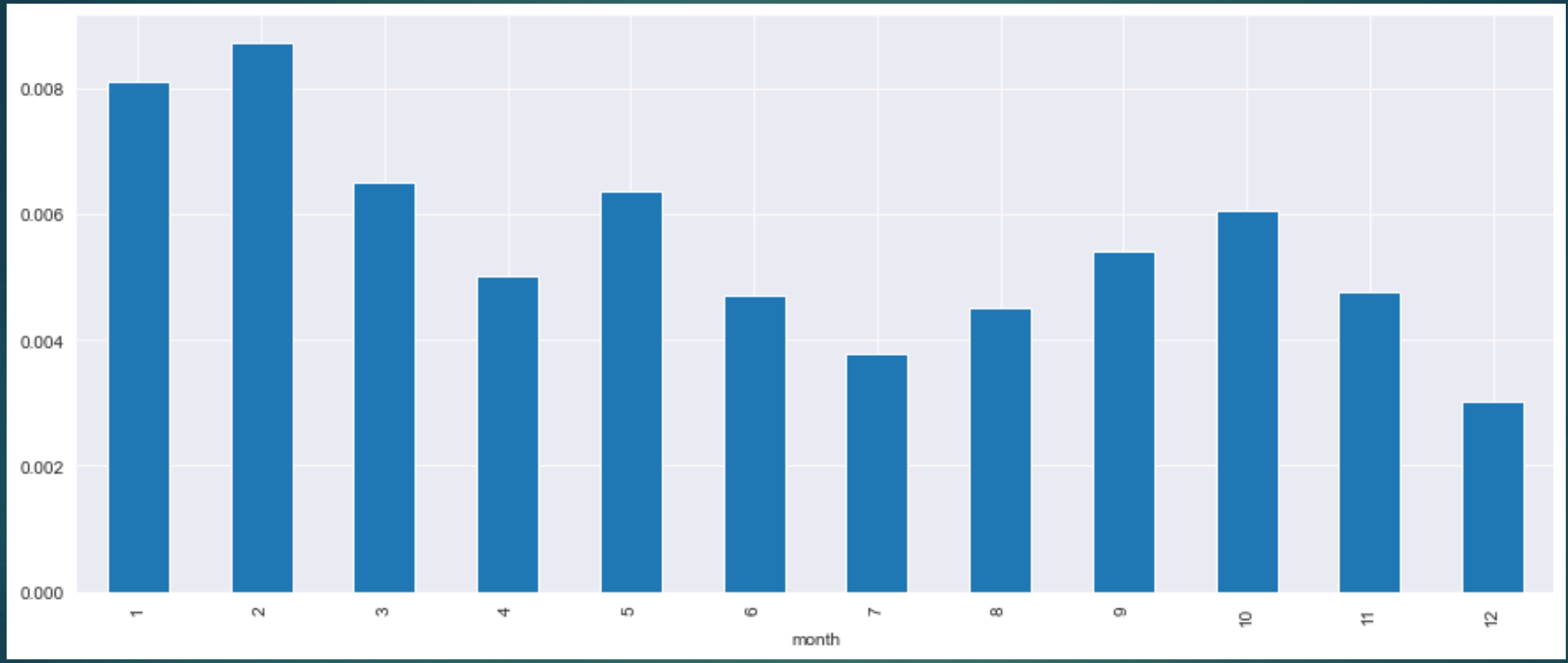


Females have performed more transactions than males.

