The More, the Better?

Dimensionality Reduction Through Exploratory Data Analysis

Xiaozhu Zhang

I.Introduction

In today's data-driven world, we are often overwhelmed by the enormous amount of available information and data. As the size of the datasets grow larger and larger, more and more complicated machine learning techniques and modeling methods emerge as a result. However, when we "feed" a dataset directly(without any processing) to a complicated model such as a neural network, more than often it takes a long time to run the code, and we end up obtaining an unsatisfying result. Therefore, this paper aims to illustrate instead of "feeding" a dataset directly to a complicated algorithm, simple techniques such as exploratory data analysis and dimension reductionality not only can give us some insights about the data, but also generate better performance in terms of prediction.

II. Method

This paper will be divided into two parts; each part will be analyzing a different dataset. Part one will focus on the dataset *Seed* and Part two will focus on the dataset *Automobile*. First, we will be learning more about the patterns that exist in the dataset through some exploratory data analysis, after that we will see how Principal Component Analysis reflects the pattern we observed. Lastly, we will compare the prediction performance between the original dataset and the dataset with reduced dimensionality to see whether the process of dimensionality reduction hinders or enhances the prediction performance.

III.Part One

A. Dataset: Seed

   a. This dataset contains 7 numerical attributes(predictive) and 1 discrete attribute(outcome). Each of the predictive attributes is a measurement of a specific kernel and the outcome is the corresponding kind of kernel. No imputation is needed since there are no missing values in this dataset. In total, there are three labels(1,2,3), each label corresponds to a specific kind of kernel.

B. Procedure

a. Correlation between predictive attributes

A preliminary way to obtain an overview of the correlations between the attributes is to plot a correlation heatmap. As Figure 1 shows, it appears that a number of predictive attributes are correlated with one another. For example, *area* and *length of kernel(len_kernel)* is highly correlated with the other variables.

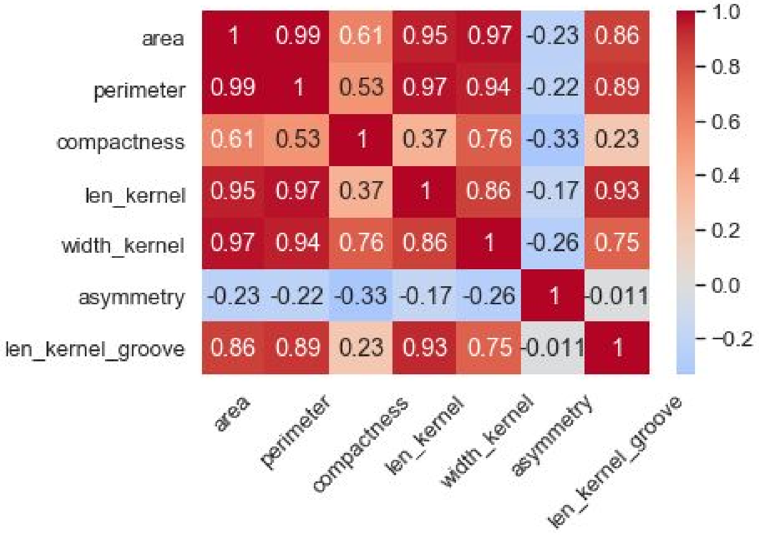


Figure 1: Correlation Heatmap of All Predictive Attributes

b. The relationship between predictive attributes and targets (labels)

After a brief glance at the correlations between the predictive attributes, now we will take a look at how these attributes are associated with the outcomes(labels). We can see in Figure 2 that the boxplot for *area, perimeter, length of kernel, width of kernel and length of kernel groove* share a similar pattern, which correspond to the high degree of correlation we observed in Figure 1. This naturally leads to the questions of whether all of the attributes are needed, and is it possible to “compress” this dataset to make it more “precise.”

