Credit Card Fraud Detection

- Mayank Jha

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 mitigate 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

	Varname	Imp
0	amt	6.741811e-01
8	category_gas_transport	1.293945e-01
10	category_grocery_pos	5.232301e-02
17	category_shopping_net	3.205304e-02
16	category_personal_care	2.711764e-02
13	category_kids_pets	2.518517e-02
15	category_misc_pos	1.750431e-02
1	gender	9.864341e-03
14	category_misc_net	9.285657e-03
19	category_travel	8.042764e-03
7	category_food_dining	4.431430e-03
9	category_grocery_net	4.091956e-03
18	category_shopping_pos	3.208406e-03
2	city_pop	1.781310e-03
12	category_home	1.009814e-03
6	trans_month	4.629258e-04
3	age_at_trans	6.080661e-05
5	long_dist	1.130483e-06
4	lat_dist	6.153246e-07
11	category_health_fitness	0.000000e+00

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 Attributes

- Snapshot of the data:
  - index - Unique Identifier for each row
  - trans*date*trans_time - Transaction DateTime
  - cc_num - Credit Card Number of Customer
  - merchant - Merchant Name
  - category - Category of Merchant
  - amt - Amount of Transaction
  - first - First Name of Credit Card Holder
  - last - Last Name of Credit Card Holder
  - gender - Gender of Credit Card Holder
  - street - Street Address of Credit Card Holder
  - city - City of Credit Card Holder
  - state - State of Credit Card Holder
  - zip - Zip of Credit Card Holder
  - lat - Latitude Location of Credit Card Holder
  - long - Longitude Location of Credit Card Holder
  - city_pop - Credit Card Holder's City Population
  - job - Job of Credit Card Holder
  - dob - Date of Birth of Credit Card Holder
  - trans_num - Transaction Number
  - unix_time - UNIX Time of transaction
  - merch_lat - Latitude Location of Merchant
  - merch_long - Longitude Location of Merchant
  - is_fraud - Fraud Flag <--- Target Class

Appendix: Data Methodology

- A random forest classifier built on top a Kaggle simulated dataset
- Class imbalance adjusted using Adaptive Synthetic (ADASYN) sampling method
- Manual hyperparameter tuning done due to extensive computational times when using Grid Search Cross Validation