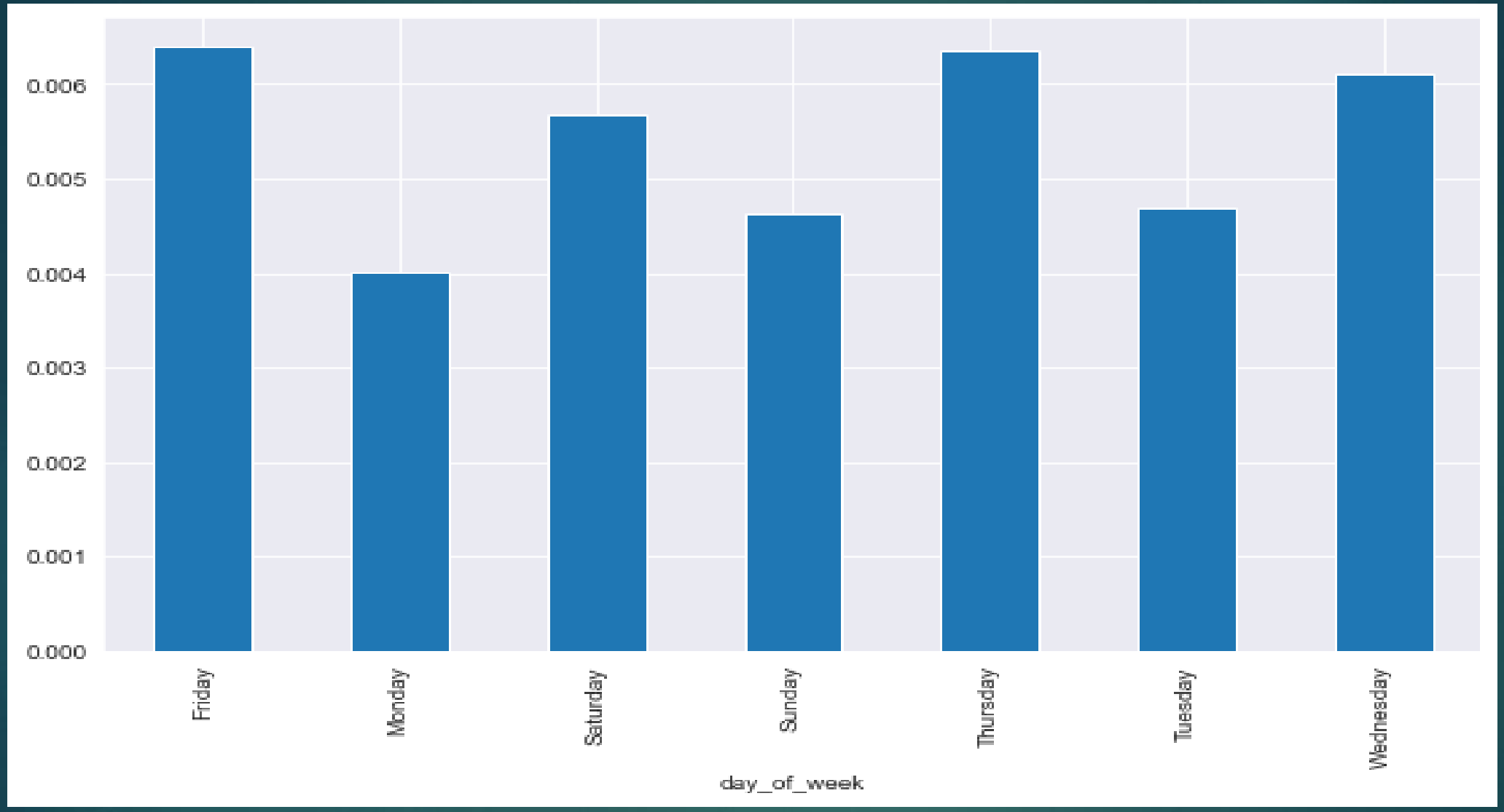
Males have performed more fraud transactions than females.



