Hou-Hung Liu, Hao-Zhi Gan, Yu-Xuan Wang, William Phan | Group 11

Data Preparation

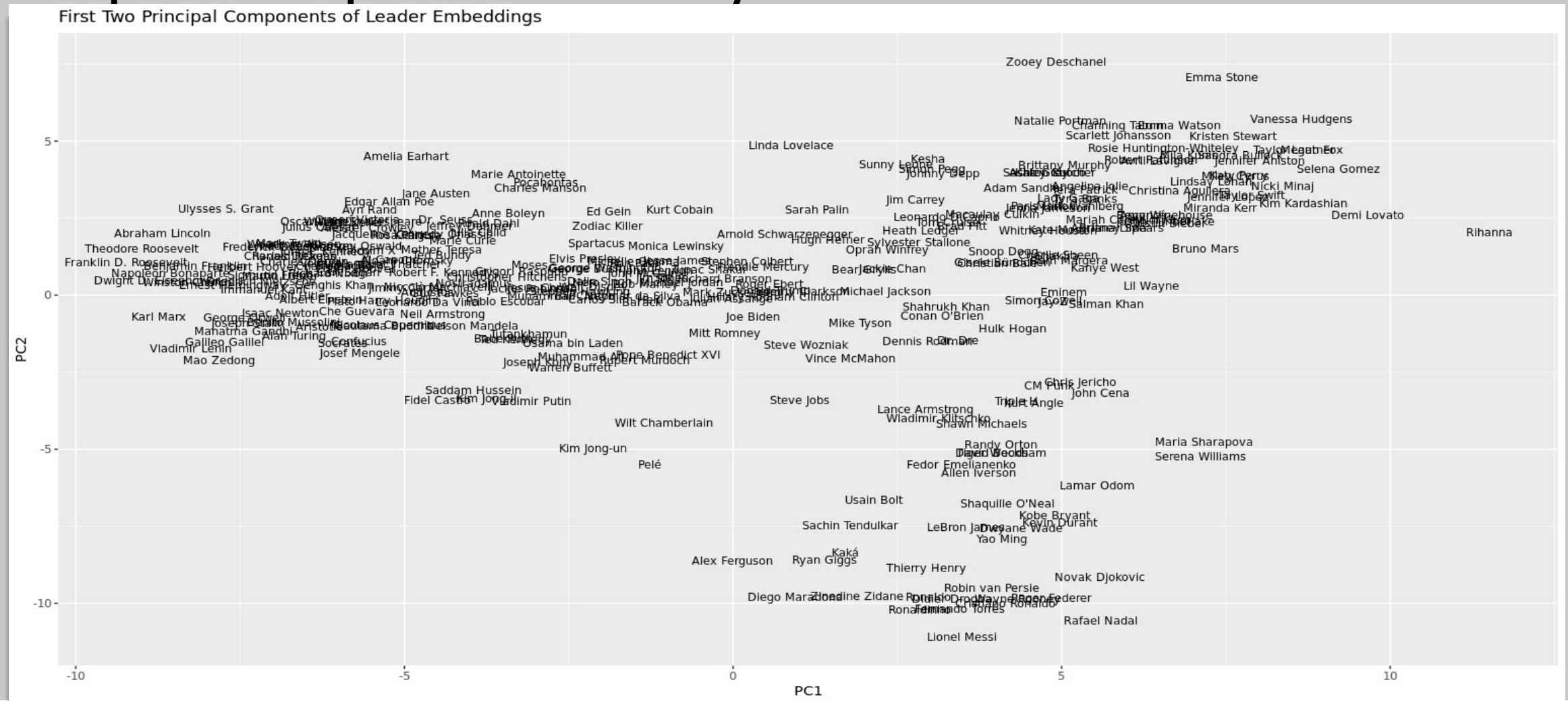
- Drop out leaders who cannot be found in the pre-trained word embeddings.
- Merge the leader's embeddings into the data frame.
- Group by the leader and calculate their average scores.

