CREDIT CARD FRAUD DETECTION

-Richa Shah

Agenda

- Objective
- Background
- Key Insights
- Cost Benefit Analysis
- Appendix:
- Data Attributes
- Data Methodology
- Attached Files

Objective

- Getting in place a credit card fraud detection system to save on incurred costs incurred.
- Huge costs are being incurred due to frauds and a manual detection system

Background

- A machine learning model has been built to detect frauds early and avoid risk of losses.
- A cost benefit analysis has been done for the deployment of the same.

Key Insights

- Transaction amount, category and gender are the most important variables
- Gas and transport, grocery and shopping are the top three categories

Current Incurred Losses

- 77,183 credit card transactions per month
- 402 fraudulent transactions per month
- $ 530.66 amount per fraud transaction
- Total costs incurred from fraud transactions is

  $ 213,392.22

After New Model Deployment

- 1720 fraudulent transactions detected by the model
- $ 1.5 cost to provide customer support to these transactions that is $ 2,580.38 in total
- 68 fraudulent transactions not detected by model which amounts to $ 35,908.09 loss
- Total cost incurred after new model deployment is $ 38,488.46
- Final savings after new model deployment is $174,903.76 that is reduction in losses by ~82%

Appendix: Data Attribute

- Snapshot of data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1852394 entries, 0 to 1852393
Data columns (total 22 columns):
 #   Column              Dtype
---  ------              -----
 0   cc_num              int64
 1   merchant            object
 2   category            object
 3   amt                 float64
 4   gender              object
 5   street              object
 6   city                object
 7   state               object
 8   zip                 int64
 9   lat                 float64
 10  long                float64
 11  city_pop            int64
 12  job                 object
 13  trans_num           object
 14  unix_time           int64
 15  merch_lat           float64
 16  merch_long          float64
 17  is_fraud            int64
 18  trans_hour          int64
 19  trans_day_of_week   object
 20  trans_year_month    period[M]
 21  age                 float64
```

Appendix: Data Methodology

- A random forest classifier built on top a Kaggle simulated dataset
- Smote sampling method
- Manual hyperparameter tuning done due to extensive computational times when using Grid Search Cross Validation

Attached Files

- Cost Benefit Analysis:
  - Cost Benefit Analysis
- Random Forest Classifier Model1:
  - CC FRAUD DETECTION