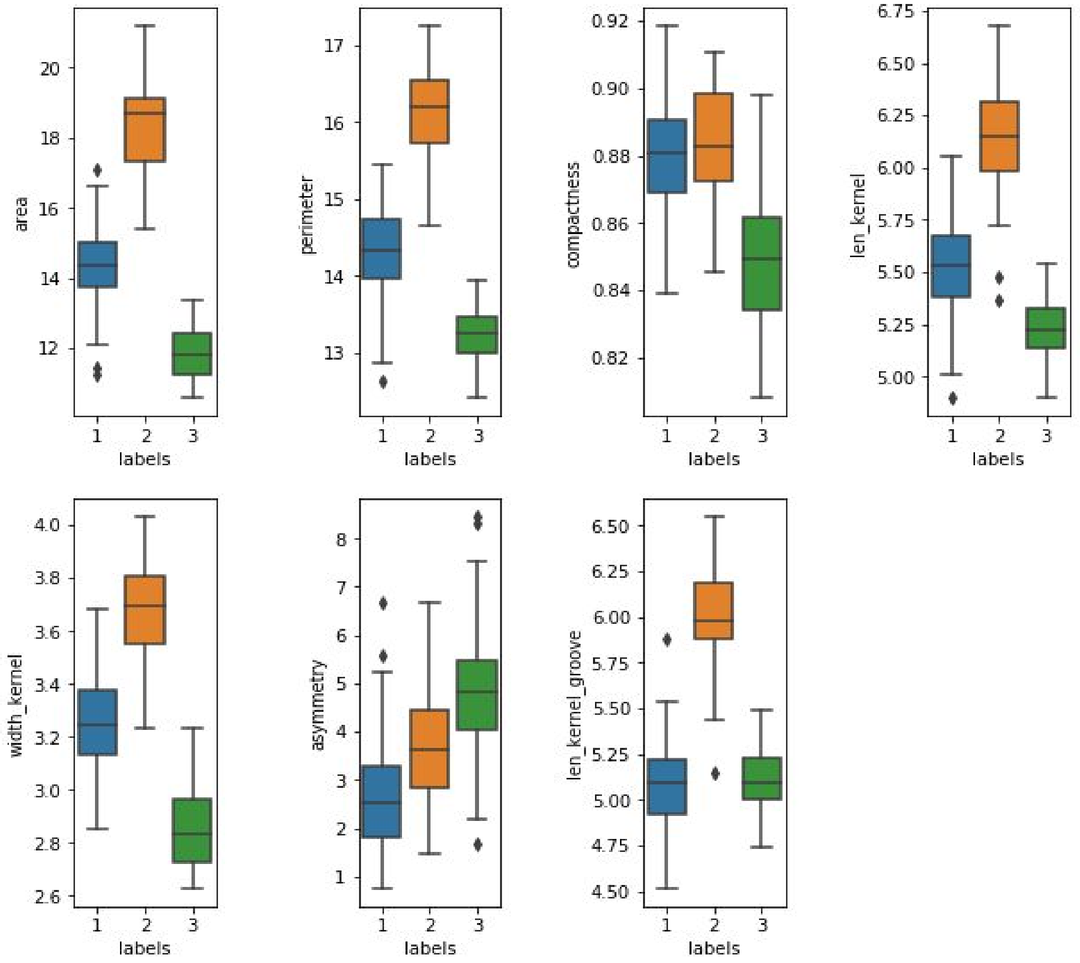


Figure 2: Boxplot: Each Predictive Attribute by Kinds of Kernel

c. Dimensionality Reduction by Principal Component Analysis(PCA)

Since the attributes of the dataset are all numerical, we can apply a common dimensionality reduction technique PCA to this dataset after the procedure of standardization(we standardize the predictive attribute to eliminate the effect of different units.) Figure 3 shows the explained variance ratio as we increase the number of components. Based on this plot, we can see that the first PC(principal component) explains 70% of the variance and the second PC explains approximately 20% of the variance. In other words, it is possible to keep 90% of the variances with only two PC's.

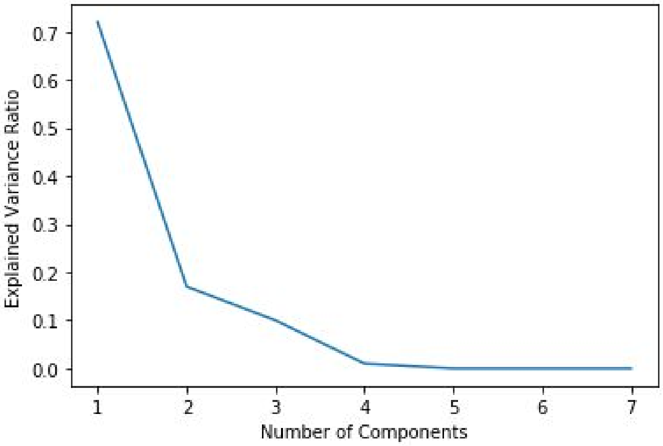
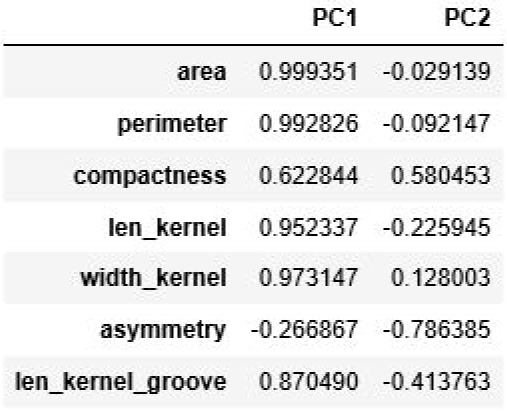


