Project Group 3

Elia Rezaeian, Nataly Jimenez, Matthew Kearney

May 3, 2024

CS 559 Spring 2024 Semester Project

**2 Company Characterization**

Using KMeans, we located clusters with similar characteristics. Based on the plot below, the Within-Cluster Sum of Squares plateaus after 6 clusters, indicating change beyond this point. Because our team consists of 3 members, we decided each member would analyze 2 clusters to predict in section 3. (Figure 1)

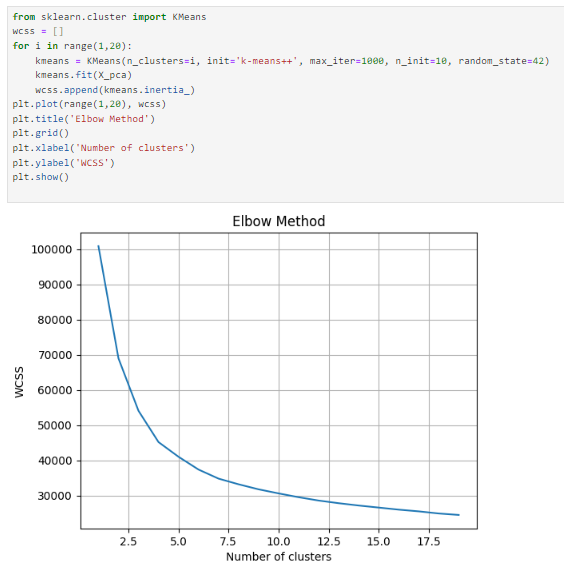


Figure 1: (Within-Cluster Sum of Squares)

After concating the Clusters column to our data frame, we began examining them.

Here we used the describe() function to obtain statistics on each of the six clusters regarding count, mean, std, etc. (Figure 2)

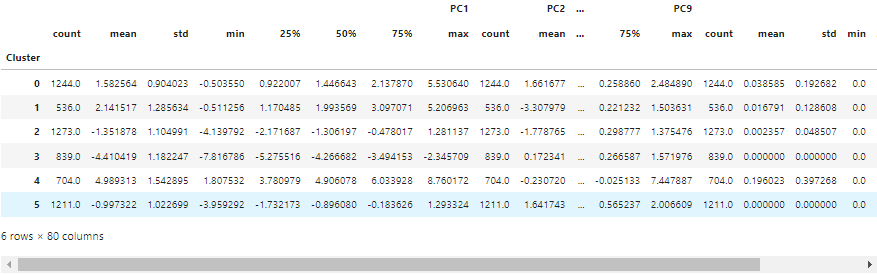


Figure 2: Description of clusters in DataFrame

Next, we used a popular visualization library, matplotlib to visualize scatterplots of the first three PCs with our clusters. (Figure 3)

First Two Components Plot:

* + Clusters are well-separated, especially noticeable for clusters like the blue (0), green (1), and yellow (5), indicating distinct groupings within the data.

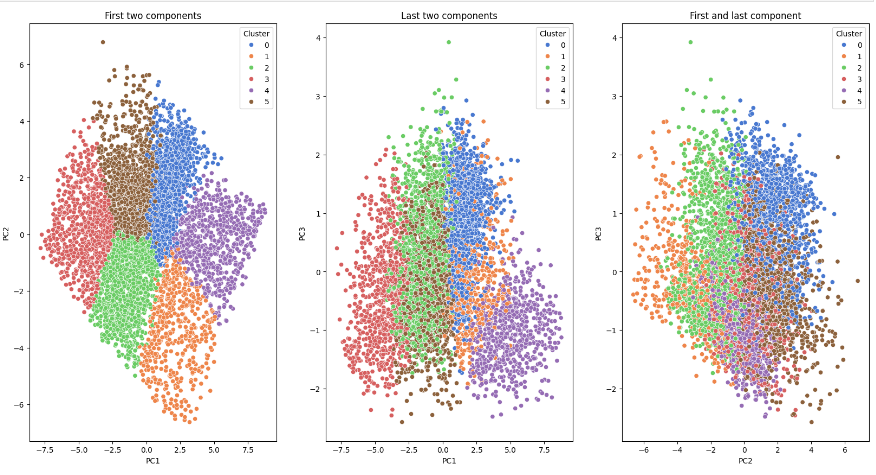
Last Two Components Plot:

* + Less separation among clusters compared to the first two components. This suggests that these components contribute less to the variance and hold less distinctive information about the data groups.
  + This plot can be useful for identifying subtler variations that are not captured by the first two components.

First and Last Component Plot:

* + Combines the most significant variance direction with one of the least, which can help in identifying any hidden structures or outliers that are not obvious in the main component plot.
  + This plot shows a moderate degree of overlap among clusters, indicating some commonalities shared across clusters in the extremities of the data.

