Names: Shyam Parikh, Ayan Mahmood, Ryan Camburn, and Lijing Li

Class: CS 559 WS-1

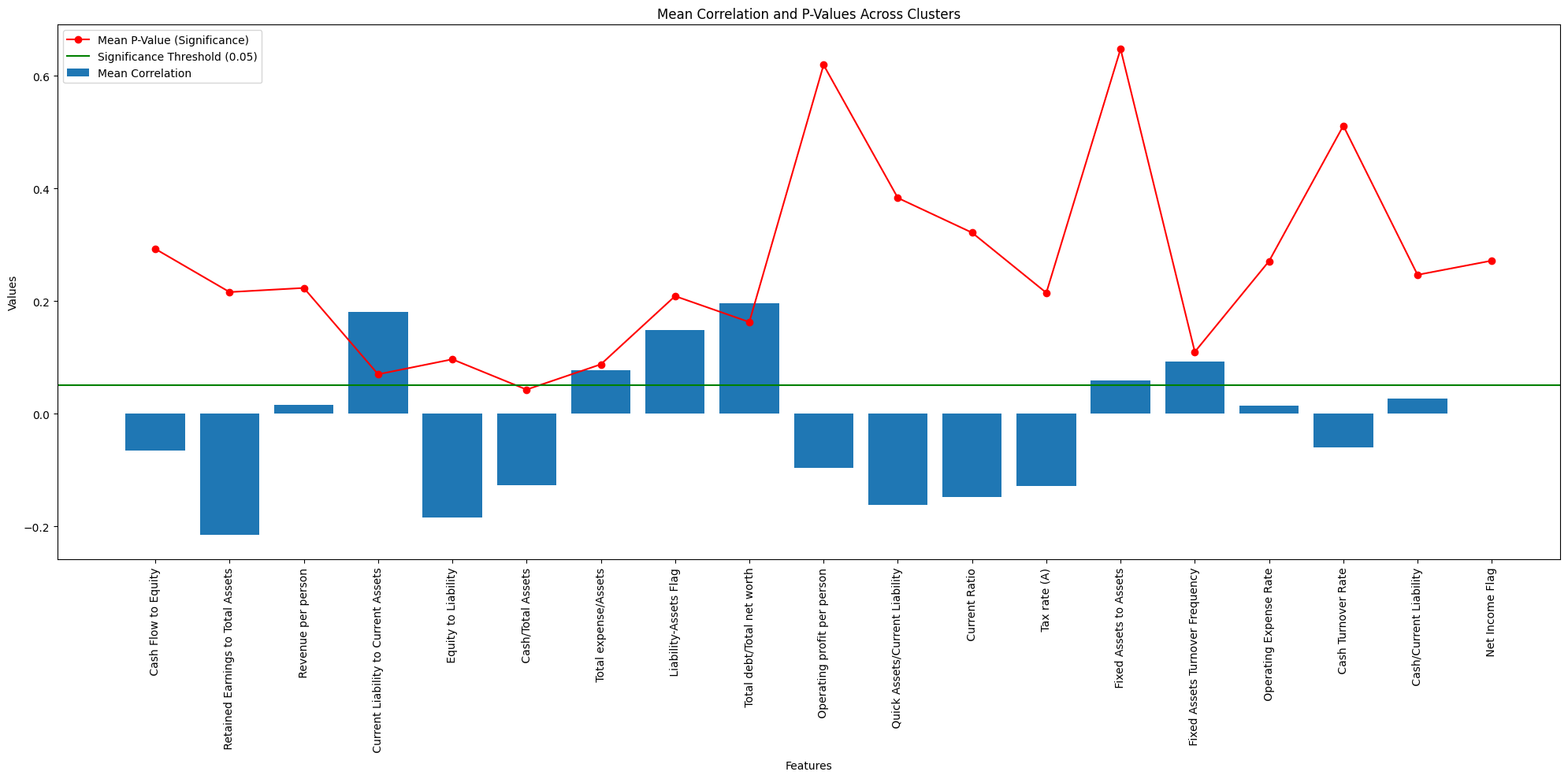
**Company Bankruptcy Prediction Results**

**3.2**

For our Company Bankruptcy Prediction project, our process consisted of a lot of preprocessing, data interpretation, and model testing. The first step of our project was the preprocessing. To train our data, we wanted to try to get most of our data to be stationary or binary. The first step was dimensionality reduction. We needed to reduce the total number of features in our dataset. To do that, we used VIF(Variance Inflation Factor) and also OLS(Ordinary Least Squares) significance testing. VIF is a metric typically used to track multicollinearity in regression analysis. For our case, if we found out that a feature had a VIF of greater than 5, then we would remove that from our feature set. This brought our feature list from 95 total factors to a list of 48 total features. Once we did that, we would go through the remaining features and see which features were significant through Ordinary Least Squares. We tested all of our features and tested their significance in predicting the bankruptcy column. If they had a significance of less than 0.05, then we proclaimed these features significant. This narrowed down our list of features from 48 to 19. After that, we needed to transform the data to attempt to make it stationary. For the scaling of our dataset, we transformed all of our columns using the Box-Cox Transformation. The Box-Cox Transformation works really well for left skewed and right skewed distributions. This transformation really made the mean and median the closest to one another as opposed to other transformations like MinMax scaling, Square Root scaling, and Log scaling. After we Box-Coxed all of our data, we decided to try to perform more dimensionality reduction using Principal Component Analysis. However, that did not produce the best results, so we chose not to move with PCA. Finally, after all of that analysis, we applied K-Means Clustering, an unsupervised learning technique to split our data into various clusters. We chose to work with 4 clusters. In cluster 0, we had a total of 2118 total companies with 2007 being not bankrupt and 111 being bankrupt. In cluster 1, we had a total of 1742 total companies with only 26 companies being bankrupt and 1726 companies not being bankrupt. In cluster 2, we have 1485 total companies with 56 total bankruptcies, and 1429 not being bankrupt. Finally, in cluster 3, we have a total of 5 bankrupt companies and 457 not bankrupt companies bringing our total to 462 total companies.

