Name: Anim Ohene, Henry Nguyen

Progress Report II

The first task completed was re-evaluating the dataset and PCA discovery. Upon performing PCA on the subset of the dataset, utilizing 7 of the features– amount, nameOrig, oldbalanceOrg newbalanceOrig, nameDest, oldbalanceDest, and newbalanceDest, the visuals were very poor, as can be shown in Figure 1. Table 1, which shows the Mutual Information scores for many of the features offers a good explanation for these results. As a result, of these poor visuals the target variable classes overlap as shown in Figure 1, and thus, no insight can be gained. This finding puts great emphasis on the feature selection and engineering stage of the machine learning pipeline, weighing great importance on preprocessing, engineering, and as in the case of the students, the careful selection of features to employ the model on. The students took the action of removing the noisy features, utilizing only 2 features– "step" and "type". To this end, it wouldn't seem necessary to employ PCA, however, the findings of Figure 1.2 show otherwise. According to Figure 1.3 which shows the use of PCA in the 2 features, more variability is captured than without applying PCA, as shown by Figure 1.2. PCA aims to maximize the variability between the features and remove noise. Despite being the "best" features, according to Table 1, there is still a low dependency between the selected features and the target variable. As such, PCA effectively served to reduce the noise and resulted in more interpretability of the dataset.

Table 1: Mutual Information Score Of features

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFlaggedFraud
0	0.010813	0.170882	0.002459	7.006855e-07	0.002611	0.000642	0.000001	0.000177	0.000092	0.0

Table 1 shows the mutual information Scores of all features concerning the target variable. Higher scores represent higher dependency with the target variable and lower scores show less dependency(less relation). * Remark: All graphs and tables are subject to change in the final report*

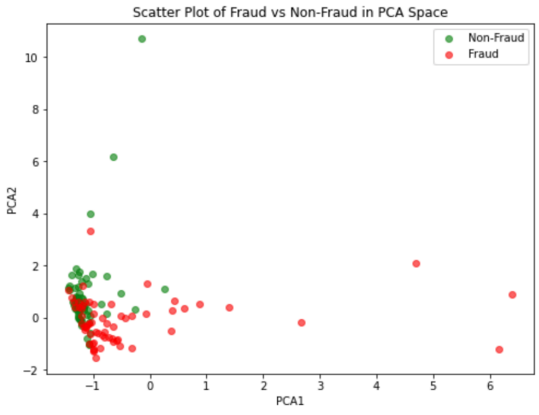

Figure 1

Shows the result of PCA to obtain 2 principal components from 7 of the features– amount, nameOrig, oldbalanceOrg newbalanceOrig, nameDest, oldbalanceDest, and newbalanceDest. The red dots represent fraud cases and the green represent Non-Fraud cases.* Remark: All graphs and tables are subject to change in the final report*

Figure 1.2

A plot of the raw preprocessed and normalized features on the two coordinate planes. Red represents fraud transactions and green represents non-fraud transactions. * Remark: All graphs and tables are subject to change in the final report*

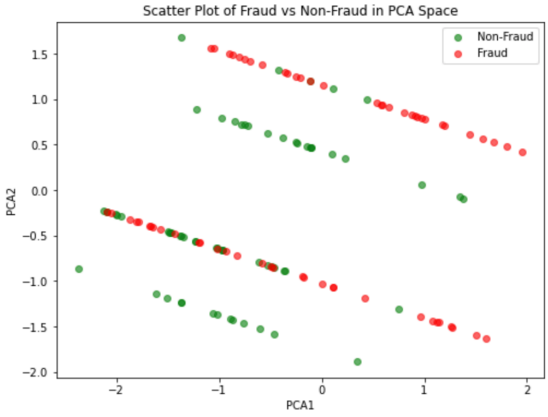

Figure 1.3

A plot using 2 principal components from the original features. Red data points represent fraud transactions and green data points represent non-fraud transactions. * Remark: All graphs and tables are subject to change in the final report*