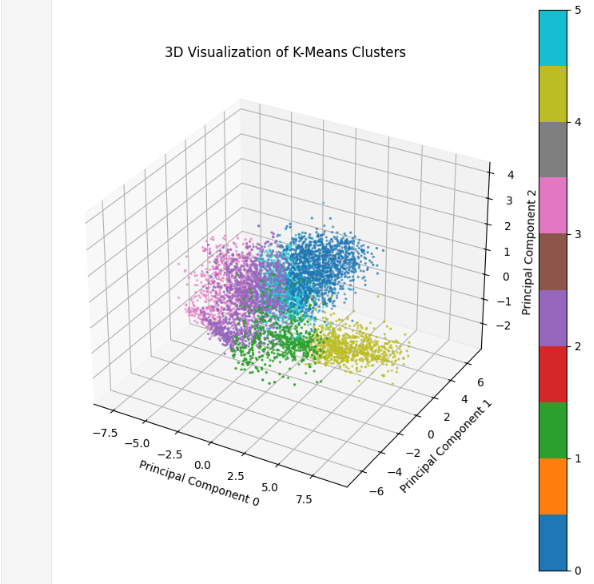
Figure 3: Scatterplots of three Principal Component Analyses and Clusters



Additionally, we used Axed3D from the mpl\_toolkits.mplot3d library to visualize the first three PCs in a 3D space. (Figure 4)

Visualization of K-Means Clusters in 3D (PC0, PC1, PC2):

* + This plot incorporates an additional dimension, offering a deeper insight into the spatial distribution of clusters. It shows how clusters spread out across three principal components, providing a fuller picture of data segmentation.
  + Like the 2D plots, clusters are distinguishable, but the 3D view helps in understanding the depth and overlap that may not be visible in 2D.
  + The spatial separation in 3D can be particularly useful for identifying which clusters might be closer in the multidimensional space, suggesting more in-depth similarities.

****Figure 4: 3D Scatterplot of Clusters in 3D

We utilized box plots to visualize the distribution of each Principal Component (PC) across all clusters. The first four PCs show the most variation with respect to Cluster IDs, exhibiting distinct quartiles (Q1, Q2, and Q3) in each. This differentiation suggests that these PCs are particularly effective in distinguishing between clusters. Variability Across Clusters: Each principal component exhibits different levels of spread across clusters, indicating how each cluster's characteristics differ in the multidimensional space shaped by PCA. (Figure 5)

* Cluster Characterization:
  + PC1: Shows a considerable spread in clusters 0 and 5, suggesting that these clusters differ significantly from others in the first principal component, which typically captures the most variance.
  + PC2: Cluster 3 shows higher variability compared to other clusters, possibly indicating that this cluster is more diverse or has less cohesion in the context of PC2.
  + Higher PCs (e.g. PC7, PC8, PC9): The variability and the range of scores tend to decrease as we move to higher numbered PCs, which often capture less variance. This is visible as generally tighter and more consistent box plots across these components.
* Outliers: There are numerous outliers across various components and clusters. This could suggest either genuine data variability or potential data issues or anomalies that might warrant further investigation.

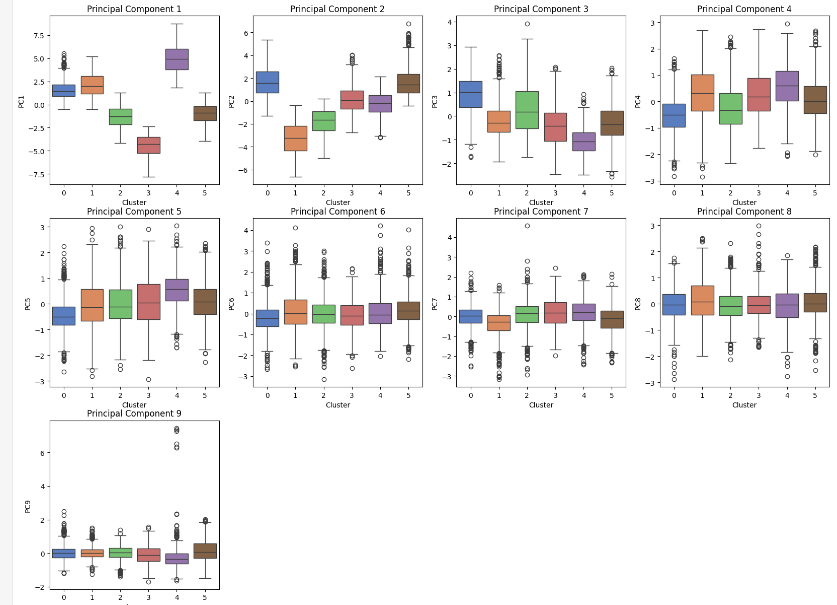
****

Figure 5: Box Plots of PCAs and Clusters

Finally, we have created density plots for the principal components (PCs) across different clusters to analyze their distributions. We observed that the distributions vary by cluster, with some showing skewness, etc. (Figure 6)

* Clusters 0 and 5:
  + Often show distinct peaks and different behaviors across several principal components, suggesting these clusters might be capturing unique or outlier behaviors in the dataset.
* Clusters 1 through 4:
  + These clusters often have more overlapping distributions, especially in higher PCs, indicating more similarity in the characteristics captured by these components.