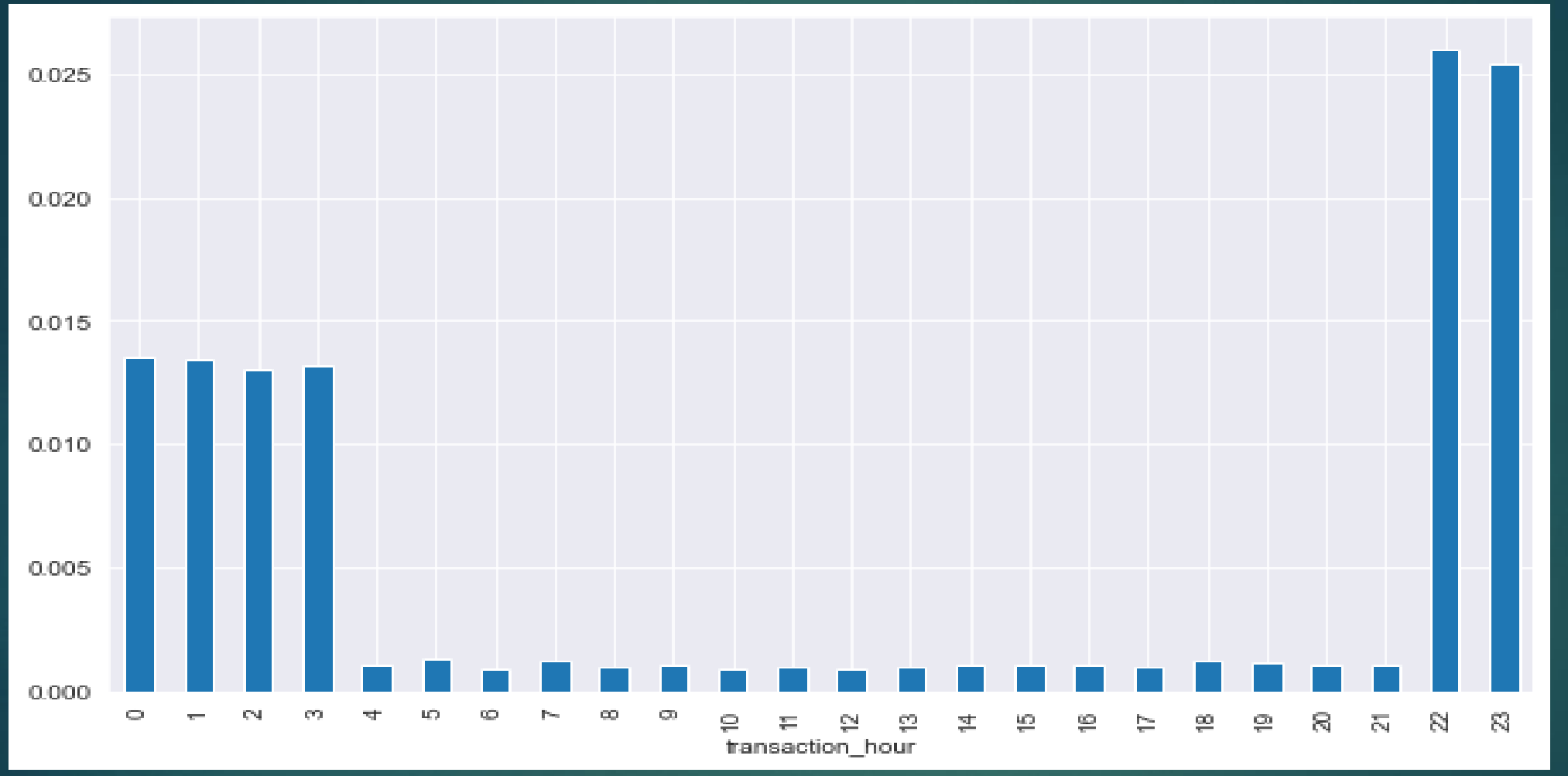
Although least transactions happened in Delaware state, those transactions are fraudulent.



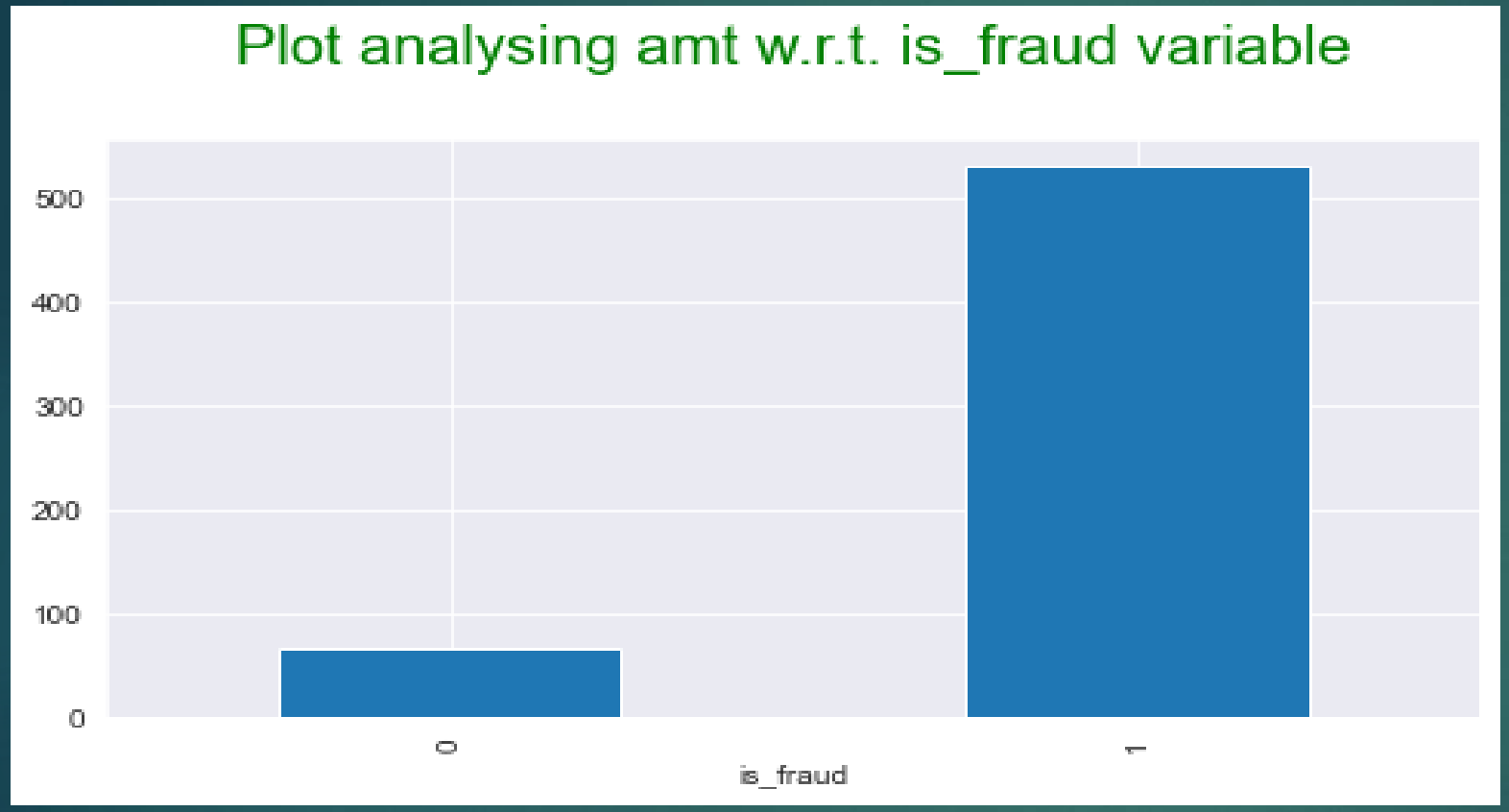
Maximum fraudulent transactions are performed in the month of February.