Figure 3: Percent of Variance Explained By Number of Components (Non-cumulative)

Table 2 shows the loadings(coefficients) by attributes and by principal components. It seems that PC1 well captures the correlation we observed in Figure 1, while PC2 focuses more on the contribution of asymmetry and compactness(which are less correlated with the other attributes.)



	PC1	PC2
area	0.999351	-0.029139
perimeter	0.992826	-0.092147
compactness	0.622844	0.580453
len_kernel	0.952337	-0.225945
width_kernel	0.973147	0.128003
asymmetry	-0.266867	-0.786385
len_kernel_groove	0.870490	-0.413763

Table 1: Loadings of the First Two PC's

Now that the original dataset is reduced to two dimensions, we can plot the transformed data to see how well the three groups are separated from one another. Based on Figure 4, it seems that the three groups are separated well with only a few exceptions.

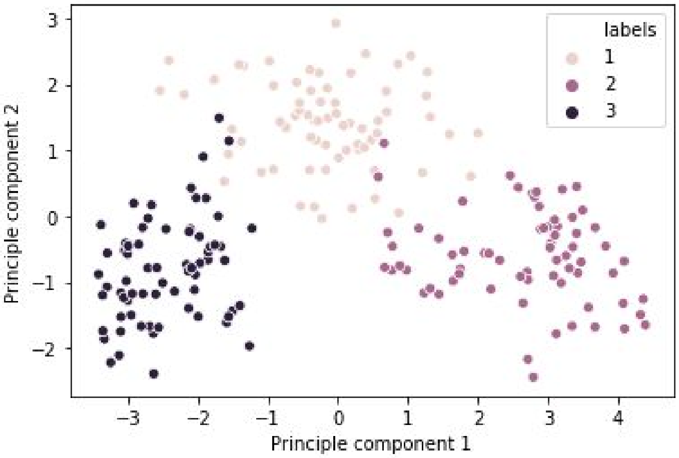


Figure 4:Visualization of the First Two PC's

d. Prediction Performance

Furthermore, if we try to build a classification model with logistic regression, the dimension-reduced dataset also generates a better result with a lower error rate(error rate: misclassified instances.) The result summarized in Table 2 is based on the average error rate of ten trials.

