Credit Card Fraud Detection system

DONE BY: AASHKA VIJAPURA

Agenda

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

Objective

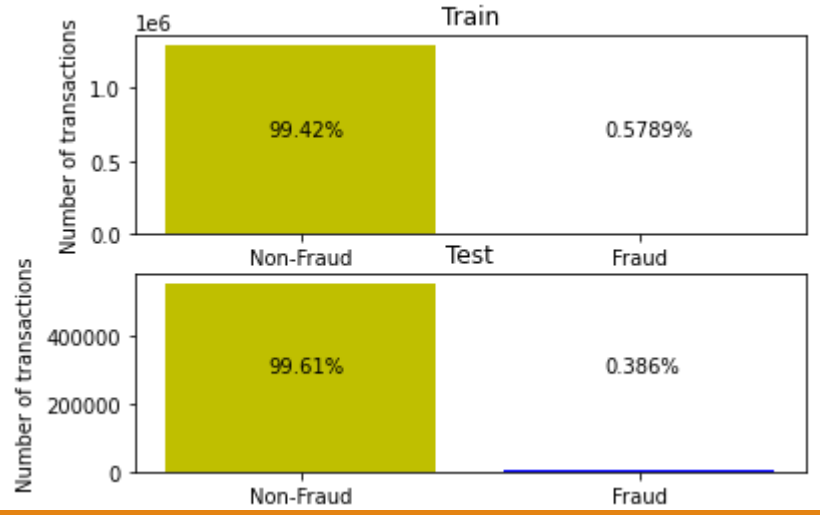
- Putting in place a system to detect credit card theft in order to reduce expenses
- A manual detection method for frauds result in significant costs

Background

- A machine learning algorithm was created to identify frauds quickly and reduce losses.
- For its deployment, a cost-benefit analysis has been conducted.

