Fraud Detection & Prediction

Avery Fulton & Zain Elsell

Feb 12, 2023

Fraud detection is a critical challenge facing businesses and organizations across many sectors of the economy. Fraudulent activities cost businesses billions of dollars each year, and detecting these activities in a timely manner is essential to minimize financial losses and protect the integrity of systems. In this project, we aim to develop a robust and effective fraud detection system by analyzing and modeling transaction data from the IEEE Computational Intelligence Society competition that took place in 2019, to detect anomalies and potential fraudulent activities. The goal of this project is to provide a reliable and efficient solution for organizations to mitigate the risk of fraud and ensure the security of their financial transactions in addition to allowing us to develop our skills in exploratory data analysis, machine learning, and applying the regression techniques covered in this course. Primarily our goal is to create a logistic predictor variable that will provide us with an output that categorizes each transaction as either fraudulent or not fraudulent.

We got the data set from a Kaggle competition hosted by the IEEE Computational Intelligence Society. The IEEE CIS is an established forum for researchers, practitioners, and students to exchange ideas and developments of Computational Intelligence. We believe that the data fits all of the criteria assigned, and is authentic. We have two sets of data, financial transactions, and declassified identification information. The data set we are more particularly interested in is the transaction data section which includes over 100+ features which we will be using only a small subset of in order to conduct our analysis and construct our models. Additionally, the data used had over 500,000 observations which each represent an individual transaction. Our data set is not a random sample as it is just a list of transactions from Visa and Mastercard holders between two points in time.

The data set contains some of the following:

- Transaction time delta, which is created from a given reference time and date time which is not an actual timestamp. (Numerical Continuous)
- Transactions amount, which is the amount of money in USD for each individual transaction. (Numerical Discrete)
- Purchase Medium, which is the transaction on which the card took place (Credit or Debit) (Categorical)
- Location data, includes distance and geographical location along with the address of the cardholder. (Numerical Continuous and Categorical)

- A product code, that functions as a UUID for each individual transaction (Numerical Discrete)
- Card Variables 1 through 6 which is the entirety of the card number used along with the CVV. (Numerical Discrete)
- Match Variables 1 through 9 which includes if entities such as cardholder name, cardholder address, etc match with the information on record. (Categorical)

In summary, we likely won't be using all of the variables listed above, as we are not sure on whether the time delta variables fall out of the scope of this project such as requiring the use of time series. In addition, there is a lot of data that we will omit as we don't have the column header and are not sure what they represent (Columns starting with "V").

Data: https://www.kaggle.com/competitions/ieee-fraud-detection/data