Data

- Credit Card Fraud - https://www.kaggle.com/datasets/dhanushnarayananr/credit-card-fraud

Related Work

1. Credit Card Fraud Detection Project using Machine Learning
    a. This project uses machine learning to differentiate between fraud and not fraud. We intend to do something very similar, only our dataset (above) is very different from the one they used - they had 30 continuous numerical features for each observation, and we have three continuous and five discrete.
2. Credit Card Fraud Detection: Top ML Solutions 2021
    a. This article does not have a code walk-through, and is not an example to follow. Instead, it has a ton of information about fraud and detecting it with ML, and all the different models and the way to use them. This article will also give us a lot more background information in the subject area.
3. Credit Card Fraud Classifier Using Undersampling
    a. In this dataset, the columns and variables were unlabeled due to confidentiality aside from the time of the transaction and amount of money spent in the transaction. In order to obtain a 97% Recall value, those analyzing the data set under sampled their data, or balanced out their test and training data in order to accurately predict fraudulent data. This is an important technique that may be needed when we continue further with our classifier.
4. Credit Card Fraud Detection Using Anti K Nearest Algorithm
    a. In this project, they used a reverse K-Nearest Neighbor algorithm and a technique named SODERNN, a modern algorithm based language. They used factors such as distances from the last transaction which we plan to do as well. They studied online fraudulent transactions as well, noting that online credit card fraud is far more likely in occurrence than in-store fraudulent transactions.

Additional Resources

1. Numpy allows us to perform linear algebra: the basis of predictions.
2. Pandas allows I/O of data as well as initial data processing.
3. Sklearn (neighbors->KNeighborsClassifier) allows the creation of the most basic KNN towards basic classification based on the evaluation of the nearest neighbors.
4. Sklearn (model_selection-> train_test_split) defines how we want to split our data for accuracy analysis.

EDA

1. Inspect Data
    a. We use the info() function on our dataframe to check for missing values and the data types of the variables

2. Remove Missing Values
    a. We find that there are no missing values in the dataset, so we do not need to perform any deletion or imputation.
3. Convert Fields to Suitable Data Types
    a. Many of the categorical variables were stored as floats, so we convert them to integers
4. Check for Invalid Data
    a. Use the min() function on our dataframe to check if there are any negative values for the fields measuring distance. We don't find any, so we can move to the next step.
5. Check for Patterns and Trends in the Data

