Hands on Data Science – Final

# Fraud Detection

You have been brought in to help a telecommunications company with their fraud problem. The company reports that *new subscriber fraud* has gotten so out of hand that they need the government to change regulations to help them stop the spike in losses. Apparently, customers can sign up for service and have several new Apple, Samsung or other phone phones shipped to them – without paying a penny. Of course, once they receive the phones the customer disappears sells the phones overseas for a small fortune. Your job is to build a model to identify fraud before the customer can walk out the door.

*Here is a little bit about the scam:*

https://techcrunch.com/2019/06/26/verizon-gets-fcc-approval-for-a-60-day-network-lockdown-on-new-phones/

https://www.verizon.com/about/news/message-ronan-dunne-verizon-consumer-group

# Tasks

Unlike previous projects, this time I want you to write up a short executive style report using the following structure. I’m not going to provide much direction, you have all the tools at your disposal. There are two datasets, new\_acocunt\_subscribers.csv and phones.csv. LEFT join them together on phone\_id. Here is a hint:

pd.merge(transaction,phone,on=’phone\_id',**how='left')**

once joined, follow the standard process we’ve used over the last couple weeks. Finally, Make sure you have a name, email and date on the report!

## Executive Summary (5pts)

Your executive summary should be brief (i.e. short like 1 paragraph ) but address the following

1. **Briefly** describe the business problem you are trying to address
2. **Briefly** summarize your key point – how fraud could you catch with your best model.

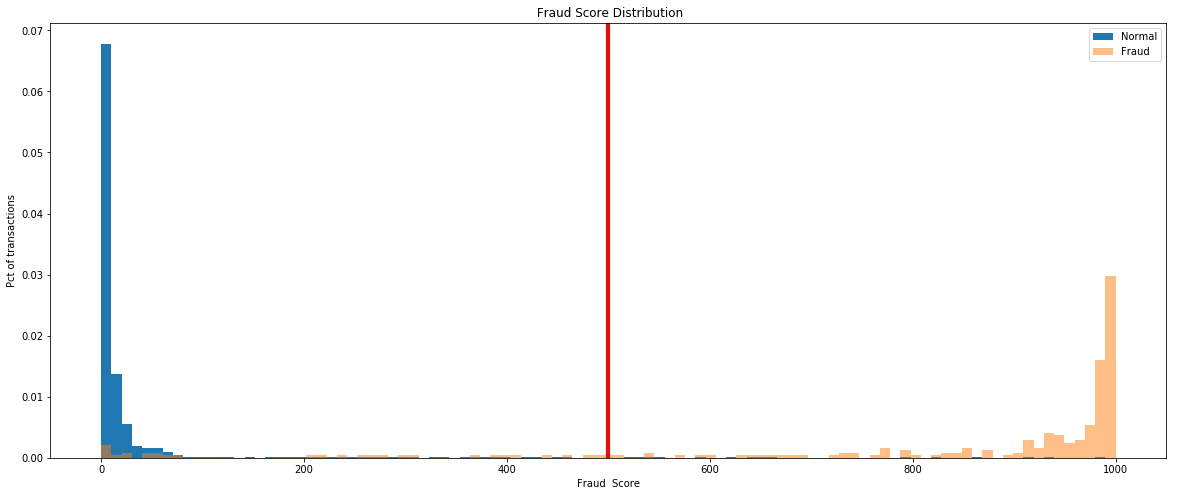
## Key Findings (25pts)

**Summarize** key findings in the data. back up these findings with charts and tables. You should be able to identify 3 numeric and 3 categorical features and explain how they relate to fraud. Back this up with appropriate summary statistics, charts, graphs and cross tabs.

## Model Performance Summary (45pts)

You will **build 3 models** of your choosing, your choice of using features, transformations etc. This section you should have a table that summarizes the following

1. For each Model
   1. What features were used in each model.
   2. What transformations you used in your pipeline (ex. Missing values, one-hot etc)
   3. What hyper parameters you decided to use, did you do any optimization
   4. Summarize Performance on Training and Test Split
      1. Accuracy (Train / Test) - what does accuracy tell you?
      2. AUC (Train/Test) – what does AUC mean and what does this tell you?
      3. Roc Chart (Train / Test)
      4. Confusion Matrix (count & pct – Test)
      5. Variable Importance – what variables are most important, does your exploration above include those? Why not?
      6. Score Distribution Fraud/Not Fraud( ex. Below), the red-line is where your suggested score threshold rule would be things to the right are fraud things to left are not-fraud



* + 1. TPR/FPR/Threshold FPR 1% - 5% ex:

--- score thresholds ---

fpr tpr threshold

0 0.01 0.89 0.23

1 0.02 0.92 0.09

2 0.03 0.92 0.08

3 0.04 0.92 0.06

4 0.05 0.93 0.05

## Recommendations (25pts)

You need to make at least 3 recommendations to the company.

1. Which model to deploy and why.
2. Recommend a False Positive Rate to operate at. At that FPR, what is the expected True Positive Rate and corresponding score threshold – should be based on TEST dataset. For every 1000 records your model identifies as fraudulent how many would be a False positive? Here is an example rule.
   1. IF fraud\_score GREATER THAN 0.xxx THEN “fraud”
3. Some other recommendation(s) of your choosing based on your findings.

## Extra Credit (10pts)

And for an extra credit, does your best model you beat an AUC of 0.95 on your **Test Data Set**?

