Jerry Cheng

**CS-559: Project 1**

**Introduction:**

This report is a description on the steps I took in order to complete the Project 1 for CS559: Machine Learning and Fundamentals. This project involved clustering and dimension reduction on a mystery dataset. Using a given data set S21\_P1\_trainset.csv, the goal was to reduce the size of the dataset in order to prepare it for modelling. The dataset contained 200 features, with 200,000 rows. With the exception of the first column ‘ID\_code’, the rest of the columns were made up of unlabeled data. The process by which I took in order to complete this assignment can be summed up in four basic steps: exploratory data analysis, feature engineering & extraction, clustering analysis, and finally dimension reduction. Exploratory data analysis refers to the approach that is used to analyze the data sets to summarize their main characteristics. Exploratory data analysis is also used to gain a better understanding of the dataset. Feature engineering is the process of using the knowledge gained in the exploratory data analysis stage in order to extract new features from the old data, which will help to improve the performance of any machine learning algorithms. Cluster analysis is the task of grouping the data in such a way that objects in the same group (known as clusters), are more similar to each other than to those in other groups (clusters). This allows us to find similarities between the observations. Finally, dimension reduction refers to transforming data from a high-dimensional space into a low-dimensional space, such that the low-dimensional representation retains some meaningful properties of the original data.

**Methodology:**

As previously noted, the process by which was taken in order to complete this assignment can be summed up in four basic steps: exploratory data analysis, feature engineering & extraction, clustering analysis, and finally dimension reduction.

The first objective in completing this assignment is to first gain a better understanding of the structure and content of the dataset. To do this, many different basic inspections had to be made, including the shape of the dataset, the head/first 10 features of the dataset, the data frame info with the info function, and some basic statistics of the data frame through using the describe function. Ultimately, doing this inspection resulted in the information the trainset.csv was a dataset of 200000 rows and 201 columns. From there, the next step is to ensure that the dataset contains no missing values, and if so, handle the missing data. In this circumstance, the dataset contained no missing data, so no further action had to have been taken. The next step is to check the correlation of the dataset. To accomplish this, the dataset was run through the correlation function, and graphed out onto a correlation heatmap. The correlation heatmap and correlation table indicated that amongst the various variable features in the dataset, there was no significant correlation. Therefore, it was concluded that extracting data could not be accomplished with the result of corr(). The next step that was done was to analyze the distribution of the data through the use of different histograms. The first histogram was created simply by mapping the various feature data into a histogram. This first histogram concluded that amongst the various ranges of the features, the features were mostly normally distributed. The next histogram was created by layering all of the different histograms on top of each other. From this histogram, we could see that the vast majority of the data seems to fall between the range (-25, 25). The next histogram was created similarly to the first histogram, with the range (-25, 25) bound. From that histogram, it was determined that some of the variables have an extremely low standard deviation or variance, and some of the variables have an extremely high deviation or variance.

Upon completing the exploratory data analysis, it was determined that there is no inherent correlation between any of the initial data frame features. As such, it would not be possible to eliminate features based on corr() alone. It was then decided to drop features based on whether or not the features’ rounded means and standard deviations were identical. To achieve this, a new data frame copy was made from the original data frame’s discrete statistics:

Python code: (newDF = pd.Dataframe(DF.describe().iloc[1:3]).

With this new data frame, the means and standard deviations were rounded. Following that, the data frame was transposed. Duplicate or identical rows were dropped, and then the data frame was transposed back. The overall effect this process had was creating a data frame with the redundant features with a very close / identical distribution being eliminated from the data frame. Variable features that were not present in this new data frame were eliminated from the old data frame and piped into a new data frame, removedDF. The end result was a data frame removedDF that had duplicates removed, effectively reducing the number of features present in the dataset from 200 to 139.

The next step that was taken was to perform kmeans clustering analysis in order to find out the similarities between observations. The first step that was performed in this step was to put the data frame, removedDF, into the plot\_inertia function. This function creates an elbow plot, which will help to determine the idea number of clusters for the kmeans clustering analysis. The function resulted in an elbow plot that showed that the ideal number of clusters would be two. The kmeans clustering was then performed on the data, making two clusters. After performing the clustering analysis, discrete mathematical statistics was performed on the clusters. Upon doing this, it was determined that further trimming of the data was possible by removing features that have the same rounded standard deviation and means, meaning that redundant clusters with the same distributions would be removed from the data frame. The aforementioned features were removed similarly to before. Now, through observing the similarities between clustering observations, the feature count has been reduced down to 123 columns. Finally, principal component analysis was performed on the final data frame.

**Results:**

The results from the principal component analysis was that in order to retain 95% of the original data, we could reduce the dimensions down to 117 principal components. This was found by running and fitting principal component analysis on a data frame that had many redundant features stripped away from the original data set.

**Discussion & Conclusion:**

The results found was the result of applying the aforementioned methodology on the S21\_P1\_trainset.csv dataset. There are many other alternative ways to perform clustering and dimension reduction on the dataset. One such idea would be to instead of looking at the redundant distribution through looking at the mean and standard deviation, to instead make use of ktesting to see if two features had a similar gaussian distribution. In other words, we could compare the various features’ gaussian distribution instead of simply looking and comparing the various features’ standard deviations and means. Upon looking at the distributions, the ranges could be analyzed and rows with outliers could be eliminated. Therefore, instead of performing feature engineering through the process detailed above, this would be a different way at performing feature engineering.

Throughout this analysis, I did not perform any elimination or comparison of specific rows in the data set. Upon looking back at the methodology, if changes were to be made, this would be one of the primary changes to be made. Looking at and performing comparisons on the individual rows could contribute to creating a more succinct and effective dataset. The usage of gaussian distribution analysis could contribute to eliminating outliers within the individual rows.

To conclude, after completing this analysis, this author would like to suggest and propose to the data provider that an appropriate final dimension would be 117 features, without any modifications to the individual rows.