	Dimension-Reduced Dataset	Original Dataset
Dimension	2	7
Error Rate	0.03	0.08

Table 2: Error rate: Using Dimension Reduced Dataset and Original Dataset

IV. Part Two

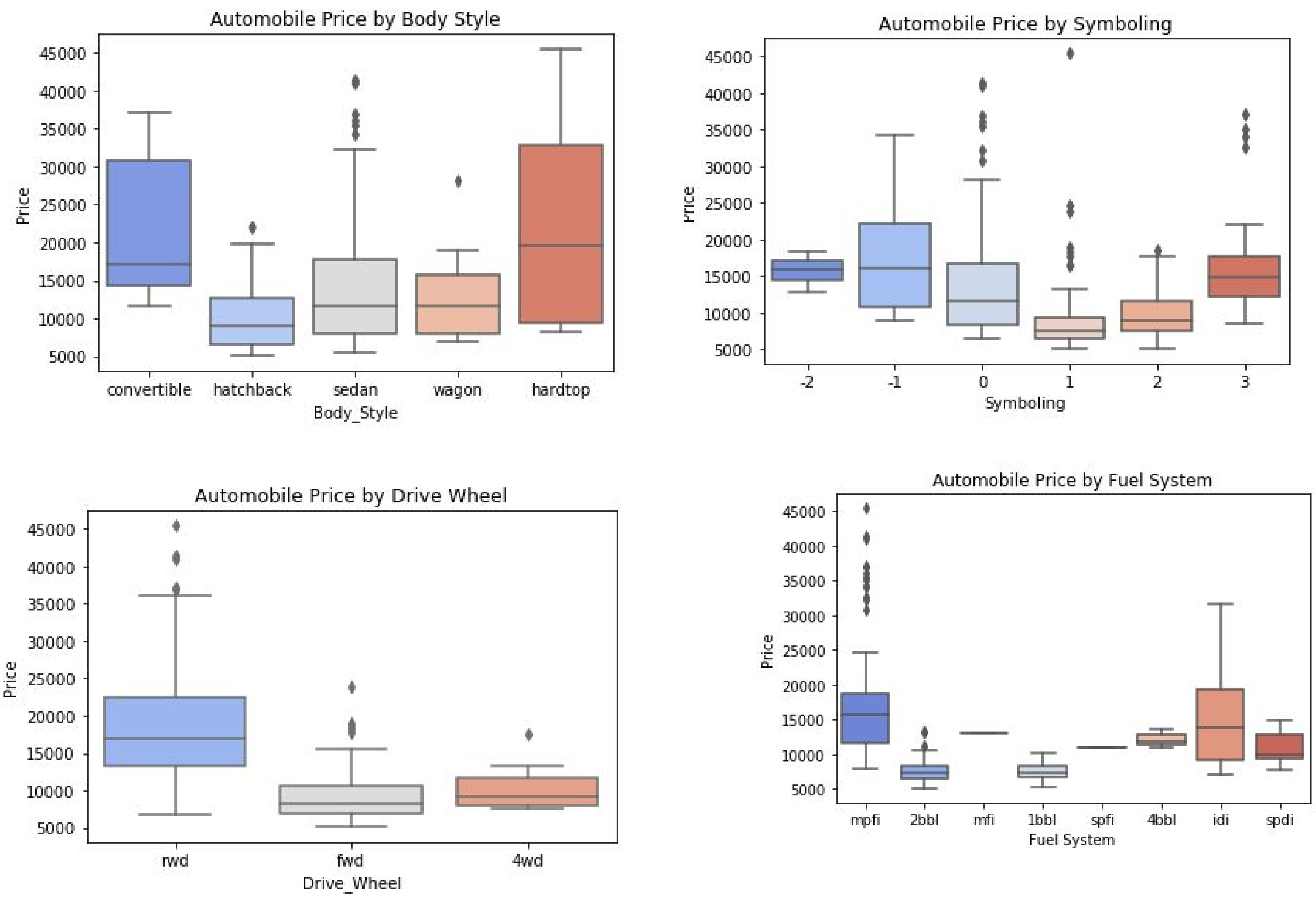
A. Dataset

a. This dataset contains 15 numerical attributes(14 predictive,1 outcome) and 11 categorical attributes. Each of the predictive attributes is a measurement of an individual automobile and the outcome is the corresponding price. Since there are missing values in this dataset, imputation is needed. Specifically, missing values in the numerical attributes are filled with mean and the missing values in the categorical attributes are filled with mode.

B. Procedure

a. Categorical variable

We will first examine the categorical variables in this dataset to capture their association with the outcome variable(price). In order to do this, boxplots of each categorical variable versus price are plotted, and the boxplots presented in Figure 5 are the ones that show obvious patterns.



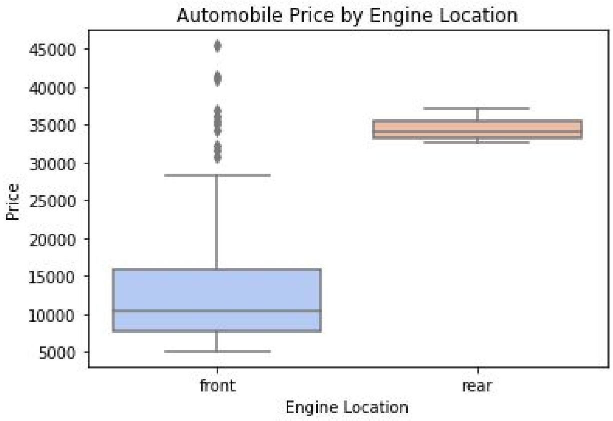


Figure 5: Boxplot for prices by groups in each attribute

b. Correlation between (numerical) predictive attributes and their association with outcome

Similar to Part One, the correlation heatmap is plotted as shown in Figure 6. It seems that the attributes in this dataset are less correlated with each other compared to the dataset in Part One. However, some of them are still fairly correlated and based on the scatterplot shown in Figure 7. Therefore, we will perform the same procedure(PCA) to the numerical attributes to see if it can be "compressed" into a lower dimension.