After Selecting the dataset, data exploratory/summary statistics, data preprocessing, and feature engineering, the next step in the pipeline was to perform the elbow method. The elbow method is a technique used to identify the optimal number of clusters to use. This is done by first obtaining the Within Cluster Sum of Squares(WCSS). WCSS value is the sum of the variance of each cluster and its data point. As the number of clusters increases WCSS decrease. This is because eventually if each data point becomes its own cluster the distance between that data point and itself is 0. Similarly, as the number of clusters decreases WCSS increases. The question becomes how can one obtain an optimal number of clusters such that clusters are not underfitted and no information is learned, but also that clusters aren't overfitted, such that the data has been partitioned into too many sub-categories. By visualizing the variation between clusters and data points as a function of the number of clusters, one can spot the point at which WCSS stops dropping rapidly–represented as an “elbow” in the graph and obtain the optimal amount of clusters. According to Figure 3, the students found that the optimal number of clusters was 4.

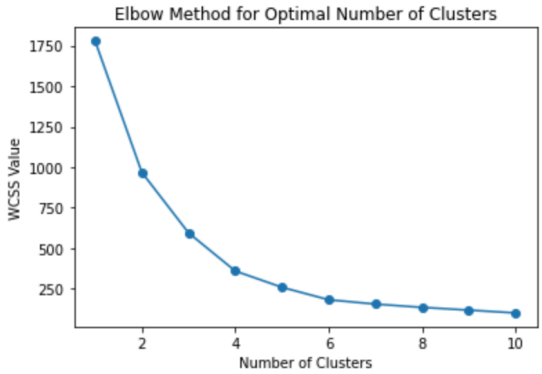

Figure 3

Figure 3 shows the elbow method. The X-axis is the number of clusters while the Y-axis represents the WCSS value for the corresponding amount of clusters.* Remark: All graphs and tables are subject to change in the final report*

After, obtaining the optimal clusters, the students performed the k-means clustering algorithm. In k-means, cluster centroids are randomly initialized among the data points. Data points are assigned to the initialized centroid based on their distance from a given centroid to other initialized centroids. Afterward, the centroids are relocated to their center of mass. This process of reassigning data points to their closest centroid and relocating each cluster to its center of mass is repeated until the center of mass of each centroid does not change signaling the final cluster assignment. Figure 4 shows the graph of using means to partition the data.

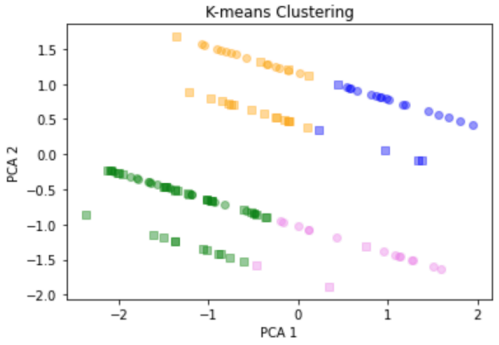

Figure 4

Figure 4 shows the use of k-means clustering with 4 clusters, represented by pink, blue, orange, and green. Data points that are boxes represent non-fraud cases and circles represent fraud cases.* Remark: All graphs and tables are subject to change in the final report*

Next was to perform Spectral Clustering on the dataset. Unlike the k-means algorithm that uses centroids, spectral clustering is a graph-based approach where each data point is a node and an edge is a connectivity between the points. In spectral clustering, the goal is to maximize the similarity between the points in the clusters. The steps are to first, construct the similarity matrix, compute the Laplacian matrix, conduct eigendecomposition to get the first C eigenvectors–U, then apply k-means to U. Figure 5 shows the result of the spectral clustering.

