Wells Fargo Campus Analytics Challenge 2021: Machine Learning Model to Predict Suspected Elder Fraud

October 11, 2021 | Mark Del Grande

Link to code:

https://github.com/Mark-DelGrande/Wells-Fargo-Campus-Analytics-Challenge/blob/main/WellsFargoCampusAnalyticsChallenge.ipynb

Feature Engineering

The first step in making a model was to understand what all of the data represented. After examining the data I had a few initial thoughts about what useful features I could create from the data and which pre-existing feature would be important. Here they are:
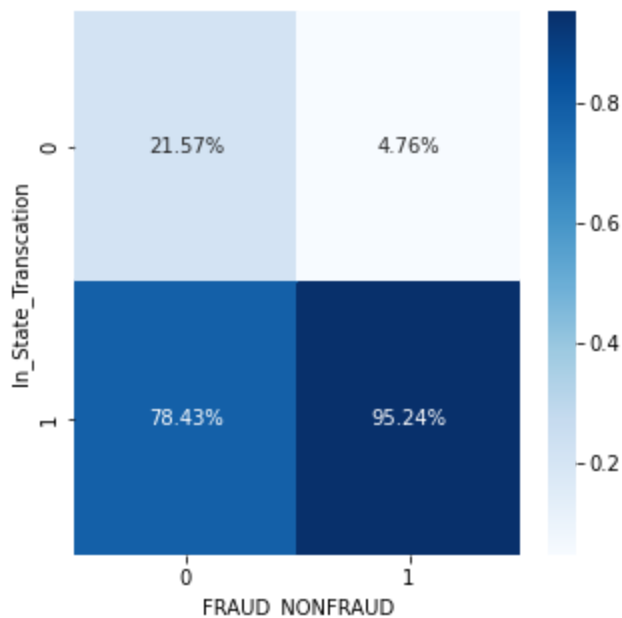
Is this an in-state transaction?

For this feature I standardized all the values in both the STATE_PRVNC_TXT, and CUST_STATE columns. I then compared the two to see if they were the same and then converted it to binary 1 for in state 0 for out of state. As you can see from the confusion matrix on the left there is some level of correlation between being an Instate transaction and a fraudulent transaction.

Customer Age

Clearly, since this whole competition was based around Elder fraud I knew from the very beginning this was going to be an important feature.

Days Between Transaction and Last Account Update

My thoughts with this feature were that people may have had their account updated without them knowing so there could be a link between a recent password or phone number update with the account and a fraudulent transaction. For this, I made two features called DAYS_SINCE_PWD_UPDT and DAYS_SINCE_PH_NUM_UPDT.

Percentage of Account the Transaction is For

Next, I wanted to look at what percentage of the account was being withdrawn. For example, I would think that if the entire account balance was being withdrawn in one transaction that would raise some red flags. In addition to this, I use the TRAN_AMT and ACCT_PRE_TRAN_AVAIL_BAL to understand how much money this percentage was worth.

Device Age Compared to Account Age

Another feature I created was looking at how long the account had been open compared to how long this device had been associated with the account. This allowed me to see if an account had just been opened and that's why the device is new or if this device had just been added to a preexisting old account. A new device on an old account may be represented by a hacker adding a device to someone's account to make a fraudulent transaction.

One Hot Encoding

The last feature engineering I tried was one hot encoding on several of the categorical columns such as AUTHC_PRIM_TYPE_CD, ALERT_TRGR_CD, DVC_TYPE_TXT, and AUTHC_SCNDRY_STAT_TXT. I dropped all of these except for the ALERT_TRGR_CD encoding for MOBL since they had negligible impacts on the results and led to a more complicated model.
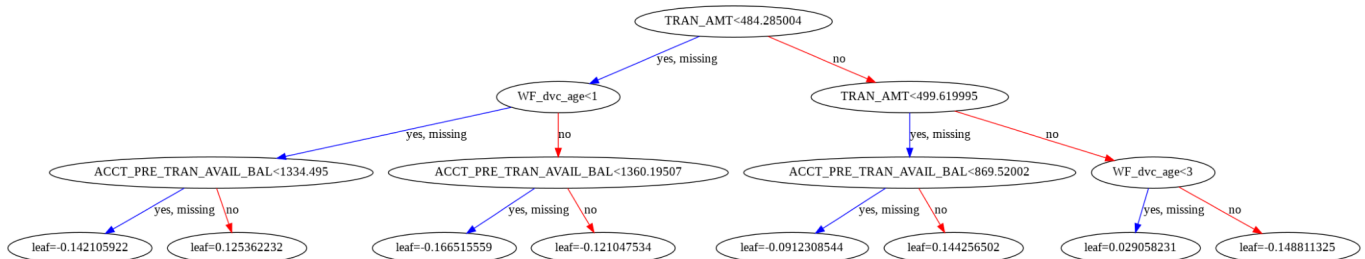
Classification Model

