Troy Quicksall

Detecting Bank Account Fraud

Introduction

With online banking increasing in popularity, so too does the prevalence of fraudulent scams. In 2022 the FTC reported 10 billion in losses which was a 1.2 billion increase from the year prior. These fraudulent transactions can take place in several different ways. Peer-to-peer payment scams are one of the methods in which scammers can conduct fraudulent transactions through applications like Zelle, PayPal, Venmo, etc. In addition to this, wire transfer, and phishing scams can lead to a compromised bank account. The good thing is that since most of these accounts and transactions take place online, the data is abundant and readily available. With access to this data, we should be able to use predictive modeling in order to gauge what accounts are vulnerable, and what transactions are most likely to be fraudulent.

When customers report fraudulent transactions to their bank, the bank will often refund their customers the amount of the fraudulent transaction. Therefore, banks would be interested in reducing this overhead cost. Additionally, payment apps like PayPal, in order to keep a good reputation, would benefit from preventing fraud.

The Data:

I sourced the data for this project from Kaggle. The data has 32000000 rows of data, each a transaction, and 32 columns of variables describing the details of the transaction and the account involved. One of the columns tells whether the transaction ended up being fraudulent. Columns include information on; fraud, income, name and email similarity, previous address, current address, age, payment type, payment amount, zip code, applications made, bank

branches, date of birth, employment status, credit risk score, email account, housing status, phone number, other bank cards, credit limit, device used, session length, and month taken place.

Methods/Results

Data Preparation:

The data provided was already mostly clean. Most of the columns are numeric values that already have entries given for missing values. For example, one column is for number of months in current address, the missing values have already been filled in the dataset. Additionally, all Boolean values were already 0 or 1. However as mentioned in previous milestones, there are a few categorical columns, which I needed to convert to dummy variables. Without converting to dummy variables these entries would have non-numeric values, making regression difficult.

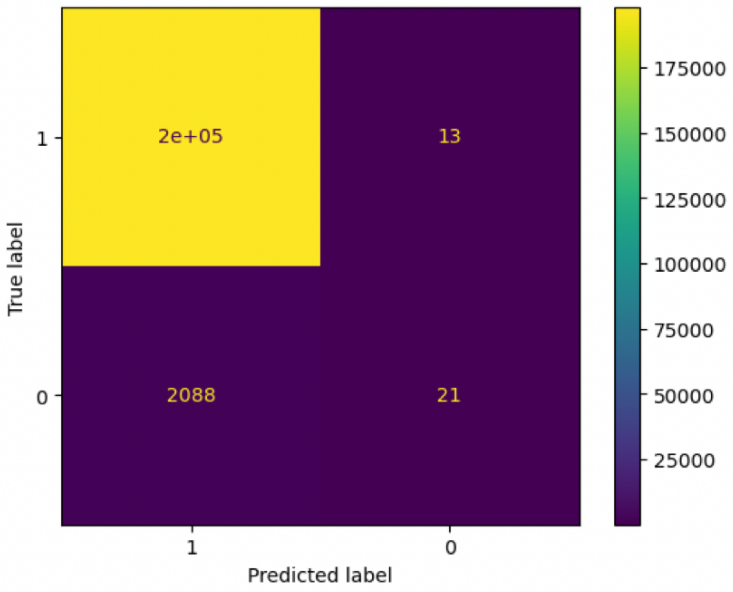
I then used a correlation matrix to identify redundant variables as described in the textbook, "Applied Predictive Analytics". Any correlations greater than .9 I considered redundant and removed from the dataframe. The results only indicated that 'source_TELEAPP' and 'source_ITERNET' were redundant.

Next all that was left to do was split the dataset into test and train sets and add MinMax Scaler. I used a MinMax scaler, because not all the metrics are on a similar scale, so standardizing them should help the model interpret the relationship more accurately.