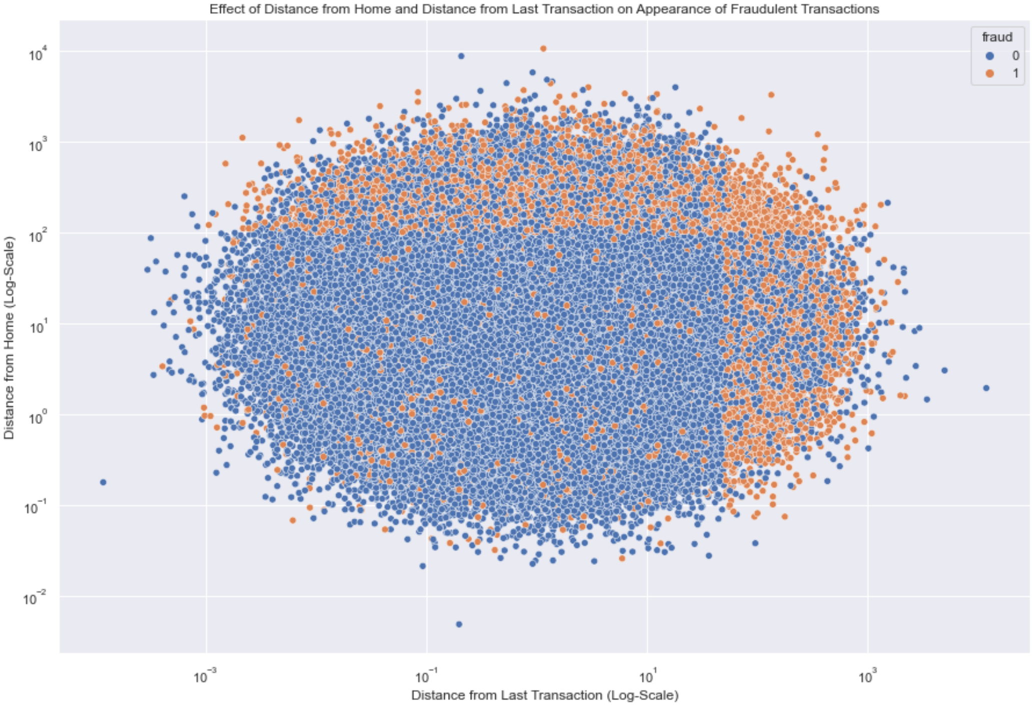

Fig. 1

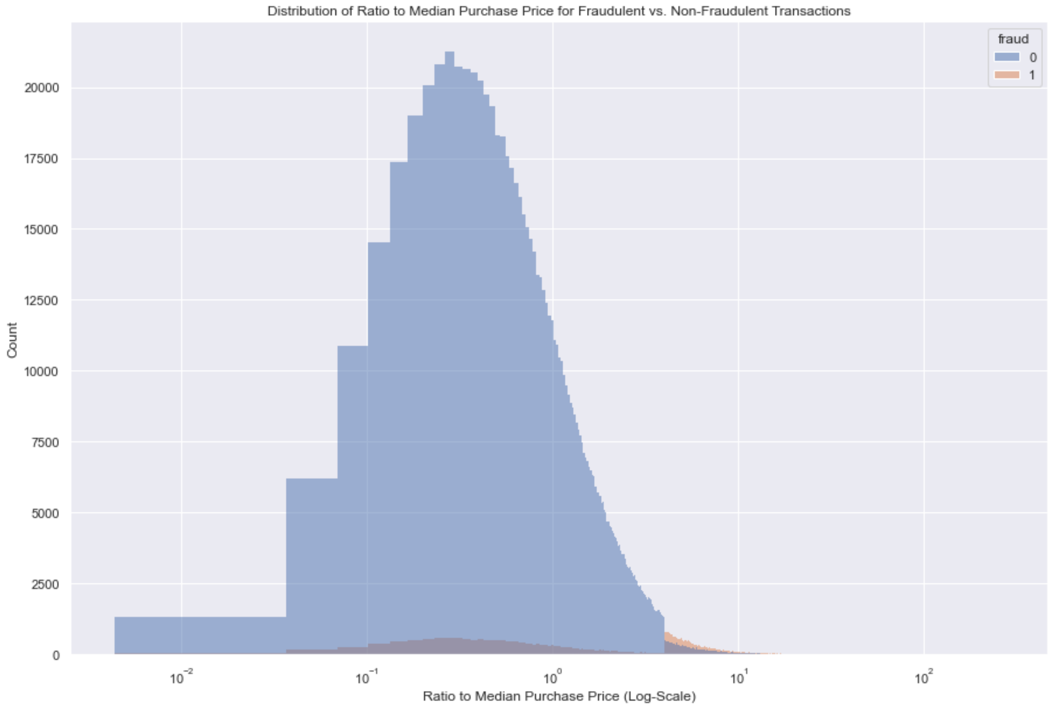

Fig. 2

a. First, we plot ‘distance_from_last_transaction’ vs. ‘distance_from_home’ on a log-log scale with different colors for transactions labeled fraud (orange) vs. non-fraud (blue). In Fig. 1, We can see that there is no correlation between ‘distance_from_last_transaction’ and ‘distance_from_home’, but there seems to be a positive association between fraudulent transactions and either/both of the variables. This association suggests that a two-sample t test between samples of fraudulent and non-fraudulent transactions could show a statistically significant difference in the distance variables.

b. Next, we attempt to check if there is a difference in the distributions of ‘ratio_to_median_purchase_price’ for fraudulent vs. non-fraudulent transactions. Fig. 2 shows that there may be a difference in the ratios at which the occurrence of fraudulent transactions are higher. This histogram also reveals that there is an imbalance in the amount of labeled transactions, as there seem to be many more non-fraudulent transactions compared to fraudulent ones. When we conduct tests and classification, we should first extract more balanced samples from the data. We may also find a statistically significant difference in the ratio to median purchase price variable if we investigate further with hypothesis testing.