**Milestone 3 - PCA**

Principal component analysis is a method that aids in determining the importance of a dataset’s features. In this case, we wanted to see which of the features in our dataset, which collected and accumulated from a randomly selected pool of posts from the politics subreddit, were necessary and which features may be of little importance when it comes to classification.

Before I even started the PCA, I noticed there were some “features” in our set that would already most likely be irrelevant in terms of classifying our data as reliable or unreliable. Therefore, I eliminated features such as the post author’s username and id, the submission link of the post, etc. Ultimately, I decided to narrow it down to the features: *'author\_total\_karma', 'author\_has\_verified\_email', 'author\_is\_employee', 'author\_is\_mod', 'author\_link\_karma', 'num\_comments', 'score','upvote\_ratio'*, as they seemed to be the ones with generally measurable and conclusive data.

**Results**

*2 Principal Components*

**How much of our variance is explained?** - [0.318, 0.218] = .536

**Which features matter most?**

***Features* =** *['author\_total\_karma', 'author\_has\_verified\_email', 'author\_is\_employee', 'author\_is\_mod', 'author\_link\_karma', 'num\_comments', 'score','upvote\_ratio']*

PCA 1 - ['0.631', '0.251', '0.000', '0.011', '0.635', '0.002', '0.324', '0.176']

PCA 2 - ['0.005', '0.067', '0.000', '0.169', '0.028', '0.763', '0.388', '0.483']

*3 Principal Components*

**Variance explained?** - [0.318, 0.218, 0.173] = .709

**Which features matter most?**

***Features* =** *['author\_total\_karma', 'author\_has\_verified\_email', 'author\_is\_employee', 'author\_is\_mod', 'author\_link\_karma', 'num\_comments', 'score','upvote\_ratio']*

PCA 1 - ['0.631', '0.251', '0.000', '0.011', '0.635', '0.002', '0.324', '0.176']

PCA 2 - ['0.005', '0.067', '0.000', '0.169', '0.028', '0.763', '0.388', '0.483']

**PCA 3** - ['0.149', '0.202', '0.000', '0.514', '0.197', '0.029', '0.526', '0.597']

*4 Principal Components*

**Variance explained?** -[0.318, 0.218, 0.173, 0.135] = .844

**Which features matter most?**

***Features* =** *['author\_total\_karma', 'author\_has\_verified\_email', 'author\_is\_employee', 'author\_is\_mod', 'author\_link\_karma', 'num\_comments', 'score','upvote\_ratio']*

PCA 1 - ['0.631', '0.251', '0.000', '0.011', '0.635', '0.002', '0.324', '0.176']

PCA 2 - ['0.005', '0.067', '0.000', '0.169', '0.028', '0.763', '0.388', '0.483']

PCA 3 - ['0.149', '0.202', '0.000', '0.514', '0.197', '0.029', '0.526', '0.597']

**PCA 4** - ['0.038', '0.698', '0.000', '0.648', '0.076', '0.108', '0.193', '0.191']

*5 Principal Components*

**Variance explained?** -[0.318, 0.218, 0.173, 0.135, 0.122] = .966

**Which features matter most?**

***Features* =** *['author\_total\_karma', 'author\_has\_verified\_email', 'author\_is\_employee', 'author\_is\_mod', 'author\_link\_karma', 'num\_comments', 'score','upvote\_ratio']*

PCA 1 - ['0.631', '0.251', '0.000', '0.011', '0.635', '0.002', '0.324', '0.176']

PCA 2 - ['0.005', '0.067', '0.000', '0.169', '0.028', '0.763', '0.388', '0.483']

PCA 3 - ['0.149', '0.202', '0.000', '0.514', '0.197', '0.029', '0.526', '0.597']

PCA 4 - ['0.038', '0.698', '0.000', '0.648', '0.076', '0.108', '0.193', '0.191']

**PCA 5** - ['0.293', '0.633', '0.000', '0.529', '0.207', '0.139', '0.340', '0.235']

**Conclusion**

The ultimate goal of a PCA is to find the number of components that can give a reliably high level of variance (information not lost from the original dataset). In this case, it took five principal components to get the variance to a truly safe spot (96%), though 4 principal components resulted in 84% variance, which is still decently reliable. When analyzing the actual features in the five principal component case, it seems that all but the author\_is\_employee features hold reasonably high importance in relation to one another. Therefore, the author\_is\_employee feature could reasonably be removed from our list.

|  |  |
| --- | --- |
| Feature | Importance |
| *Author\_total\_karma* | 1.116 |
| *Author\_has\_verified\_email* | 1.851 |
| *Author\_is\_employee* | 0 |
| *Author\_is\_mod* | 1.871 |
| *Author\_link\_karma* | 1.143 |
| *Num\_comments* | 1.041 |
| *Score* | 1.771 |
| *upvote\_ratio* | 1.682 |