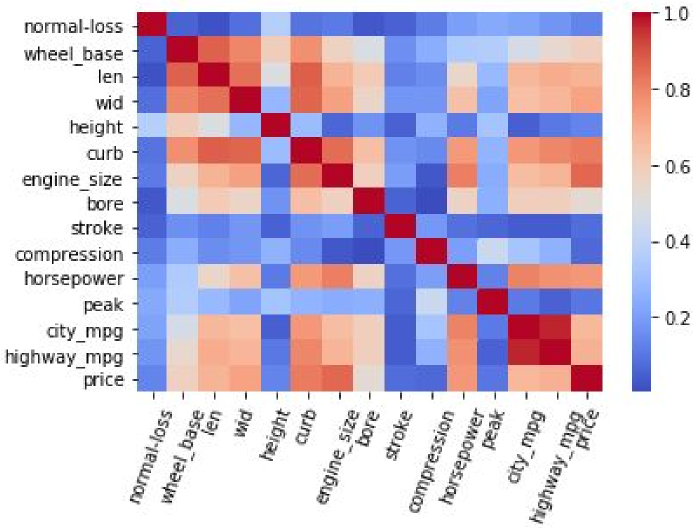


Figure 6: Correlation Heatmap of All (numerical) Predictive Attributes

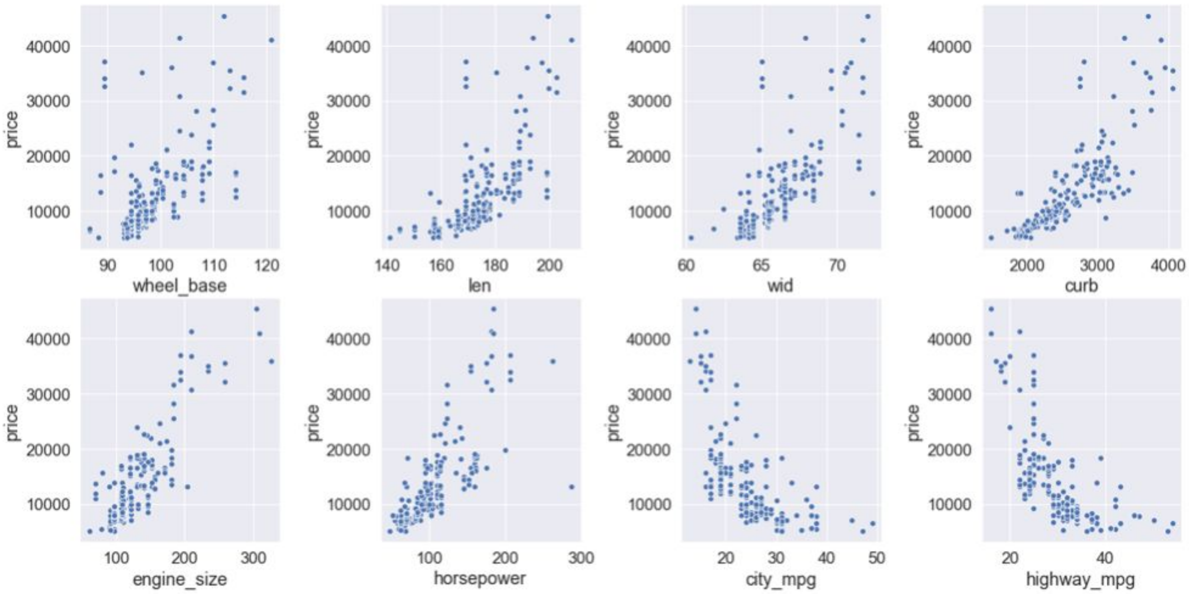


Figure 7: Scatterplot of All (numerical) Predictive Attributes vs Price

c. Dimension Reduction by Principal Component Analysis(PCA)

By applying PCA to this dataset after the procedure of standardization, we can see from Figure 8 that the first PC(principal component) explains 50% of the variance, the second PC explains approximately 20% of the variance. The explained variance ratio keeps dropping and stabilizes at the 6th PC. In other words, it is possible to maintain most of the variances with 6 PC's.