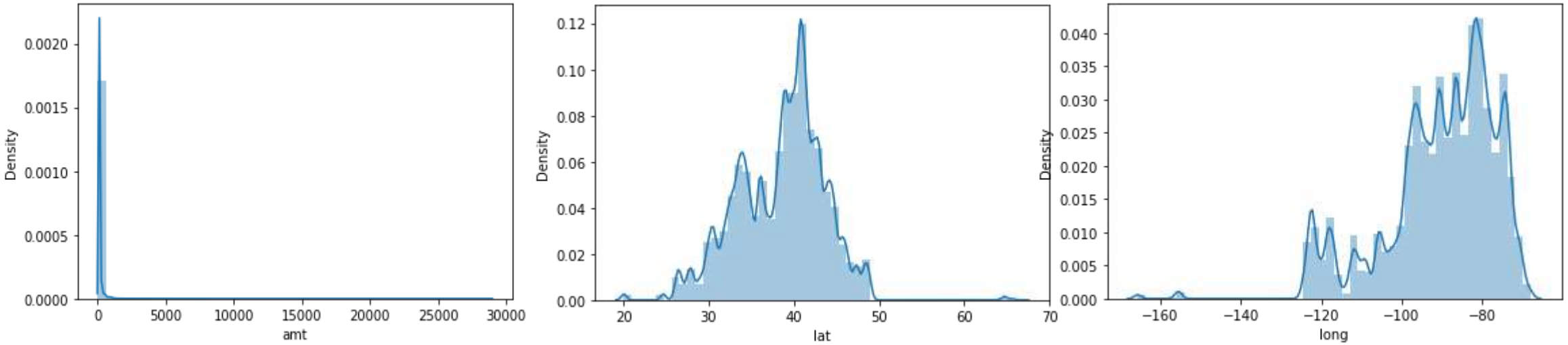
Key Insights

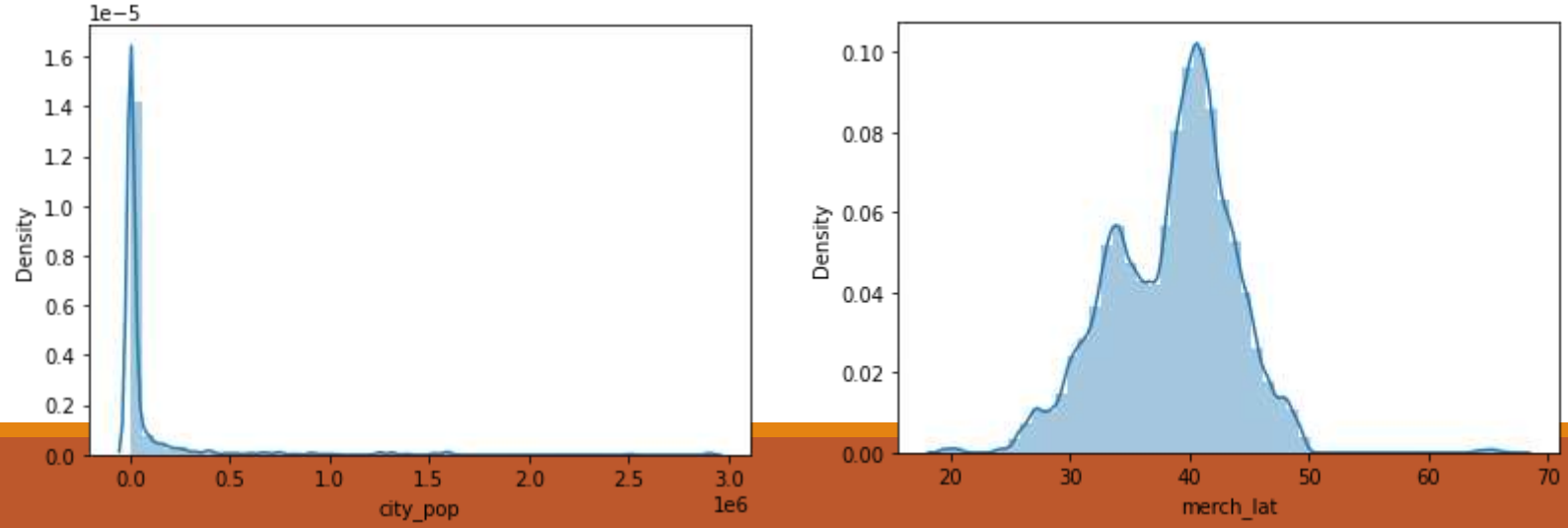
- Transaction amount, category and gender are the most important variables
- Gas and transport, kids and pets, home are the top three categories
- Highly imbalanced dataset



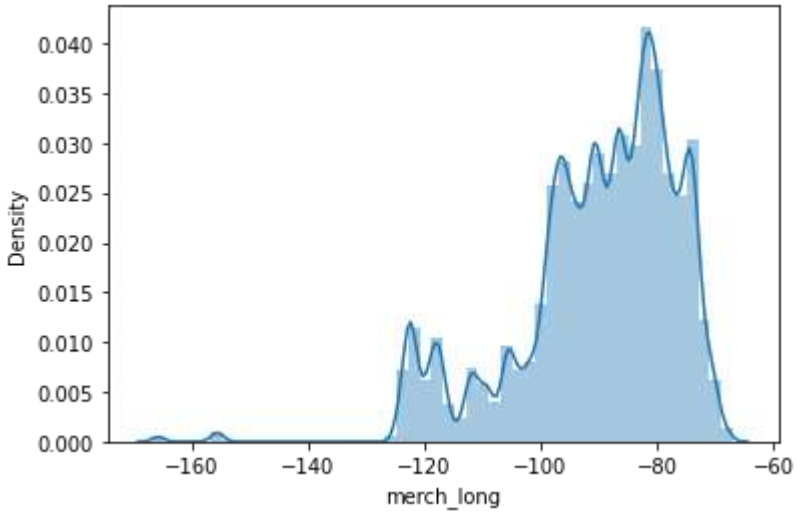
	Varname	Imp
0	amt	0.876727
13	category_kids_pets	0.029171
8	category_gas_transport	0.023109
12	category_home	0.013885
19	category_travel	0.010977
18	category_shopping_pos	0.010669
10	category_grocery_pos	0.009082
15	category_misc_pos	0.008953
7	category_food_dining	0.005615
17	category_shopping_net	0.003883
1	gender	0.002981
3	age_at_trans	0.002070
2	city_pop	0.002070
11	category_health_fitness	0.000412
9	category_grocery_net	0.000194
14	category_misc_net	0.000098
4	lat_dist	0.000092
6	trans_month	0.000013
5	long_dist	0.000000
16	category_personal_care	0.000000