Once the data was formatted and ready for predictions I started by training an XG boost classification model. The data was split into 67% for training and 33% for testing. With this data, split my model achieved an accuracy rate of 95.09%.
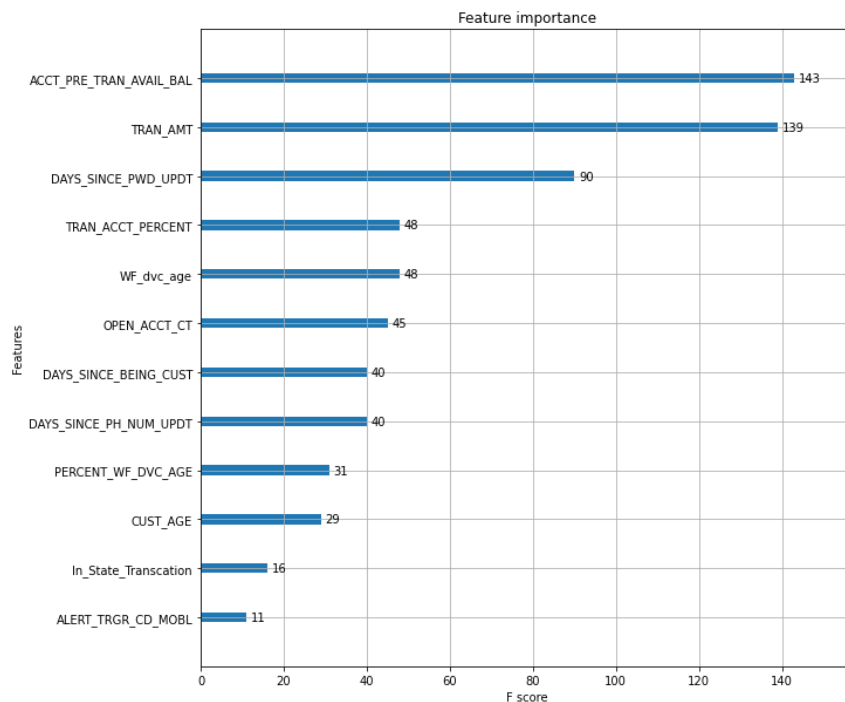
Here is the importance of all the features. In my final model, I found that these 12 features were the most beneficial to my predictions. Therefore, I dropped all other features I had made to keep the model relatively simple and still have accurate results.

Above is what the tree plot looked like once the XG boost Classification model was trained.

```
TRAN_AMT                            54.250000
ACCT_PRE_TRAN_AVAIL_BAL              0.000000
CUST_AGE                            23.000000
OPEN_ACCT_CT                         4.000000
WF_dvc_age                         634.000000
In_State_Transcation                1.000000
DAYS_SINCE_PWD_UPDT                112.000000
DAYS_SINCE_PH_NUM_UPDT             915.000000
TRAN_ACCT_PERCENT                    0.002297
DAYS_SINCE_BEING_CUST             3611.000000
PERCENT_WF_DVC_AGE                   0.175575
ALERT_TRGR_CD_MOBL                   1.000000
```

y=0 (probability **0.999**, score **-6.668**) top features

Contribution?	Feature
+2.924	ACCT_PRE_TRAN_AVAIL_BAL
+1.255	TRAN_AMT
+1.247	<BIAS>
+0.742	TRAN_ACCT_PERCENT
+0.187	DAYS_SINCE_PH_NUM_UPDT
+0.169	ALERT_TRGR_CD_MOBL
+0.068	OPEN_ACCT_CT
+0.060	PERCENT_WF_DVC_AGE
+0.053	DAYS_SINCE_PWD_UPDT
+0.045	WF_dvc_age
+0.026	DAYS_SINCE_BEING_CUST
-0.050	CUST_AGE
-0.057	In_State_Transcation

Lastly, here is an example of a test prediction. The top picture shows the data that was inputted into the model and the bottom pictures show the score given to this data and the reasoning behind the score. It scored a -6.688 and this means the model was 99.9% sure that this was not a fraudulent transaction.

Flow Chart