Abstract:

Credit card fraud is a serious social issue. In this research, we would like to explore how transaction amount and time of fraudulent and non-fraudulent cases are affected by various variables. We also did clustering algorithms to build the models that help us better understand the difference between fraudulent transactions and non-fraudulent transactions were possibly related to clusters of variables. As a result, we found the most effective explanatory variables and a difference in variables groups for fraudulent transactions and non-fraudulent transactions, however, our study is limited, and it could be improved by expanding the dataset.

Introduction:

Credit card fraud is the criminal use of some else's credit cards. In this lab, we will find the variables that are most related to the credit fraud amount and time, and use clustering and classifications to obtain the models that explain the relationship between variables. We also used ensemble learning to improve the result of our models. We have three main findings. We found that the most variables that explained the most of the amount and time of credit card fraudulent cases are V13, V15, V19, V24 and V26. We did four models using K-means++, DBSCAN, Logistics Regression and Zero-inflated, and we also explored the performances of four models. As a result we found that DBSCAN has the best performance of all models. Last but not least, we found that the most credit card fraudulent cases happened either in a really short amount of time or took a while. And most fraudulent transactions either stole a small amount of money.

Data:

There are 284807 observations in our datasets and there are 31 variables. The variables include Time, Amount, Class and twenty-eight anonymous variables from V1 to V28.

Variable	Description
Time	The transaction duration
Amount	Transaction amount
Class	A label that 1 is fraudulent transactions and 0 is non-fraudulent case
V1-V28	28 different features about users and being concealed due to privacy issues

Results: A summary of the analyses you completed for the lab + a discussion on the ***results*** that talks about these topics:

There are several findings in our research, and we classified them into four groups

First, we found the time records for two sections. For the feature engineering section, we found the PCA test took 0.51 seconds to compute and Calculating pairwise Manhattan and Euclidean distances between fraudulent and non-fraudulent data takes 0.5 second. Calculating the TSNE test took the longest time, which is 62.3 seconds. For the clustering and classification section, we found the computation time for K-means++ is 4.71 seconds, the computation time for DBSCAN is 81.6 seconds, and the computation time for Logistics Regression is 0.04 seconds. Finally, the computation time for Zero-inflated is 3.6 seconds. Thus, the DBSAN took most time to calculate and Logistics Regression took least time to calculate.

Second, we use feature engineering methods to find the most effective variables for time and amount variables. Based on the original dataset, we found V1, V2, V3, V4 and V6 have the greatest mean and median pairwise differences along time and amount. We did a min-max normalization to scale our dataset into the range of 0 to 1. Based on the scaled dataset, we found new top five variables that have the highest mean and median pairwise differences, and they are V13, V15, V19, V24 and V26. We did a PCA test and visualization on a randomly drawn 1000 samples. There are at least two distinguishable groups in the plot. One is the non-fraudulent group, which is at the lower left end of the scatterplot, Another is the fraudulent group, which is at the upper right end of the scatterplot. We then did a T-SNE test, and found that there are two clusters. After reading the t-SNE plot, I think the dimensionality reduction method is better than other methods for the dataset, since it explains the difference between the fraudulent group and non-fraudulent group.

We did a Euclidean and Manhattan distance between fraudulent and non-fraudulent data. Using both Manhattan and Euclidean calculation, we found that the distance between fraudulent and non-fraudulent data is higher than the distance between fraudulent and fraudulent data. We then divided amount and time variables into 10 deciles for both fraudulent and non-fraudulent data. For fraudulent data, we found that most fraudulent data are found in the top 10% of amount and 60% of total time, which means that the most credit card fraudulent cases happened either in a really short amount of time or took a while. We also found that most fraudulent data are found in the top 10% of amounts. For non-fraudulent data, we found it doesn't have a significant pattern in terms of money amount and time.

Third, we found the best model for clustering the data. We used the top five scaled variables that have the largest distance to Time and Amount, and we did four types of clustering and classification algorithms and obtained the *precision*, *recall, F1 score, Silhouette Index, and Calinski-Harabasz Index* for our models. First, we did K-means++ algorithm to divide them into two clusters. The precision is 0.034, recall is 0.331, F1 score is 0.061, Silhouette Index is -0.028, and *Calinski-Harabasz Index* is 32.874. Precision, recall, and F1 tells us the model is not strongly relevant to Time and Amount, and the most relevant data could be explained. According to the Silhouette Index and Calinski-Harabasz Index, our points in clusters are not very tightly grouped and the variance is relatively large. Second, we did the DBSCAN algorithm, and we

found the precision is 0.147, recall is 0.097, F1 score is 0.117, Silhouette Index is 0.628, and *Calinski-Harabasz Index* is 73.339. For the linear regression model, we expect to get a set of similar measurements like other models', however, we did not get what we expected in this study. We got all 1 for precision, recall, F1 score, Silhouette Index and *Calinski-Harabasz Index* . Finally, we did a zero-inflated algorithm, and we found that the precision is 0.2, recall is 0.04, F1 score is 0.007, Silhouette Index is 0.721, and *Calinski-Harabasz Index* is 5.251. Among these models, we found the model based on DBSCAN is the best since the accuracy of the model is highest. For the third part of our study, we also did ensemble learning. We expected to see the precision, recall, F1 score, Silhouette Index and *Calinski-Harabasz Index* of the combination of our previous four models using Max-Voting and Stacking methods. However, we did not get what we expected in this study, and we got 1 instead.

Lastly, we did a pairplot for all variables and two histograms for time and amount. We found that the most credit card fraudulent cases happened either in a really short amount of time or took a while. And most fraudulent transactions either stole a small amount of money.

- Which variable(s) seem(s) to explain the outcomes the most?

V13, V15,V19,V24 and V26

- What would be some of the ways to improve this analysis?

If there are equally large sizes of fraudulent and non-fraudulent data, we might be able to better differentiate clusters and measure variables' performances

We could use cross-validation to test the models ability.

- Do you see any typical causal inference problems (***selection bias***, ***simultaneity***, ***omitted variable bias***) in this analysis?

There might be selection bias, because we don't know if these data are from the same transaction sources or multiple transaction sources. Also the risk level of the transaction sources are different. So there might be the issue that data was not randomized.