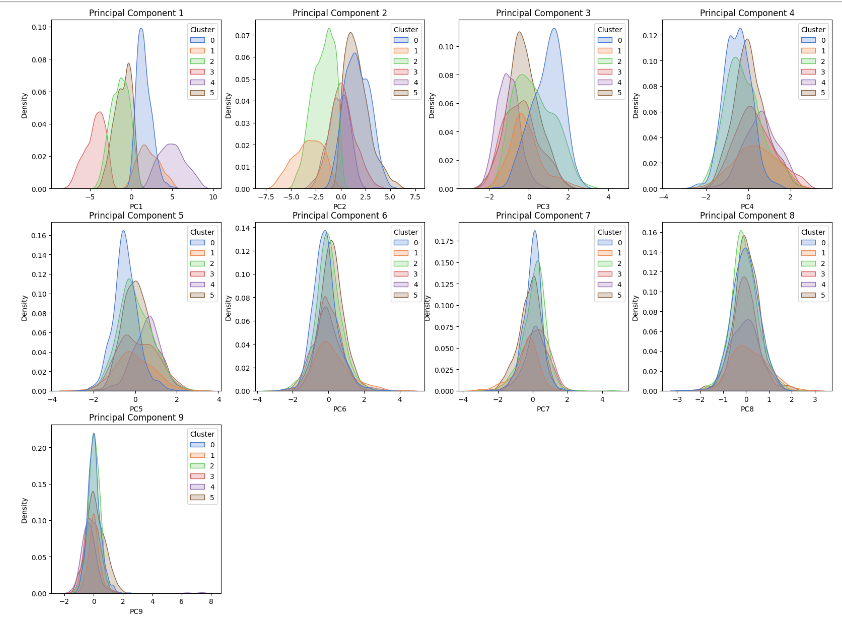
****

Figure 6: Density Plots of PCs and Clusters

**3 Train Model 1: Stacking Method & 4 Train Model 2: k-fold Cross Validation**

| Subgroup ID | Name of Student | Accuracy Score Base Models [TT(TF)] | Avg Accuracy Score Base Models | Accuracy Score Meta Model [TT(TF)]) | Accuracy Score K-fold CV [TT(TF)] | N Features |
| --- | --- | --- | --- | --- | --- | --- |
| 0 | Elia | LR: 75% [36/(12)] |  |  |  |  |
|  |  | SVM: 62% [30/(18)] | 79% | 100% | 75% | 9 |
|  |  | NN: 100% [48/(0)] |  |  |  |  |
| 1 | Elia | LR: 80% [8/(2)] |  |  |  |  |
|  |  | SVM: 60% [6/(4)] | 80% | 100% |  | 9 |
|  |  | NN: 100% [10/(0)] |  |  |  |  |
| 2 | Nataly | LR: 0% [0/(3)] |  |  |  |  |
|  |  | SVM: 0% [0/(3)] | 33% | 100% | 97% | 9 |
|  |  | RF: 100% [3/(0)] |  |  |  |  |
| 3 | Nataly | N/A |  | N/A | N/A | 0 |
| 4 | Matthew | KNN: 80.5% [565(137)] |  |  |  |  |
|  |  | RF: 80.5% [565(137)] | 80.5% | 100% | 96.6% | 9 |
|  |  | GB: 80.5% [565(137)] |  |  |  |  |
| 5 | Matthew | N/A |  | N/A | N/A | 0 |

Figure 7: Table 2 Representation of Group 3

Note: Cluster 3 and 5 had no bankrupt companies so classification was not attempted.

**5 Generalization**

The best model from Sections 3 and 4 was developed by Elia, who worked on the first cluster containing 48 companies that filed for bankruptcy. These results were achieved using the stacking method. We employed this model to predict the outcomes of the test data provided to us. Elia wrote a function for model tuning and applied those models to different data, as well as calculated scoring using Stratified Kfold and GridSearch CV. The 3 best models, which were Logistic Regression, SVM, and Neural Network respectively, were used as base models for stacking. The meta-learner used was logistic regression.

**6 Class Competition**

The accuracy scores, acc train, from the meta-model from Section 3 was 100%. The number of features N features used was 9. These were the results of the best model selected in Section 5 and the best k-fold cross-validation model in Section 5 respectively.