After performing analysis on all of the clusters, we were able to notice information about the correlations. From looking at all of our cluster datasets what we see is that retained earnings to total assets always seems to be negatively correlated with our target, no matter what subgroup we use. Similarly, that is the same case with Equity to Liability which is also somewhat negatively correlated with the target. On the other hand, we see that the top 3 positively correlated variables are Current Liability to Current Assets, Liability-Assets Flag, and Total debt/ Total Net worth. The one main difference here though is that these correlations all have mixed rankings upon the top three correlations. It's likely that these correlations will probably be most helpful to us in our testing.

To identify the individual significance of each factor into predicting bankruptcy in that respective cluster, we implemented OLS again and found some significant results. In clusters 0 and 2, we found that 10 features out of the 19 were significant. For cluster 1, there were only 8 features that were significant and finally, in cluster 0, we found that only 2 features were significant. What made this even more interesting was that there was no shared factor among all of the clusters. All of the clusters ended up being composed of different significant factors. Here is a plot that takes a combined average of all of the significance of each of the features and an average of the correlation of the features against our Y variable of bankruptcy.



From this plot, we can see that the factors on the left hand side seem to have a lot more of an average significance than factors on the right hand side.

3.3

| Subgroup ID | Name of Student | Average accuracy score base models [TT(TF)] | accuracy score Meta model [TT(TF)] | N\_features |
| --- | --- | --- | --- | --- |
| 0 | Shyam Parikh | 1.00[111(0)] | 1.00[111(0)] | 6 |
| 1 | Ayan Mahmood | 0.67[17(9)] | 1.00[26(0)] | 10 |
| 2 | Lijing Li | 0.98[55(1)] | 0.8214[46(10)] | 5 |
| 3 | Ryan Camburn | 0.67[3(2)] | 1.00[5(0)] | 7 |
| Team |  | 0.83 [186(12)] | 0.96 [188(10)] | 7 |

Up above, we have a report of all of our reported accuracies of our base models and our meta models.

Subgroup 0 Base Models Confusion Matrices

|  | 2007  0 | 0  111 |  |
| --- | --- | --- | --- |

Above is the Confusion Matrix for DecisionTreeClassifier on Cluster 0

|  | 2007  0 | 0  111 |  |
| --- | --- | --- | --- |

Above is the Confusion Matrix for ExtraTreesClassifier on Cluster 0

|  | 2007  0 | 0  111 |  |
| --- | --- | --- | --- |

Above is the Confusion Matrix for RandomForestClassifier on Cluster 0

Subgroup 0 Meta Model Confusion Matrix

|  | 2007  0 | 0  111 |  |
| --- | --- | --- | --- |

Subgroup 1 Base Models Confusion Matrices

|  | 1716  1 | 0  25 |  |
| --- | --- | --- | --- |

Above is the Confusion Matrix for RandomForestClassifier on Cluster 1

|  | 1716  26 | 0  0 |  |
| --- | --- | --- | --- |

Above is the Confusion Matrix for SVC on Cluster 1

|  | 1716  0 | 0  26 |  |
| --- | --- | --- | --- |

Above is the Confusion Matrix for DecisionTreeClassifier on Cluster 1

Subgroup 1 Meta Model Confusion Matrix

|  | 1616  0 | 100  26 |  |
| --- | --- | --- | --- |

Subgroup 2 Base Models Confusion Matrices

|  | 1413  0 | 16  56 |  |
| --- | --- | --- | --- |

Above is the Confusion Matrix for RandomForestClassifier on Cluster 2

|  | 1429  0 | 0  56 |  |
| --- | --- | --- | --- |

Above is the Confusion Matrix for GradientBoostingClassifier on Cluster 2

|  | 1257  3 | 172  53 |  |
| --- | --- | --- | --- |

Above is the Confusion Matrix for SVC on Cluster 2

Subgroup 2 Meta Model Confusion Matrix

|  | 1190  10 | 239  46 |  |
| --- | --- | --- | --- |

Subgroup 3 Base Models Confusion Matrices

|  | 457  0 | 0  5 |  |
| --- | --- | --- | --- |

Above is the Confusion Matrix for RandomForestClassifier on Cluster 3

|  | 457  0 | 0  5 |  |
| --- | --- | --- | --- |

Above is the Confusion Matrix for ExtraTreesClassifier on Cluster 3

|  | 457  5 | 0  0 |  |
| --- | --- | --- | --- |

Above is the Confusion Matrix for KNeighborsClassifier on Cluster 3

Subgroup 3 Meta Model Confusion Matrix

|  | 367  0 | 90  5 |  |
| --- | --- | --- | --- |