Method

- Using two methods:

1. Analysed by total individual survey responses
   - Principal Component Analysis
   - Multiple linear regression
2. Analysed by average rating of leaders
   - Ridge Regression & Lasso Regression
   - Elastic Net Regression

Principal Components Analysis



We select 24 principal components, each variance explained are higher than 1%. The total proportion of variance explained are higher than 50%.

Summary

From this plot, most of the politicians are clustered at left top side; athletes are found at the bottom, and singer at the right top according to their features represented by PC1 and PC2.

Multiple Linear Regression

Model:

$$Value = \beta_1 race\ African\ American_i + \beta_2 raceOther_i + \beta_3 raceWhite_i + \beta_4 genderMale_i + \beta_5\ genderOther_i + \beta_6 Political\ Orientation\ Other_i + \beta_7 Political\ Orientation\ Republican_i + \beta_8 PC_1 + \beta_9 PC_2 + \cdots + \beta_{31} PC_{24} + \varepsilon_i$$

Variable	Coefficients	Std. Error	t-statistic	p-value
Race: White	-4.06	0.97	-4.20	0.00***
Race: African American	-12.43	1.68	-7.38	0.00***
Race: Other	-9.99	1.48	-6.74	0.00***
Gender: Male	-2.17	0.63	-3.44	0.00***
Gender: Other	1.68	1.89	0.89	0.37
Political orientation: Republican	-0.91	0.96	-0.95	0.34
Political orientation: Other	-1.17	0.67	-1.75	0.08
Number of Observations: 7373 Adjusted R-squared: 0.24				
Note: ”if p-value<0.05, ” if p-value<0.01, ‘’ if p-value<0.001 , ‘***’**				

- This statistical model will help define which characteristics of participants and the features of leaders are significant to determine the evalutaion of leadership effectiveness.
- The results shows that the categorical variables ”Race” and ”Gender” are statistically signicant. Hence, these participants individual characteristics will affect the evaluation of leadership effectviness regardless of participants’ political orientation.

Different Regression Models and Metrics Analysis

- **Linear Regression**

The linear regression equation can be expressed in the following form:

$$Y = a_1x_1 + a_1x_2 + a_3x_3 + \cdots + a_nx_n + b$$

Where the following is true:

- Y is the average leadership effectiveness for correspondant leader.
- $x_1, x_2, x_3, \ldots, x_n$ are the features derived from word vectors for correspondant leader.
- $a_1, a_2, a_3, \ldots, a_n$ are the coefficients.
- b is the parameter of the model.

- **Lasso Regression**

Loss function = OLS + alpha * summation (absolute values of the magnitude of the coefficients)

In Lasso, the loss function is modified to minimize the complexity of the model. The modification is done by limiting the sum of the absolute values of the model coefficients (l1-norm)

- **Ridge Regression**

Loss function = OLS + alpha * summation (squared coefficient values)

In Ridge, the loss function is modified to minimize the complexity of the model. The modification is done by
adding a penalty parameter equivalent to the square of the magnitude of the coefficients.

- **Elastic Net Regression**

Elastic Net combines the properties of both Ridge and Lasso regression by penalizing the model using both the l2-norm and the l1-norm.

Different Regression Models and Metrics Analysis

Model	Linear Regression		Lasso Regression		Ridge Regression		Elastic Net Regression	
Data Sets	Training	Test	Training	Test	Training	Test	Training	Test
Root Mean Squared Error	4.395e-14	27.097	2.671	17.635	0.234	25.981	4.536	12.505
R-squared value	1.0	-1.148	0.973	0.090	0.999	-0.974	0.921	0.543

- Based on this table, we observed that there is a significant improvement of goodness of fit using Lasso, Ridge and Elastic Net Regression compared with the original linear Regression as there is an increase of R-squared value and a decrease of RootMean sqaured Error on test sets. In addition, the Elastic Net Regression model is preferred among all the models as it registered the lowest Root Mean Square Error (12.505) and highest R-squared value (0.543).