Module 7 Assignment: Credit Card Fraud Detection

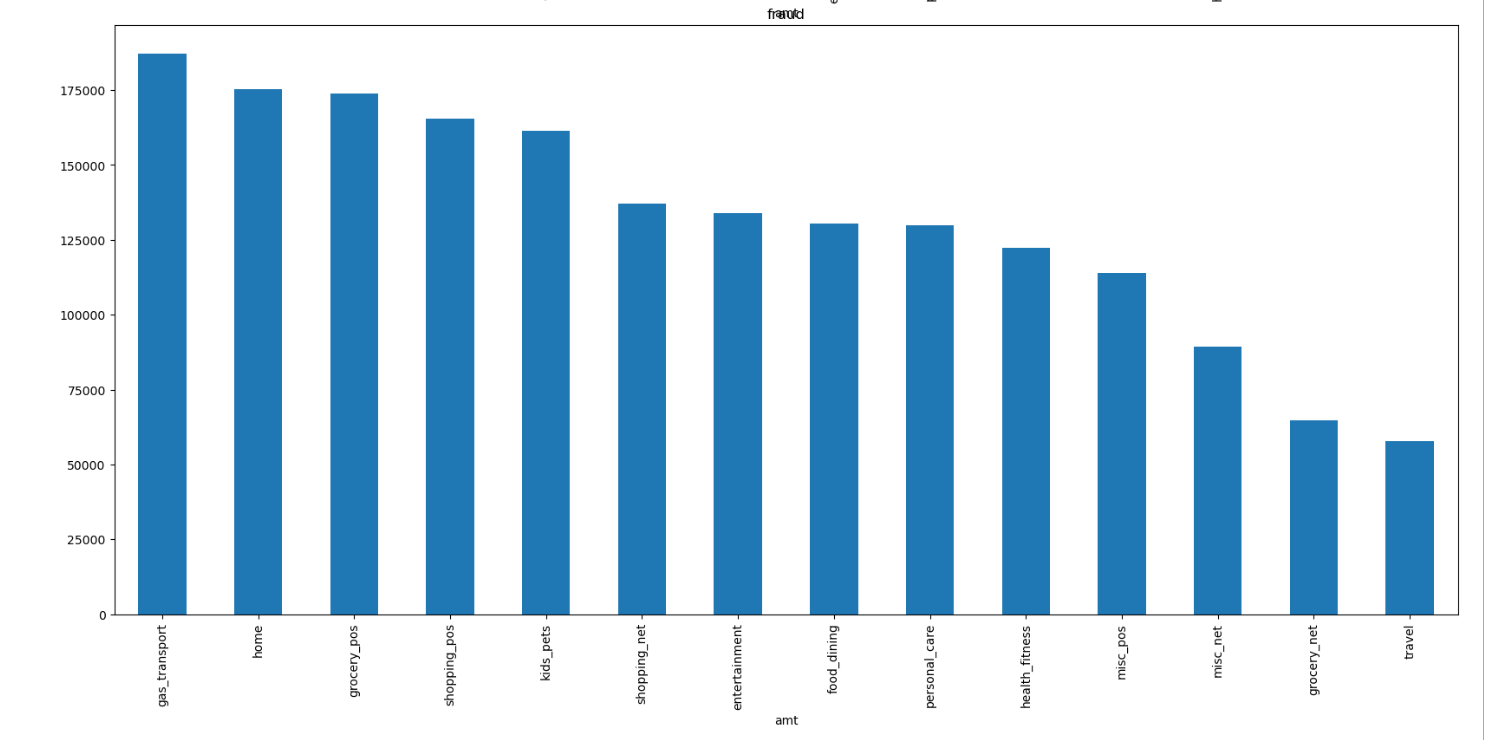
Business Problem: The consulting team’s approach addressing Broom Solutions inaccuracy at identifying fraud transactions is to design, test and build a predictive model whose accuracy and performance will be higher than the current model being used by the Company to ensure transactions are not flagged as a false positive and lead to loss of customers and their loyalty.

