CSCI 6223 Practical Data Science

Homework #2

Due 11/08/2023

In this homework, you will use ***Principal Component Analysis***.

You will work together with your team to solve this problem.

Turn in one pdf document for the whole team!

PCA Best Practices

There are a few things to keep in mind when using PCA.

PCA only works with **numeric features**, like continuous quantities or counts.

PCA is **sensitive to scale** so you it's good practice to standardize your data before applying PCA.

Remove **outliers** since they will skew the results.

The Business Problem

You are a data scientist working at a startup electric vehicle company. Your new electric vehicle is called the SUN and will be released to the public at the end of the year. Your company has been awarded a government grant if you set the price of your vehicle comparable to gas powered vehicles with similar performance. Unfortunately, "similar performance" has not been specified so you don't know where to set the price. All you have is a set of performance and price data for gasoline powered vehicles manufactured last year.

Project Deliverables (100 points)

1. Using the auto_data.csv file and the PCA best practices above,
    a. Determine the number of principal components that would explain 90% of the variance.
    b. Include a plot showing the explained variance versus the number of dimensions.
    c. Include reasoning for your decision.
2. Build two linear regression models to predict the price of the car.
    a. The first regression model should predict the price of the car using the following features.
        i. 'num-of-doors','wheel-base','length','width','height','num-of-cylinders','engine-size','peak-rpm','city-mpg','highway-mpg'
    b. The second regression model should predict the price of the car using your selected principal components.