## Extra Credit Pt 2. (10pts)

And for another extra credit opportunity. Can you provide me with a simple set of rules to explain my predictions with – think surrogate tree. And how much variability in my model score do these rules explain on your **test data set**. (hint: R-square)

## Turn in:

1. Executive Report \_yourname.docx - Clear, *concise* and accurate write up.
2. Notebook\_yourname.ipynb - Clear, repeatable notebook.
   1. I should be able to simply run your notebook by pointing at the data.
3. Notebook\_yourname.html – an example run of your notebook with the outputs.

## Appendix About the Data

No, this data doesn’t contain real names, emails or addresses though they may look the part.

new\_subscribers.csv

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| feature\_name | feature\_type | dtype | count | nunique | nunique\_pct | feature\_suggestion | Description |
| is\_fraud | TARGET | int64 | 10000 | 2 | 0.0002 | LIKELY CATEGORICAL, NUMERIC w. LOW CARDINALITY | target t variable fraud not fraud. 1 = fraud, 0 = legit |
| transaction\_id | NUMERIC | float64 | 9997 | 9997 | 0.9997 | EXCLUDE, GT 90% UNIQUE | unique identifier of transaction |
| accessory\_count | NUMERIC | float64 | 9997 | 10 | 0.001 | LIKELY CATEGORICAL, NUMERIC w. LOW CARDINALITY | number of accessories ordered |
| phone\_count | NUMERIC | float64 | 9995 | 5 | 0.0005 | LIKELY CATEGORICAL, NUMERIC w. LOW CARDINALITY | number of phones ordered |
| phone\_id | NUMERIC | float64 | 9999 | 16 | 0.0016 | LIKELY CATEGORICAL, NUMERIC w. LOW CARDINALITY | join to phone to get the name of the phones ordered |
| billing\_postal | NUMERIC | float64 | 9998 | 1336 | 0.1336 | NO WARNING | postal code where the person is getting the bill |
| total\_lifetime\_value | NUMERIC | float64 | 9999 | 3485 | 0.3485 | NO WARNING | expected life time value of this customer |
| shipping\_amt | NUMERIC | float64 | 9998 | 76 | 0.0076 | NO WARNING | amount of shipping charged |
| promo\_discount | NUMERIC | float64 | 9996 | 331 | 0.0331 | NO WARNING | amount of promotion discount |
| credit\_score | NUMERIC | float64 | 9998 | 311 | 0.0311 | NO WARNING | customer's credit score |
| email\_name | CATEGORY | object | 9997 | 9438 | 0.9438 | EXCLUDE, GT 90% UNIQUE | email name |
| billing\_state | CATEGORY | object | 9991 | 51 | 0.0051 | NO WARNING | billing state |
| billing\_address | CATEGORY | object | 9998 | 9998 | 0.9998 | EXCLUDE, GT 90% UNIQUE | billing address |
| customer\_name | CATEGORY | object | 9996 | 9411 | 0.9411 | EXCLUDE, GT 90% UNIQUE | name of customer |
| event\_timestamp | CATEGORY | object | 9995 | 9995 | 0.9995 | EXCLUDE, GT 90% UNIQUE | transaction time stamp |
| user\_agent | CATEGORY | object | 9998 | 2233 | 0.2233 | NO WARNING | web browser user agent |
| phone\_plan | CATEGORY | object | 9996 | 5 | 0.0005 | NO WARNING | what plan they signed up for |
| contract\_term | CATEGORY | object | 9997 | 5 | 0.0005 | NO WARNING | how long of a contract |
| phone\_protection | CATEGORY | object | 9999 | 2 | 0.0002 | NO WARNING | did they buy phone protection |
| phone\_number | CATEGORY | object | 9994 | 9994 | 0.9994 | EXCLUDE, GT 90% UNIQUE | what is their current phone number |
| ip\_address | CATEGORY | object | 9998 | 2476 | 0.2476 | NO WARNING | what ip address did they use when ordering the phone over the internet |
| avs\_code | CATEGORY | object | 9995 | 24 | 0.0024 | NO WARNING | what was the address verification code when they paid for the accessories |
| email\_domain | CATEGORY | object | 9991 | 1932 | 0.1932 | NO WARNING | what is the email domain |
| phone\_trade\_in | CATEGORY | object | 9999 | 2 | 0.0002 | NO WARNING | did they trade in a phone? |

phones.csv

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| feature\_name | feature\_type | dtype | count | nunique | nunique\_pct | description |
| product\_id | NUMERIC | int64 | 305 | 305 | 1 | Product id joins to product id on the transaction, be sure to left join |
| product\_name | CATEGORY | object | 305 | 305 | 1 | phone name and details |
| product\_mfg | CATEGORY | object | 305 | 12 | 0.03934 | phone manufacturer |

**Example Score Distribution Code**

# -- assign predictions based on threshold --

predictions.loc[predictions['score'].astype(float) > score\_threshold, "predicted\_fraud" ] = 1

predictions.loc[predictions['score'].astype(float) <= score\_threshold, "predicted\_fraud" ] = 0

fraud = predictions.loc[predictions[FRAUD\_LABEL] == 1 ]

legit = predictions.loc[predictions[FRAUD\_LABEL] == 0 ]

bins = np.linspace(0, 1000, 100)

plt.figure(figsize=(20,8))

plt.hist(legit['score'].astype(float) , bins, alpha=1, density=True, label='Normal')

plt.hist(fraud['score'].astype(float) , bins, alpha=0.5, density=True, label='Fraud')

plt.legend(loc='upper right')

plt.title(" Fraud Score Distribution")

plt.xlabel("Fraud Score")

plt.ylabel("Pct of transactions");

plt.axvline(x = score\_threshold ,linewidth=4, color='r')

plt.show()