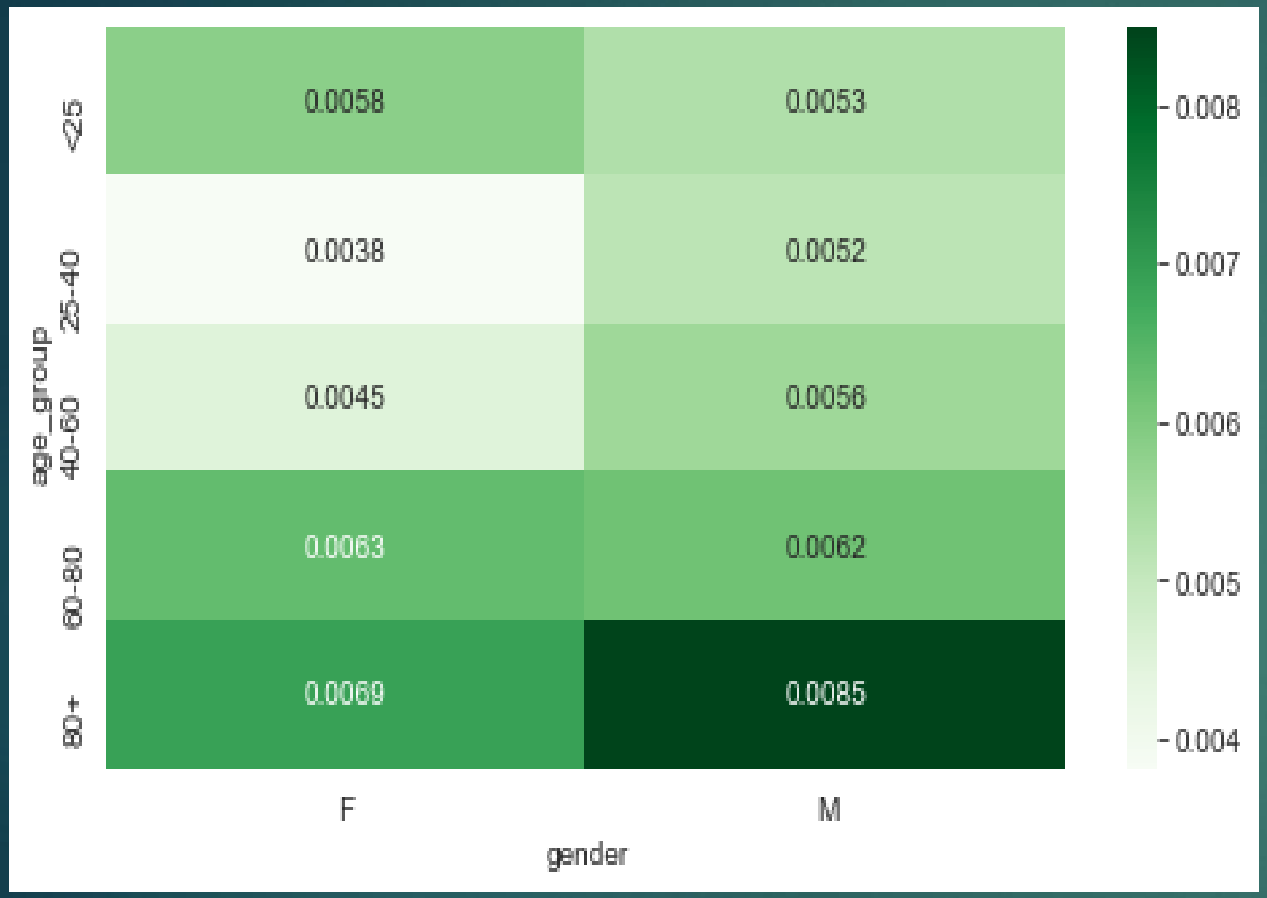
Maximum fraudulent transactions take place on a Friday and a Thursday