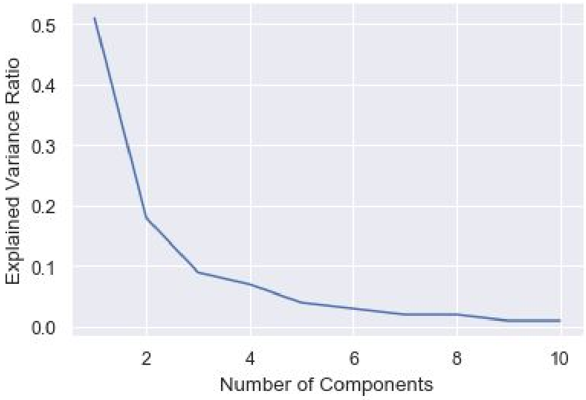


Figure 8: Percent of Variance Explained By Number of Components (Non-cumulative)

Figure 9 shows the loadings(coefficients) of the first six PC's. Similar to what we observe in Part One, the first PC well captures the correlation between the numerical variables and the rest of the PC's focus on variables which are less correlated to the outcome.

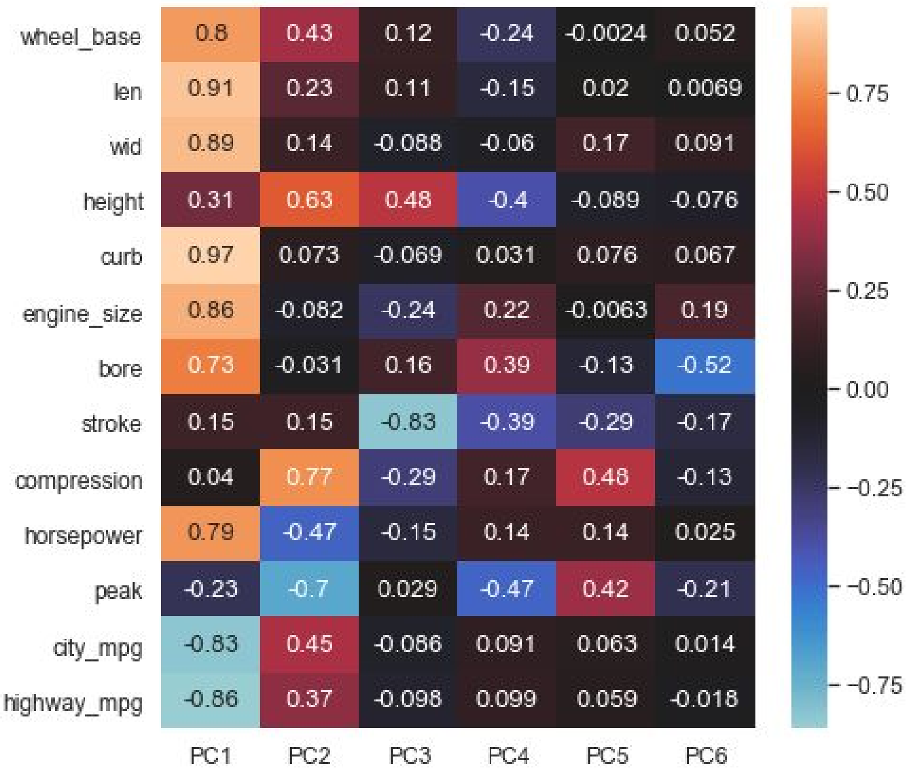


Figure 9: Loadings of the First Six PC's

d. Prediction Performance

If we try to build a classification model with a regression tree, the dimension-reduced dataset also generates a result with a fairly close score(score:ratio of reduced error.) The result based on the average score of ten trials is summarized in Table 3.

	Dimension-Reduced Dataset	Original Dataset
Dimension	11(6 num+5 cat)	25
Score	0.82	0.85

Table 3: Score: Using Dimension Reduced Dataset and Original Dataset

V. Discussion

In summary, sometimes it is plausible to apply dimensionality reduction techniques to a dataset with attributes that are obviously correlated with one another. The higher the correlation, the better we are able to preserve the information contained in the original dataset. However, projecting the dataset onto a lower dimensional space may make it difficult to interpret the new attributes(the transformed data by PCA.) Therefore, the decision of applying dimensionality reduction method depends on whether the purpose is prediction or interpretation. Furthermore, there are also other dimensionality reducing methods such as model selection, and shrinkage which can balance the issue between interpretation and prediction but are not discussed in this paper.

Reference

1. Processing Seed data: Piazza thread 31 by Shozen Dan.
2. Visualization: https://seaborn.pydata.org/index.html
3. for the code of statistical method used(PCA, logistic regression, regression tree)

   https://scikit-learn.org