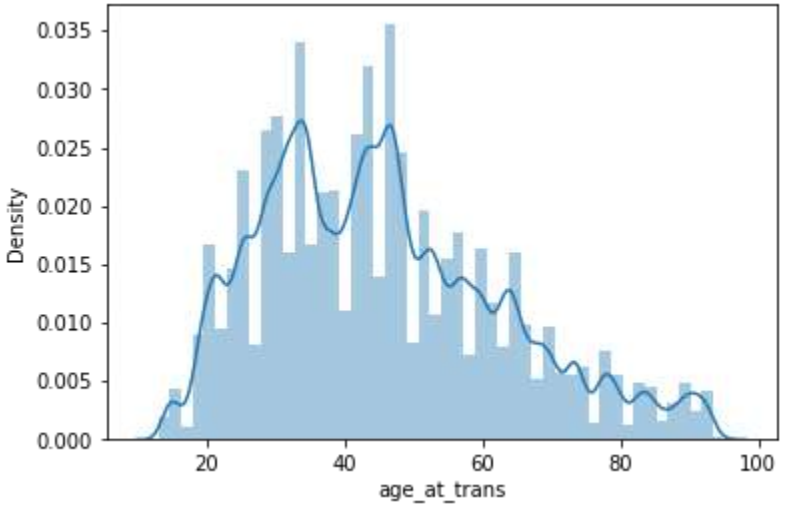
Key Insights

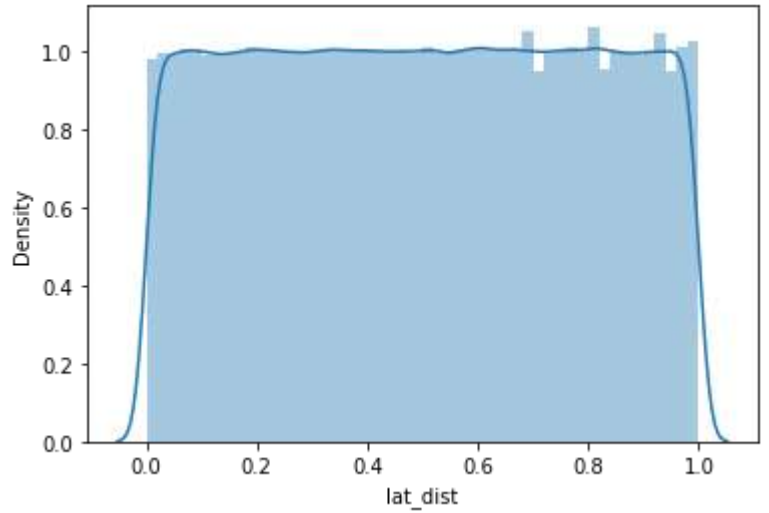


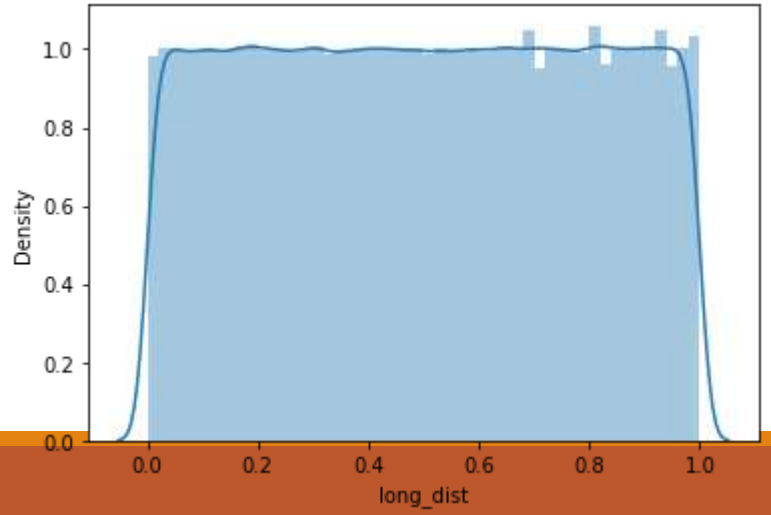


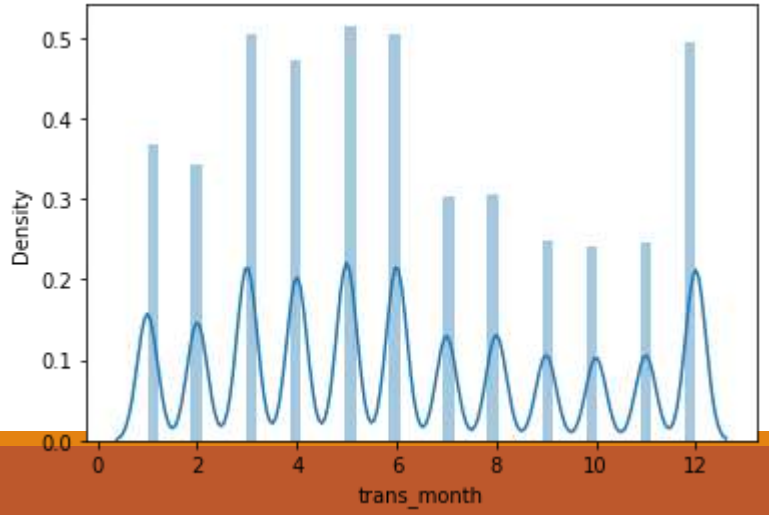
Key Insights



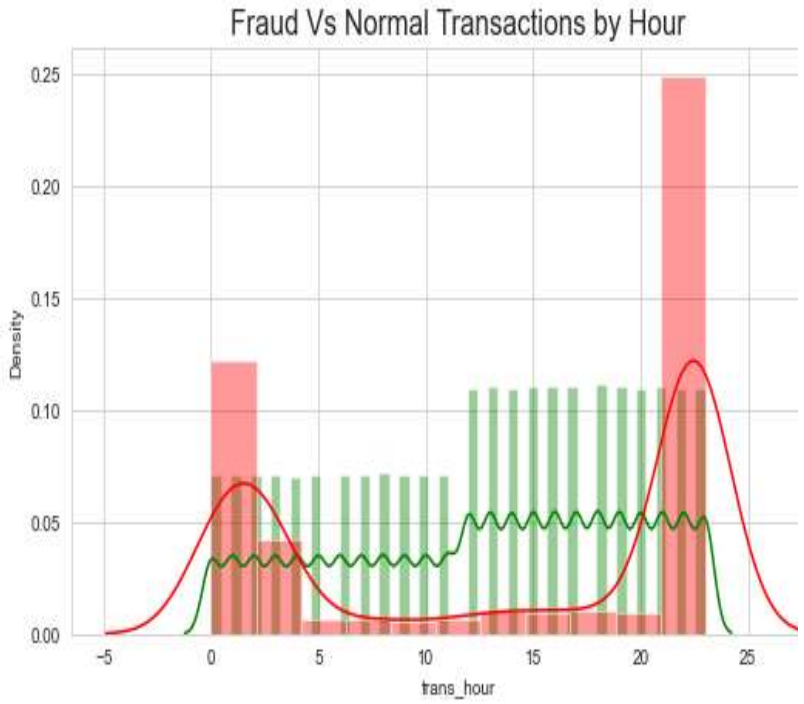


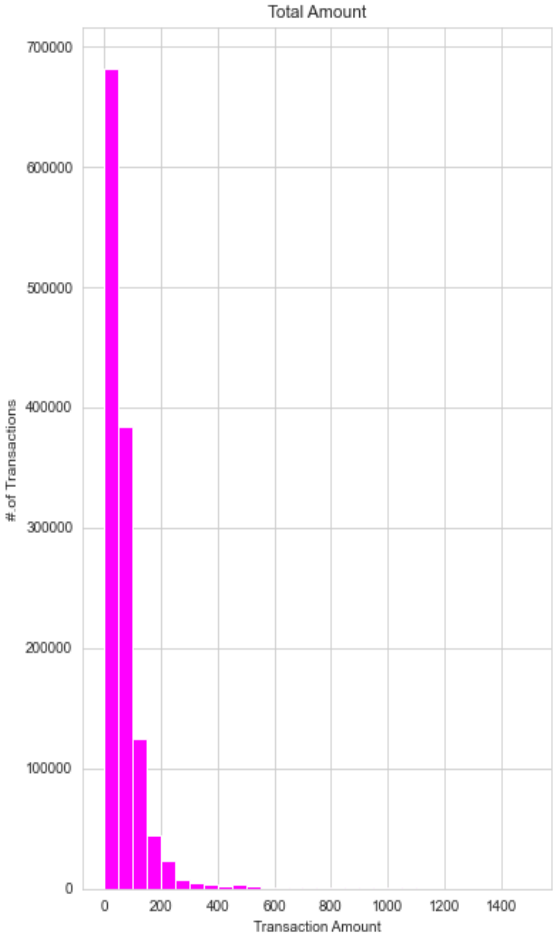