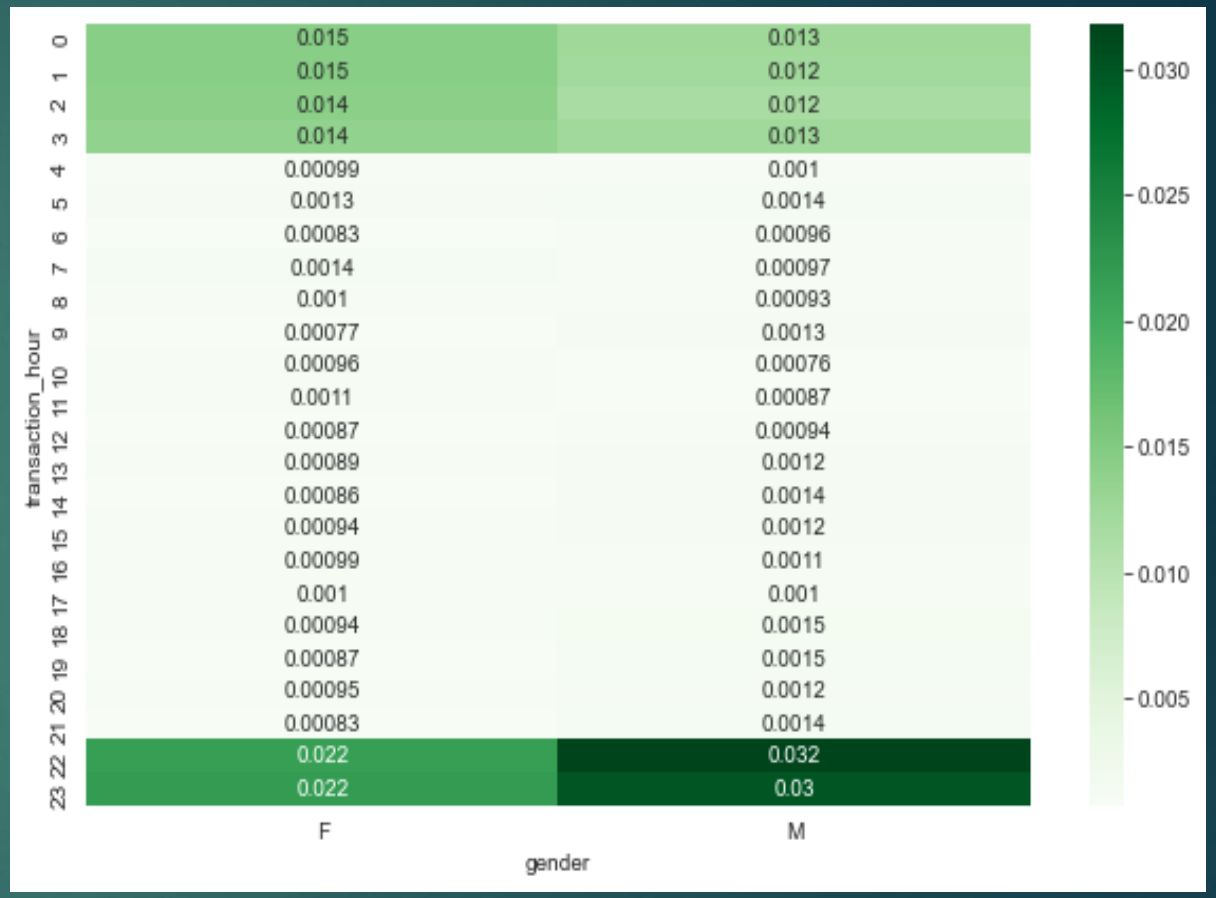


Maximum fraudulent transactions take place during ODD HOURS that is late at night.



