Midterm Project

(R for Data Scientists)

Jaya Padma Sri Maddi (jxm166230)

Venkata Kartheek Madhavarapu (vxm153830)

Arpita Mothukuri (axm163631)

Lakshmi Priyanka Parimi (lxp160730)

PART 1 – Data Wrangling

Title: Assignment #Data Wrangling

Purpose: The purpose of this part is to perform data wrangling on "**RegularSeasonDetailedResults**" dataset to obtain statistics for each team in every season.

Datasets: RegularSeasonDetailedResults dataset has been used for this part. "RegularSeasonDetailedResults" dataset contains details regarding game results covering seasons 2003-2016.

Approach:

- First, we loaded RegularSeasonDetailedResults dataset.
- Next we created separate dataset for winning team by selecting season,wteam,wscore and the remaining statistics columns.
- Then we created one more dataset for losing team by selecting season,lteam,lscore and the remaining statistics columns.
- Then we renamed column names for both the datasets so that both of them have same column names.
- Next, we combined both the datasets using rbind and applied group by to get cumulative statistics for each team for every season.

PART 2 - Clustering

Title: Assignment #kmeans#PCA

Purpose: In this part we need to perform clustering by applying **kmeans** on the wrangled dataset obtained in Part 1.

Datasets: Dataset obtained by performing wrangling on "**RegularSeasonDetailedResults"** dataset in part 1 has been used for this part.

Approach:

- First, we applied kmeans on wrangled dataset with 8 clusters. We found between_ss / total_ss to be 59.5%. We tried out with different number of clusters and found out that using 8 clusters gave optimal value for between_ss/total_ss.
- Then we scaled the data and applied kmeans again with 8 clusters. We found between_ss / total_ss to be 43.4%. (Refer to graph #1)
- Next,we applied PCA on the scaled data followed by PVE. We applied elbow method and observed that 8 is the optimal number of clusters. (Refer to graph #2)

PART 3 - Regression

Title: Assignment #Linear Regression#Ridge#Lasso#PCR#KNN

Purpose: In this part we need to apply regression techniques **Linear Regression, Ridge, Lasso, PCR and KNN** to predict the score for each team in every tournament.

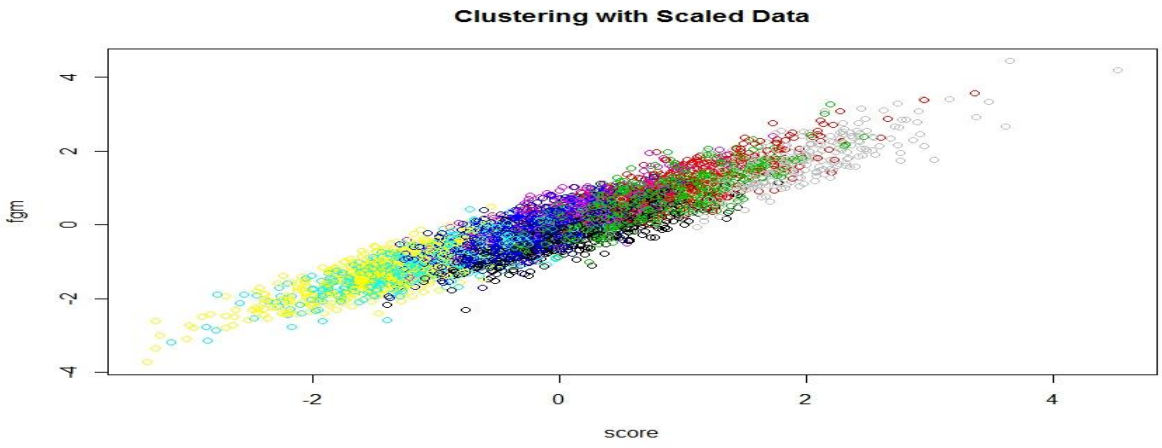
Datasets: Two datasets have been used for this assignment. Wrangled data in Part1 is combined with **"TourneyCompactResults"** dataset for regression part. "TourneyCompactResults" dataset contains game-by-game NCAA tournament results for all seasons of historical data.

Approach:

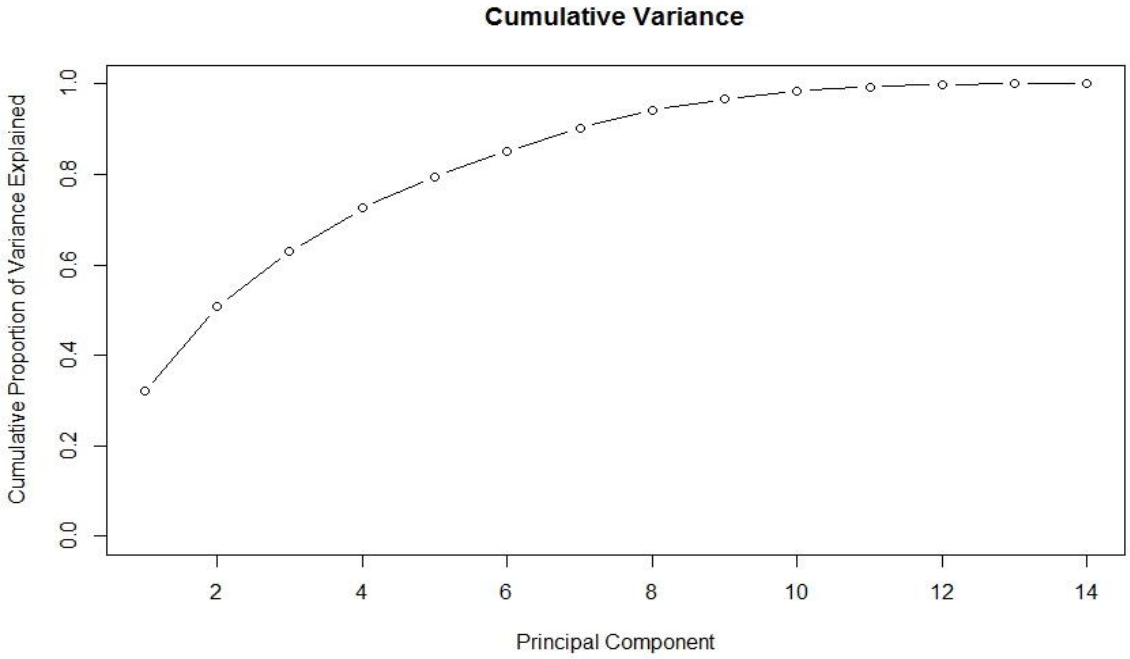
- First, we did wrangling on TourneyCompactResults data to get season, team and score columns. Then we merged this data with "RegularSeasonDetailedResults" wrangled data set. Prior to merging "score" column of wrangled dataset in part 1 is dropped. The resultant dataset consists of season, team, statistics and score.
- Next, we applied linear regression and cross validation on the dataset and found out the RMSE value to be 11.06137.
- Then we applied ridge and found out the best lambda value to be 0.4926468. RMSE of ridge is 11.03713. (Refer to graph #3)
- Next, we applied Lasso and found out RMSE value as 11.03128. (Refer to graph #4)
- Next, we applied PCR by dividing the dataset into training and test data. We figured out the RMSE value to be 10.9596 (Refer to graph #5).
- Next, we applied KNN to figure out the RMSE value to be 12.19363.

GRAPHS:

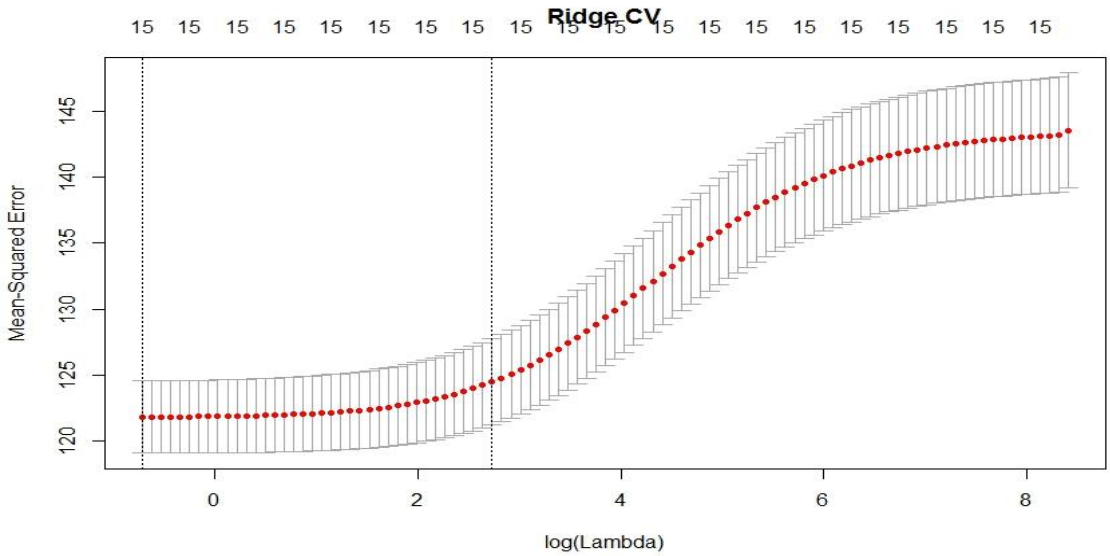
1.Clustering with Scaled Data



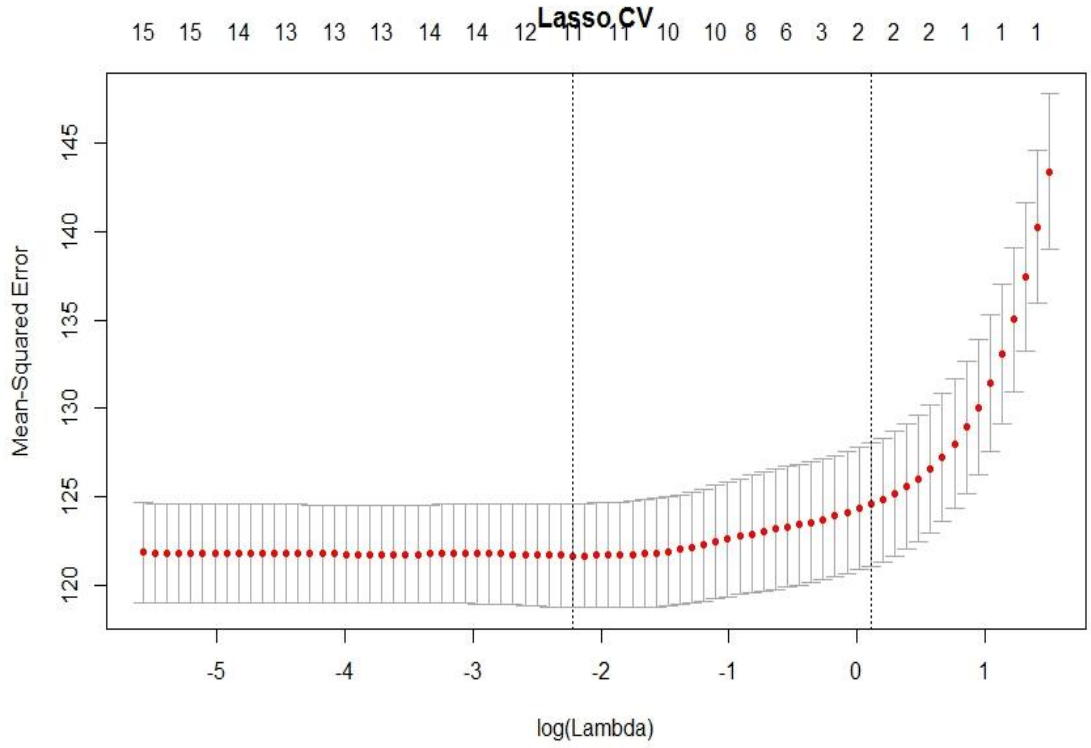
2.PCA


Cumulative Variance
Cumulative Proportion of Variance Explained
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10
12
14
Principal Component

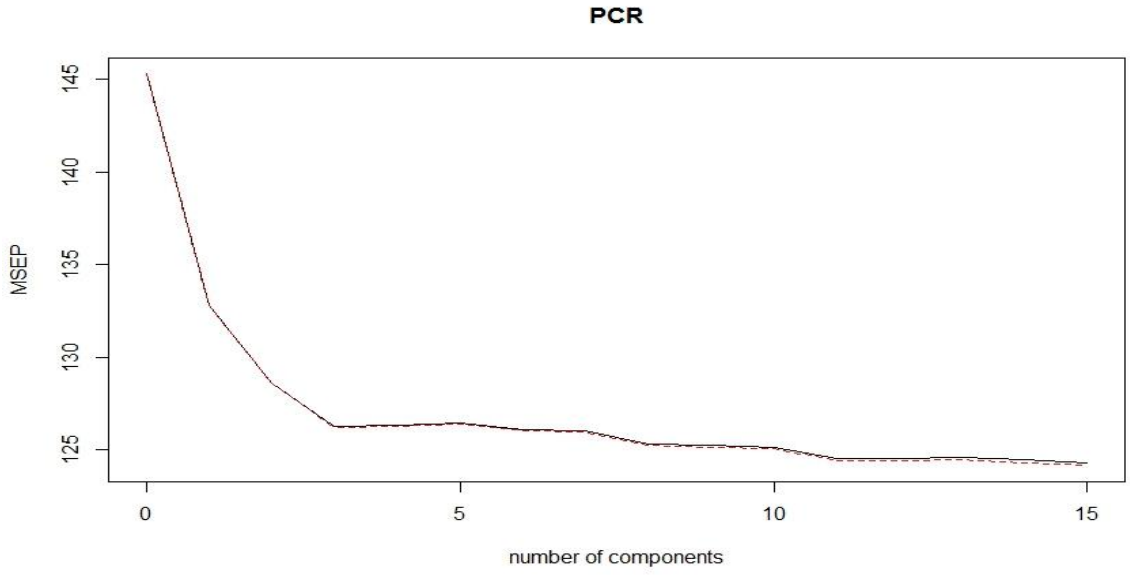
3.RIDGE


Ridge CV
15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15
Mean-Squared Error
120
125
130
135
140
145
0
2
4
6
8
log(Lambda)

4.LASSO


Lasso.CV
15 15 14 13 13 13 14 14 12 11 11 10 10 8 6 3 2 2 2 1 1 1
Mean-Squared Error
120
125
130
135
140
145
-5
-4
-3
-2
-1
0
1
log(Lambda)

5.PCR


PCR
MSEP
125
130
135
140
145
0
5
10
15
number of components

Summary:

Given below are couple of points which we have learned in this project:

- Without scaling the results are bad for pca.
- The quality of results in decreasing order is 'without scaling and without pca', 'pca with scaling', 'pca without scaling', 'with scaling and without pca'.
- We used elbow method to find the optimal number of components.
- Reducing dimensionality helped us achieving a better regression model.
- PCR had given the best results among all the regression techniques applied.