Bank Fraud technical analysis

By Jocelyn Luo

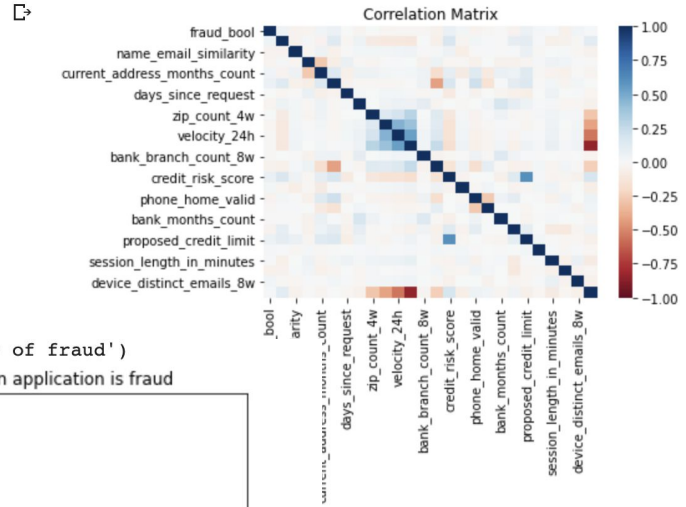
About the dataset

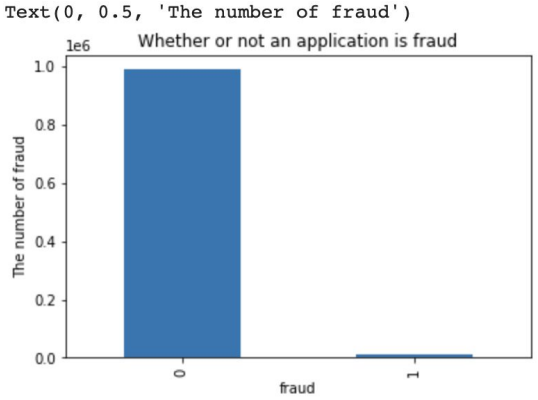
- A very clean dataset(row, column, duplicates)
- High correlation between velocity 4w and months

  (some months has more days)

- Imbalance dataset regarding to fraud boolean

  (considering collecting more data)





Process

Which group/features of customers might be an alert of fraud application?

EDA

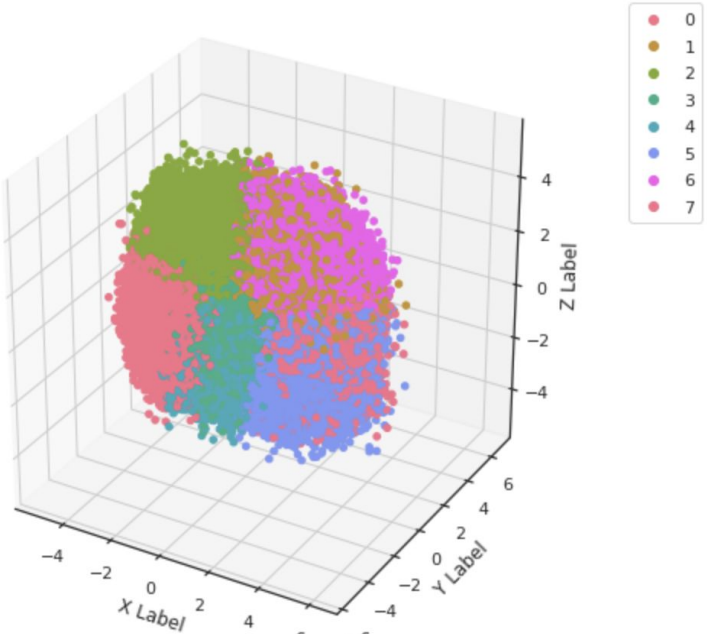
- By age
- By income

Classification

1. Logistic regression (low accuracy)
2. Random forest
3. Feature importance

Clustering (8 groups)

4. K-means (pca) – overall sense
5. Gaussian mixture models --specific description of features



'Cluster 6'

	fraud_bool	income	name_email_similarity	prev_address_months_count	current_address_months_count	customer_ag
count	19726.000000	19726.000000	19726.000000	19726.0	19726.000000	19726.0000(
mean	0.388827	0.643501	0.475026	-1.0	124.184680	38.37878
std	0.487496	0.275549	0.299237	0.0	87.916359	12.8316(
min	0.000000	0.100000	0.000113	-1.0	0.000000	10.0000(
25%	0.000000	0.400000	0.183553	-1.0	53.000000	30.0000(
50%	0.000000	0.700000	0.475287	-1.0	100.000000	40.0000(
75%	1.000000	0.900000	0.760774	-1.0	177.000000	50.0000(
max	1.000000	0.900000	0.999996	-1.0	398.000000	90.0000(

8 rows × 27 columns

Clustering

- 8 groups by elbow method
- K–means (with pca for 3 principal components):

  our customers are closely clustered (with rare outliers)

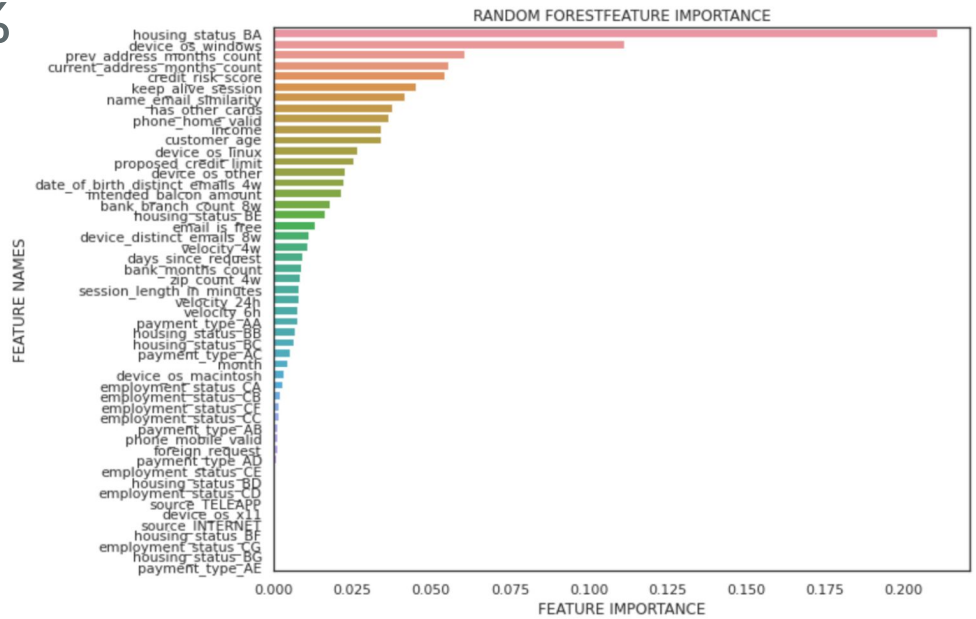
- Gaussian Mixture Model (without pca)

  Cluster 6 has the highest mean for fraud bool and highest value of current_address_months_count

What do we know from model?

Accuracy (random forest): 98%

- Housing_status_BA
- device_os_windows
- prev_address_months_count
- current_address_months_count
- credit_risk_score



Thank you

Analysis found at:

https://colab.research.google.com/drive/1kYarykG5T_aad6TFcrXHudnZk8-ffZSX#scrollTo=aNtkAlOeKMOi