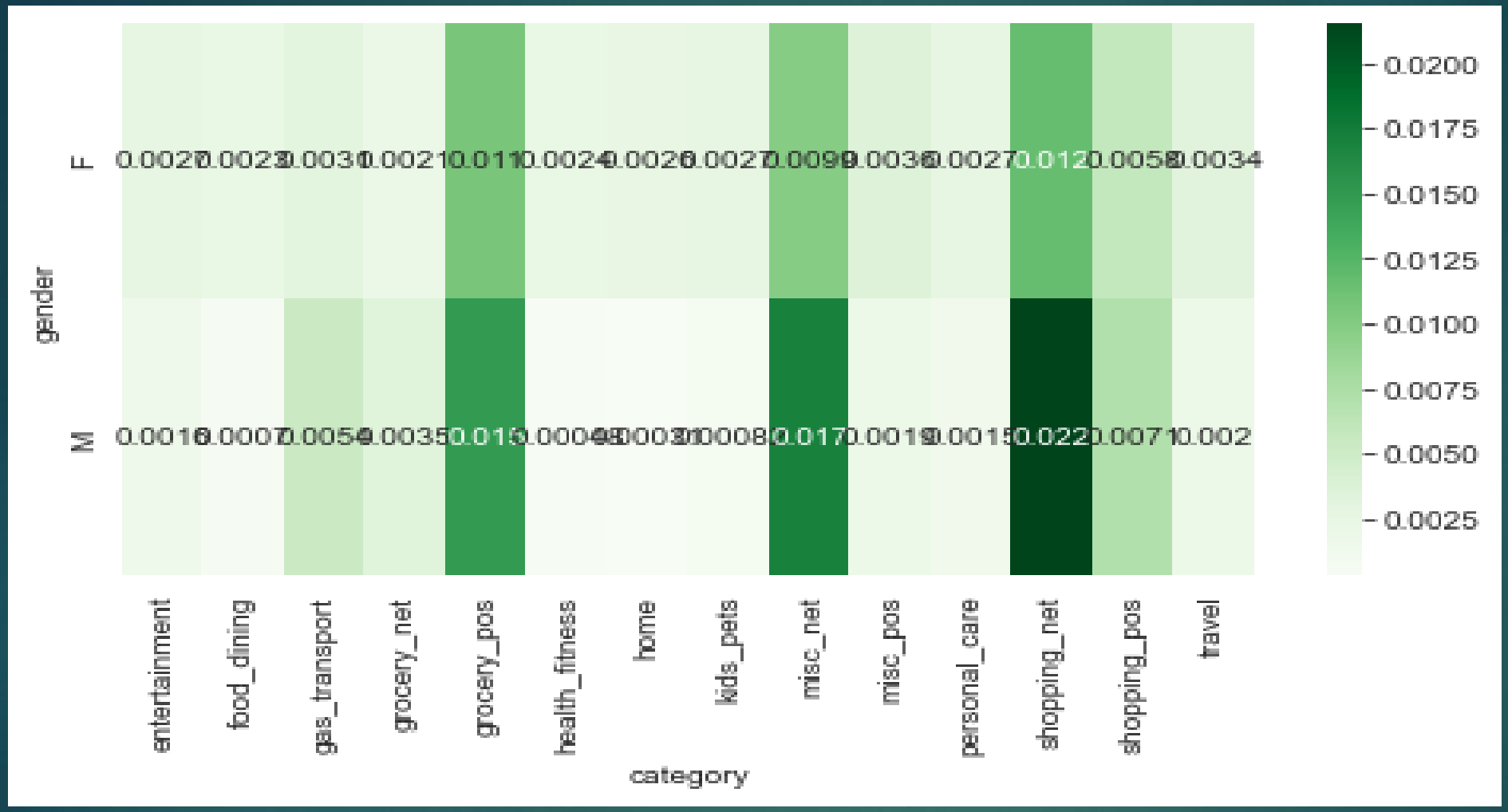
Average amount of a FRAUD transaction is more than 500 dollars.





Transactions performed through cards issued to 80+ years males are fraudulent.

Transactions performed at odd hours by male customers are fraudulent



Most fraudulent transactions are performed by males in the category shopping_net

Model Building

- Out of 18,52,394 transactions, 9,651 transactions are fraudulent.
- Data is split into train and test sets such that each set has equal percentage of fraudulent transactions.
- Skewness in data is removed using appropriate transformation techniques.