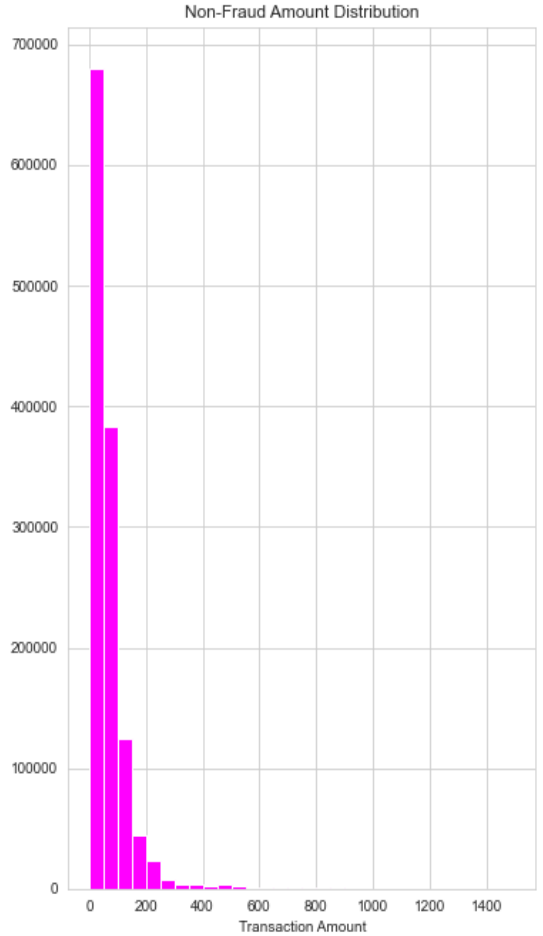


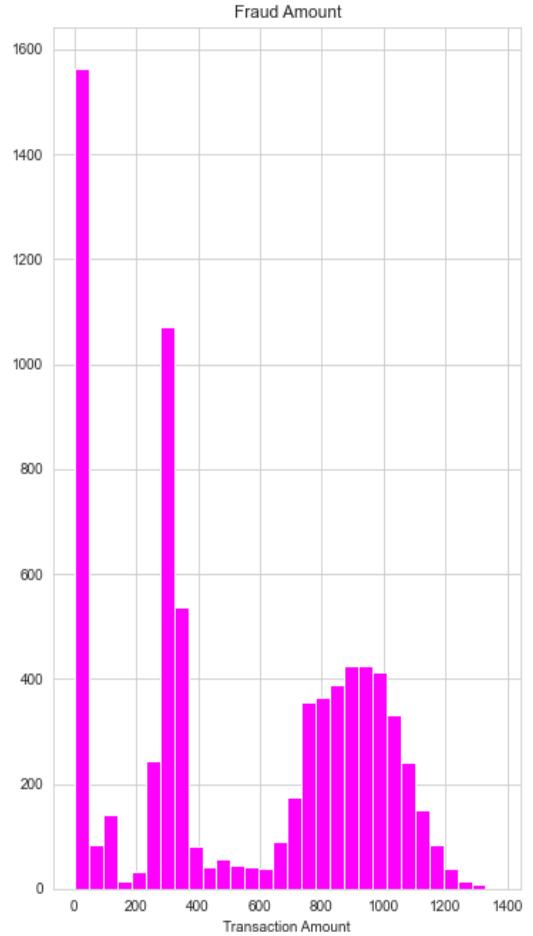


Key Insights









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
- 27 fraudulent transactions not detected by model which amounts to $ 14,283.64 loss
- Total cost incurred after new model deployment is $ 27410.51
- Final savings after new model deployment is $185981.71 that is reduction in losses by ~ 87%

Appendix: Data Attributes

Snapshot of the data:

- 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

THANK YOU