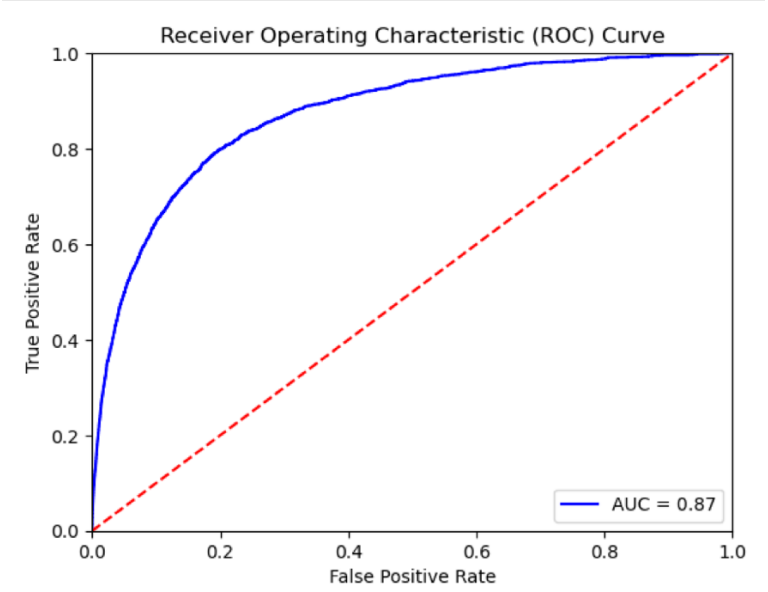
Modeling/Results:

Since the outcome variable is binary (either fraud or not), I used logistic regression, as it will provide the best results. Logistic regression determines the dependent variable based off the relationship between all the independent variables. Many of the independent variables are categorical, therefore those will need to be converted to dummy variables in order to fit the qualitative data into the regression.

First, I looked at the F1 score. The F1 score can be interpreted as the mean of precision and recall. Recall is how often the model correctly identifies positive instances. In our case how often a fraudulent transaction was detected (or non-fraudulent). Therefore, the closer to 1 the F1 score, the better. This model has a great F1 score at .98.



The confusion matrix shows us the predicted true and false values versus the actual true and false values. The top left tells us the correct true predictions, the top right shows us false positives, the bottom left shows us false negatives, and the bottom right shows us true negatives. As we can see our largest number is where we want in, in the top left under true positive predictions.



The ROC curve shows us the rate a model predicts the true outcome correctly versus incorrectly. The more area under the curve the better. As we can see based off the ROC-Curve the model performed well.

Cross-validation is useful to estimate the success of a machine learning model on unseen data. This means to use a limited sample in order to measure how the model is expected to perform on data not used to train the model. During cross-validation the model performed well, holding an accuracy of 98%.

Conclusion

The goal of this project was to build a model that could predict whether a transaction is fraudulent. The interpretation of my results shows that the model performs very well in doing that. The model accuracy (normal and standardized), and F1 Score are both .98. With an accuracy of 98%, and cross-validation having similar results, I feel this model can predict a

fraudulent transaction given the appropriate metrics. With all the various analysis done, I also feel confident the model is not overfit, and will perform well on live data.

Given that most all the metrics in the dataset are available at the time of the transaction, this model could be implemented to predict fraud on a transaction-by-transaction basis. Banks could use this model as a part of their payment processing pipeline to reduce overall levels of fraudulent transactions.

Take Aways:

Through the course of this project, I learned more about evaluating logistic regression models. I was very comfortable with linear regression models, and models that predict non-categorical outcomes. There are plenty of metrics I am familiar with in Linear Regression like R2, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). However, I'm less familiar with the evaluation of categorical predicting models. By learning more about evaluating these models, I believe I gained a better understanding in general about what makes them successful.

References:

Fraud facts and statistics. John Marshall Bank. (2024, March 12). https://www.johnmarshallbank.com/resources/security-center/fraud-facts-and-statistics/

Bennett, K. (n.d.). *5 common types of bank account fraud and how to protect yourself*. Bankrate. https://www.bankrate.com/banking/common-types-of-bank-account-fraud/

Assumptions of logistic regression. Statistics Solutions. (2024, April 17). https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/

How to evaluate a logistic regression model?. Tutorialspoint. (n.d.). https://www.tutorialspoint.com/how-to-evaluate-a-logistic-regression-model