Model Building

In order to handle class imbalance in data, following sampling techniques are used:

- Random Over-Sampling
- SMOTE - Synthetic Minority Oversampling Technique
- ADASYN - Adaptive Synthetic Sampling Method

Model Building

Following 4 types of models were built using each of the 3 sampling techniques :

- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost

Model Building

Following are the results and comparison of various evaluation metrics for each of the models built.

	Logistic Regression				Decision Tree				Random Forest				XGBOOST			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
Base	0.99	0.84	0.28	0.42	0.99	0.67	0.68	0.67	0.99	0.75	0.68	0.719				
Random Oversampling	0.76	0.018	0.83	0.035	0.99	0.647	0.675	0.66	0.99	0.67	0.7	0.68	0.97	0.179	0.91	0.3
SMOTE	0.77	0.018	0.827	0.036	0.95	0.089	0.84	0.16	0.95	0.088	0.834	0.159	0.96	0.119	0.92	0.21
ADASYN	0.84	0.022	0.69	0.044	0.93	0.06	0.86	0.12	0.93	0.063	0.847	0.11	0.92	0.059	0.94	0.112

Model Building

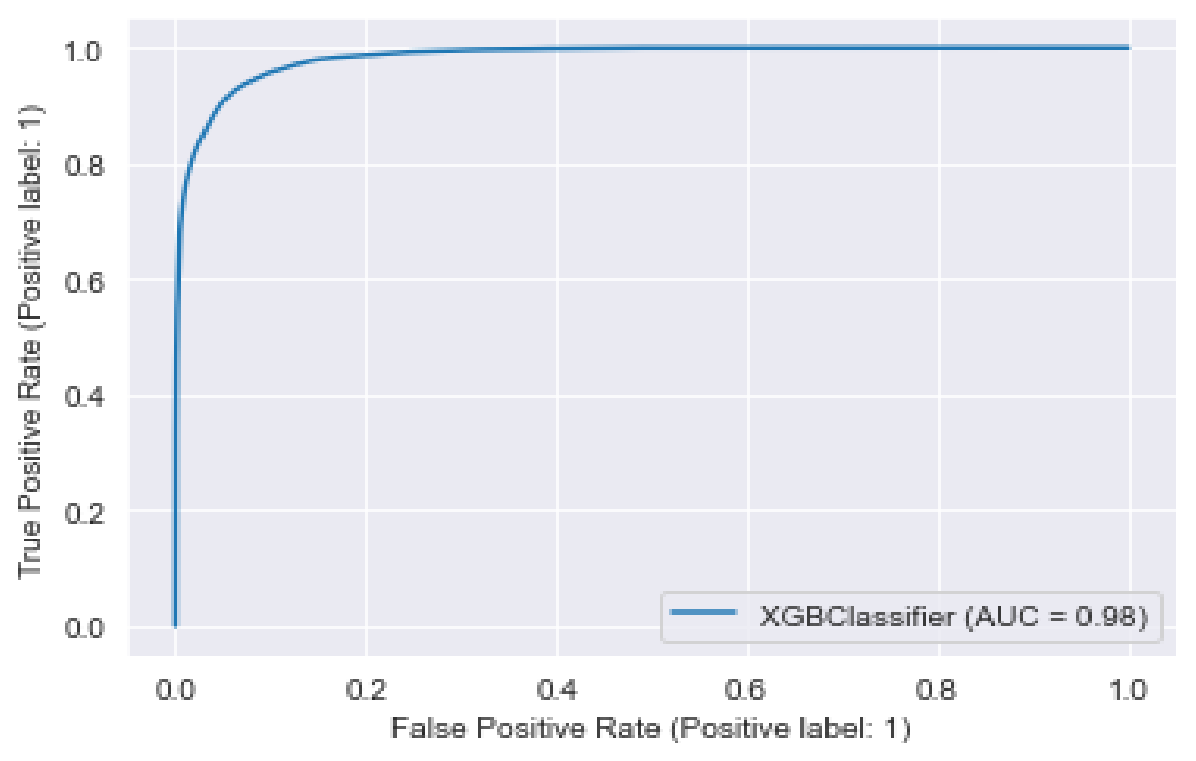
- We need to fine tune the model that gives us the Best combination of ACCURACY and RECALL.

