Capstone 1 Project : Statistical Data Analysis
Fraud Detection in Mobile Payment Data
Springboard Data Science Career Track
Eric Cruz

Think of the following questions and apply them to your dataset:

- Are there variables that are particularly significant in terms of explaining the answer to your project question?

There may be variables or combinations and functions of variables that seem significant in explaining the question of whether a transaction is fraud. In particular, the "Type" and "Amount" come forward in the data visualizations for reasons discussed below.

- Are there significant differences between subgroups in your data that may be relevant to your project aim?

There is a significant subgroup in the "Type" variable because only 2 of the 5 types are represented for the target cases (e.g. isFraud=1). The more important subgroup is the target variable itself, the "isFraud" boolean indicator. The dataset is highly imbalanced with only 0.13% of approximately 6 million transactions.

- Are there strong correlations between pairs of independent variables or between an independent and a dependent variable?

There is a strong correlation between "Amount" and "Old Balance Orig" which can be seen in the fraud transactions but not the non-fraud transactions, and higher but not quite as significant between "Amount" and "New Balance Orig". This is plausible given the provided description "In this specific dataset the fraudulent behavior of the agents aims to profit by taking control or customers accounts and try to empty the funds by transferring to another account and then cashing out of the system."

- What are the most appropriate tests to use to analyze these relationships?

The Paysim dataset is a synthetic simulation of mobile payment data, because actual data is generally not available due to proprietary concerns. This means the creation of underlying data can be considered a bootstrap approach. The test to determine fraud is inferential, but not exactly along the lines of traditional hypothesis testing due to the imbalance observed in the target boolean variable. Rather than a linear regression analysis, a more appropriate approach for binary classification is logistic regression. It also seems plausible to consider a Bayesian approach, such as the GaussianNB model.

However, there are several other models that may be appropriate given the binary classification problem, which will be considered as we learn about models in upcoming sections of the course. Some potential candidates are Random Forest Classifier, Support Vector Machine, and KNN just to name a few.