Math 461 - Final Project

2024-04-23

## 1. Dataset

* Your dataset should be selected with a target variable in mind. The target variable is the variable you will build models to predict.
* Describe your dataset. What is the dataset about? What does each row in the dataset present? How many variables/rows in the dataset? What is the source of the dataset?

## 2. Data Visualization

* Give at least 5 meaningful visualizations/graphs for the dataset. Give your comment for each graph.
* The visualization could be done in Excel, SAS or any other software. This video illustrate how to do visualization in SAS Enterprise Guide.

## 3. Unsupervised Techniques

Unsupervised learning Techniques are techniques used to discover the underlying structure of a dataset. This section will ask you to do the unsupervised learning techniques covered in the class including clustering (k-means and Hierarchical), Principal Component Analysis and Factor Analysis.

### 3.1 Principal Component Anaysis (PCA)

* Explain PCA
* Implement PCA on your dataset
* Plot the scree plot of the percentage of the total variances captured in the PCs.
* How much variance is captured by the first two PCs? by the first PC?
* What are the contribution of the original variables in the first PC?
* How many PCs are needed to capture at least 90% of the original dataset?

### 3.2 K-means Clustering:

* Explain k-means clustering
* Implement k-means clustering on the dataset
* Plot the total sum squares within clusters and use the elbow method to decide the number of clusters.
* Visualize the data with the selected number of clusters. Give your comments on the clustering results.
* Visualize the data using PCA colored by clusters.
* Report the cluster means

### 3.3 Hierarchical Clustering

* Explain Hierarchical clustering
* Explain the different between linkages
* Apply Hierarchical clustering with the complete linkage. Plot the dendrogram.
* Plot the total sum squares within clusters and use the elbow method to decide the number of clusters.
* Replot the dendrogram that with the clusters boxed
* Visualize the data using PCA colored by clusters.

### 3.4 Factor Analyis

* Explain factor Analysis
* Run factor analysis on the dataset with the number of factor being 2
* How many variance of the dataset are explained by the model? Which variables are least relevant to the factor model?
* What variables are affected by factor 1? What variables are affected by factor 2? Could you name the two factors?
* Rerun the analysis with your own selected number of factors.

## 4. Modelling

Sample Codes

### 4.1. With Binary Target

In this section, you are asked to run several models to predict your selected target variable. Your target variable should be *binary.* If you do not have a binary variable in the dataset, you can create a binary variable from an interested numeric variables.

The models are as follows.

* Generalized Linear Model (Logistic Regression)
* Linear Discriminant Analysis
* Quadratic Discriminant Analysis

For each of the model, do the follows.

* Explain the model
* Implement the model
* Report the training accuracy of each models.
* Comment on the final results. What is the best model in term of the training accuracy?

### 4.2. With Numeric Target

In this section, you are asked to run several models to predict your selected target variable. Your target variable should be *numeric*.

The models are as follows.

* Linear Regression
* PCA Regression
* Generalized Linear Model (Poisson Regression)

For each of the model, do the follows.

* Explain the model
* Implement the model
* Report the each models.
* Comment on the final results. What is the best model in term of

## 5. Presentation

* Introduce the dataset: where/how do you get the data, how many observations; how many variables. Description of variables. How did you clean the data?
* Present the results of the implementation of techniques in the written component. Give give your comments on each results.