- Following 4 models were selected :
  - Logistic Regression SMOTE model
  - Decision Tree SMOTE model
  - XGBoost ADASYN model
  - Random Forest ADASYN model

- Hyperparameter Tuning - Cross Validation technique is used to extract 10 most important features that are better predictors of fraud.

Model Building

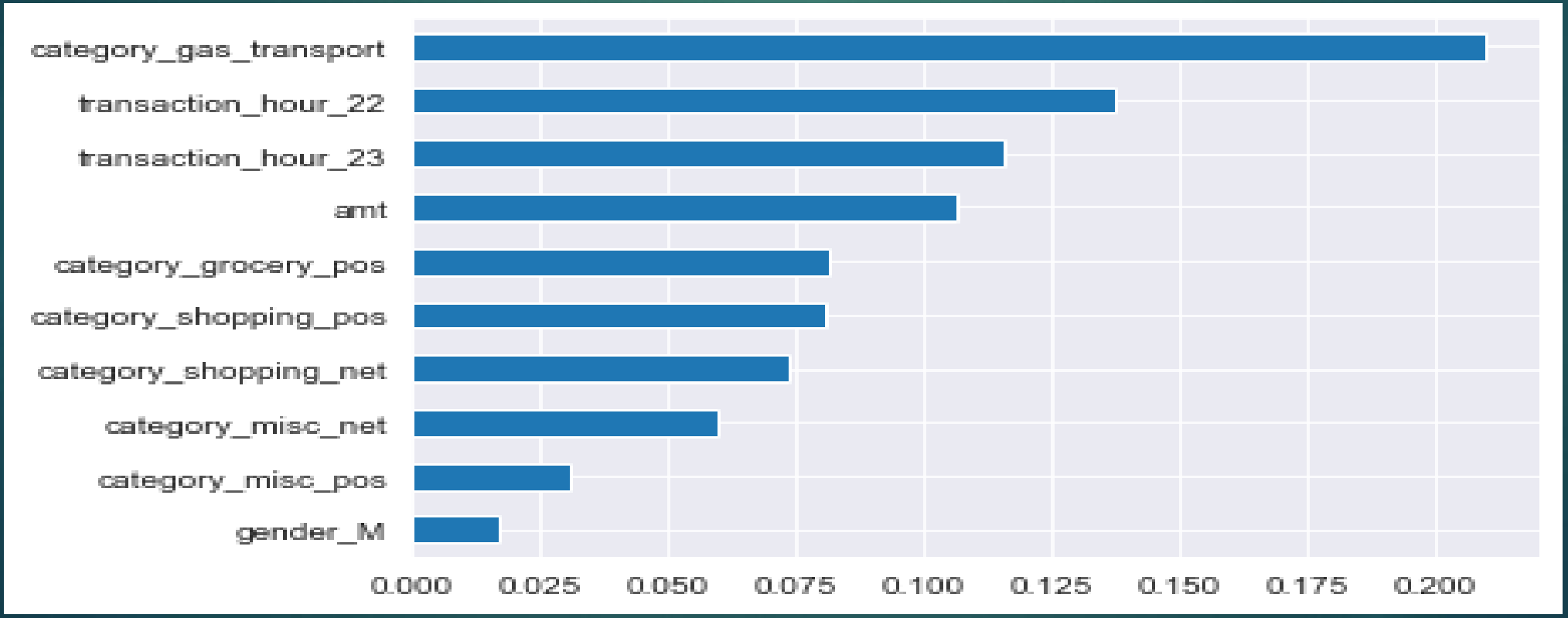
Hyperparameter Tuning using Cross Validation for XGBoost ADASYN model gives best results.




```
Accuracy:  0.9245949841556614
F1 score:  0.11437991377124018
Recall:  0.9347150259067357
Precision:  0.06091713378807321

 clasification report:
                precision    recall  f1-score   support

           0       1.00      0.92      0.96    552824
           1       0.06      0.93      0.11      2895

    accuracy                           0.92    555719
   macro avg       0.53      0.93      0.54    555719
weighted avg       0.99      0.92      0.96    555719


 confussion matrix:
 [[511109  41715]
 [   189   2706]]
```

Model Building

Top 10 most important features predicting fraud :



Cost Benefit Analysis

- If this model is deployed and if a payment gets flagged by the model, an SMS will be sent to the customer requesting them to call on a toll-free number to confirm the authenticity of the transaction.
- A customer experience executive will also be made available to respond to any queries if necessary.
- Developing this service would cost the bank $1.5 per fraudulent transaction.

Cost Benefit Analysis

- Cost before model deployment : $213391.6
- Cost after model deployment : $103071
- **FINAL SAVINGS : $110320.6**

Thank You