**METHODOLOGY**

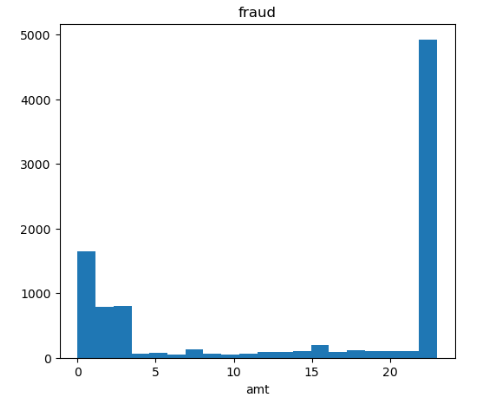
1. The given data was split into two parts – Train and Test. However, for the purpose of ensuring the randomness and accuracy of the model, the two were combined into one dataset and read through pandas a dataframe.
2. The data was then further inspected to check for null values and how many rows and columns. The data set has 1852394 rows and 22 columns. The basic descriptive stats of numerical values 'amt', 'lat', 'long', 'city\_pop','unix\_time', 'merch\_lat', 'merch\_long' were calculated. It can be seen that count is same across all, amount has a mean of 70.064 and the average city population is 88643.675.
3. The categorical variables were further described using histogram visuals, splitting the transaction time into year, day and time to help with understanding the cases of fraud against time. The age of the customers was extracted using their date of birth.
4. **EXPLORATORY DATA ANALYSIS**
   1. Based on external research and the data given, the below variables were identified as the business drivers:

|  |  |
| --- | --- |
| **INTERNAL FACTORS** | **EXTERNAL FACTORS** |
| Store security practices | Customer’s security practices |
| Use authorization techniques like ID, CVV | Population |
| Locations, Operations | Location |
| Staff, merchant | Software used |
| Software used | Customer Age |
|  | Time of transactions |

* 1. In order to avoid fraudulent transactions, the retailer would have to incorporate good security measures especially with increase in digital transactions, the gateways used must ensure that authorization of the customer is done accurately using their ID and card details.
  2. On the other hand, while plotting the category of fraud against the transactions which are fraud and not, it can be seen that a huge contribution to these are from transactions done for gas transportation, grocery shopping, shopping and from home. This might be due to ‘Skimming’ which involves reading card details from gas stations through bogus card payment terminals and using those details to make online purchases without the customer’s knowledge. The next highest category is ‘home’ which might mean real estate. Since real estate deals with huge sums of money, transactions are easily diverted to fraudulent accounts. The next few categories are related to grocery and shopping which includes the use of fake identification, fraudulent or stolen credit cards, counterfeit funds obtained through counterfeit money, altered checks, counterfeit coupons, and phony travelers/gift checks as well as return receipt fraud. These all affect the retailers drastically.



* 1. A histogram plot of the hour of transactions against fraud shows that most of the fraudulent transactions occur early morning or late night, majorly as banks are closed during these hours which make it easier to perform the fraud while the reporting of the fraud takes time as police staff and cybersecurity staff will be very less leading to low security.



* 1. A look at the gender classification of the population shows that there are more female customers and around 2.6% of them contribute to the fraud transactions, however, it’s very close to the stats shown by the male population.

1. **MODELLING:**
   1. A correlation between the parameters and fraud shows that transaction amount is the most highly correlated with fraud while the remaining parameters don’t have a strong enough relationship.
   2. As the data set is highly imbalanced, the data needs to be oversampled to ensure balancing.
   3. The data is then split into sets for training and testing the models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Logistic Regression** | **Decision Tree** | **Random Forest** | **XGBoost** |
| **F-measure** | 85% | 89% | 88% | 95% |
| **AUC** | 0.9256 | 0.9447 | 0.955 | 0.9447 |
| **Most important feature** | category\_shopping\_pos | transaction amount | transaction amount | transaction amount |
| **Second important feature** | category\_shopping\_net | transaction hour | transaction hour | category\_gas\_transport |
| **Third important feature** | category\_gas\_transport | category\_travel | 60d | transaction hour |
| **Model chosen** |  |  | Yes |  |

1. **INSIGHTS:**
   1. Based on the above, the best model that is predicting accurately is the random forest model with an accuracy of ~95% and logistic regression showed the least accuracy of ~92%.
   2. The prediction shown by majority of the models is that transaction amount is an important feature that causes fraud. This may mean that a check on high valued transactions should be flagged for both the consumer and the retailer. On the consumer side, the customer could be given a phone call for a verification to be done on call before the transaction is processed. On the retailer’s side, the customer can be verified and checked with to ensure that they verified the transaction.
   3. Since transaction hour is significantly important, if a transaction is to be processed at late night hours or early morning hours, a team could be outsourced to monitor unusual transactions. With the help of the model, profiling can be done to check which customers or merchants are more prone to fraud. More monitoring and security measures can then be applied there.