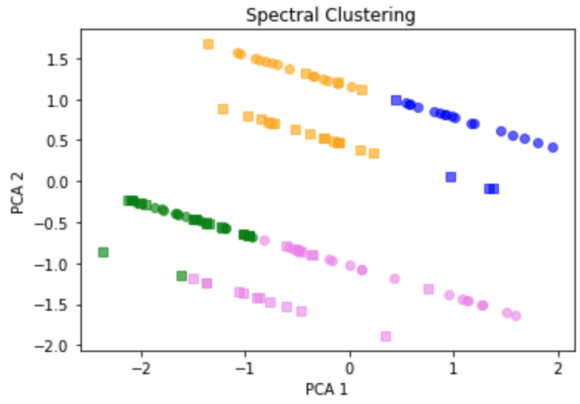

Figure 5

Figure 5 shows the use of spectral clustering with 4 clusters, represented by pink, blue, orange, and green. Data points that are boxes represent non-fraud cases and circles represent fraud cases.* Remark: All graphs and tables are subject to change in the final report*

The final clustering algorithm the students used was Aggolrmerative clustering. Agglomerative clustering is a form of hierarchical clustering that initializes each sample as a cluster, merges the nearest two clusters into a new cluster, and repeats until one cluster remains. The clustering results can be seen in Figure 6.

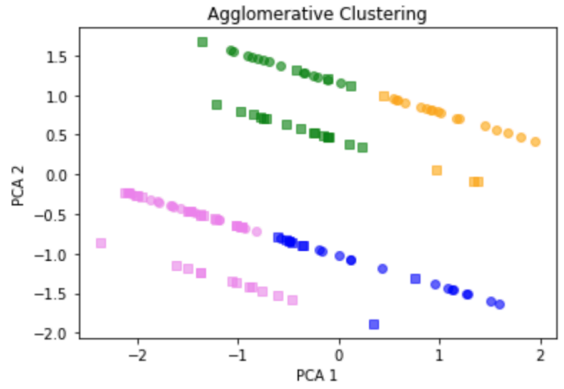

Figure 6

Figure 6 shows the use of Aggloermative clustering with 4 clusters, represented by pink, blue, orange, and green. Data points that are boxes represent non-fraud cases and circles represent fraud cases.* Remark: All graphs and tables are subject to change in the final report*

The last task performed was the evaluation of the clusters. The students used silhouette scores to measure the integrity of the clusters. Silhouette score measures how closely related the data points are to their assigned cluster, and how dissimilar they are to other clusters. The range of silhouette score is between -1 and 1, with a value closer to 1 meaning better clustering results. According to Figure 7, the k-means algorithm had the best partitioning, followed by agglomerative clustering, and lastly spectral clustering.

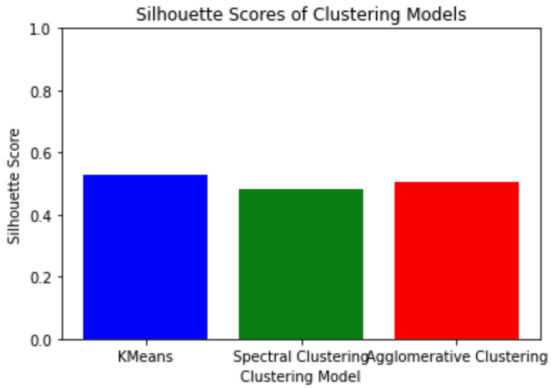

Figure 7

Figure 7 shows a bar plot of the silhouette scores for each of the clustering models. Blue is the score for k-means, green for spectral clustering, and red for agglomerative clustering. * Remark: All graphs and tables are subject to change in the final report*

To conclude, what hasn't yet been done is the interpretation of the clustering results, identifying ranges of susceptibility in fraudulent transactions, and the Final Report. Throughout this pipeline, the students have been able to select a dataset, data explore, preprocess, engineer, perform a selection from the raw data, and perform evaluation metrics. Despite the low dependencies, there is still much knowledge and insight to be gained from this fraud dataset, which the students will continue to analyze. Advantageously, this dataset is the result of a thesis written by the author, so the students will also read parts of the thesis for guidance. In the following week, the students will interpret the clustering results, Identify ranges of susceptibility in fraudulent transactions, and start on the Final Report.