Final Project - Fortune 1000 Companies

Ryan Benner

Introduction

This data provides various statistics/metrics about Fortune 1000 companies, and how they have changed over the past year (2021-2022). The data set was found on Kaggle, and there are originally 1000 rows, each representing one company from the top 1000 companies. When the dataset was first used, it was actually unusable, because all the columns were objects, rather than floats or continuous variables (which they should have been, e.g. Revenues: $26,000.0, instead of 26000.00). Because of this, the various sklearn libraries were unable to comprehend the data. In order to circumvent this, it was necessary to clean the data by getting rid of all commas, $ signs, % signs, and then changing each column to the correct data type. After cleaning, there were still some missing values, so there are 688 usable rows out of 1000.

The variables being utilized in the data are:

- Name
- Revenues
- Revenue_percent_change
- Profits
- Profits_percent_change
- Assets
- Market_value
- Change_in_rank
- Employees

Question #1: When predicting company profitability, which predictor has the biggest impact on profits? How does excluding certain predictors affect the model's accuracy in predicting profits?

Methods

In order to determine what variables have the largest impact on predicting a company's profitability, a series of linear regression models are being trained using a train/test split with the dataset. In each of the models, one variable is excluded from the predictors array and the respective train/test R2 scores are stored to be displayed and compared graphically across all models. Originally, it was planned to use a simple linear regression model, but with the R2 scores being very good in training but poor in testing, it was decided that it was necessary to use a model that gives penalties to model complexity. Because of this, ridge regression is being used, so that any added model complexity is penalized.

Results

After the analysis concluded, the models were found to be fairly successful, with an average R2 score around ~0.75 for the training models, and around ~0.7 for the test models. With an R2 score of 1 signifying a perfect fit, this is indicative that the models are a good match to the data. As expected, a few testing R2 scores were lower than others, which would signify that the variable being excluded in that model has a greater impact on the model's performance than other variables. Market_value and Employees had the largest dip in model performance when excluded from the model's predictors.

Discussion

Following the completion of the analysis results, the more notable variables were found to be market value and employees, where their exclusion led to noticeable reductions in R2 scores for those models. This outcome suggests that there is strong correlation between these two variables and a company's profitability, likely because of their direct relationship with the company scale and asset valuation. Further investigation of these variables could allow for deeper insights as to how these factors contribute to a company's overall success and profitability. In order to better visualize the data relationships and R2 scores found, two graphs were made: A bar graph of all training and testing R2 scores, and a scatter plot of

market_value versus company profits. With these two graphs, it is evident that market_value exclusion drastically decreases the model's performance, and the scatter plot makes it evident as to why - There is a high linear increasing correlation between market_value and profitability.

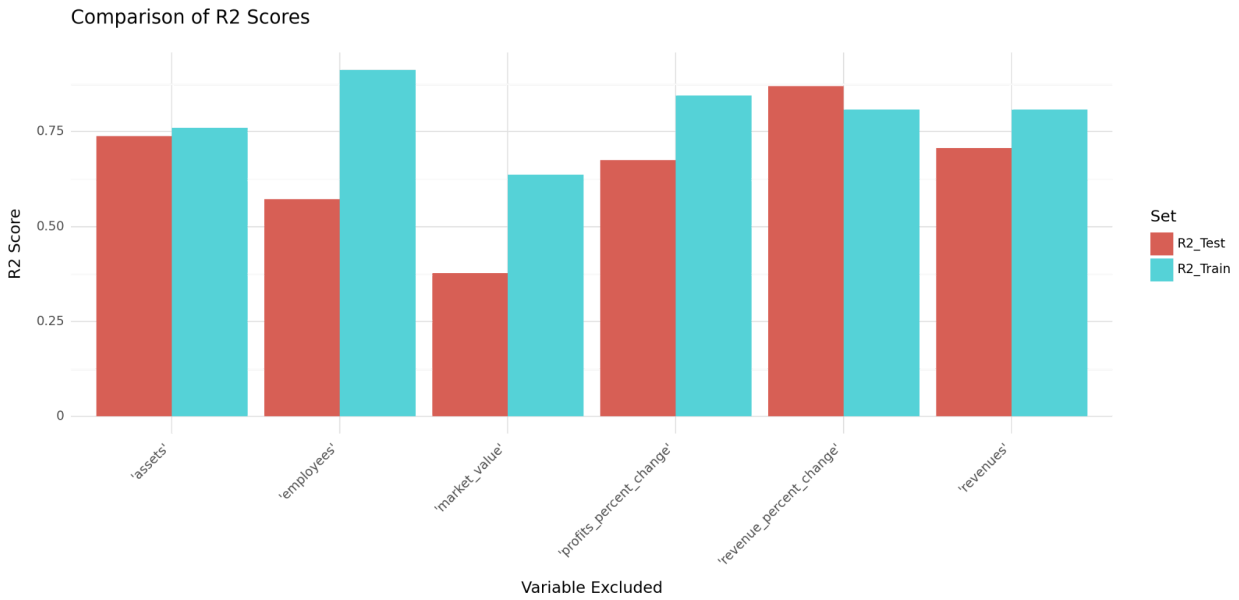

Figure 1: Bar Graph of Train/Test R2 Score with Different Variable Exclusion

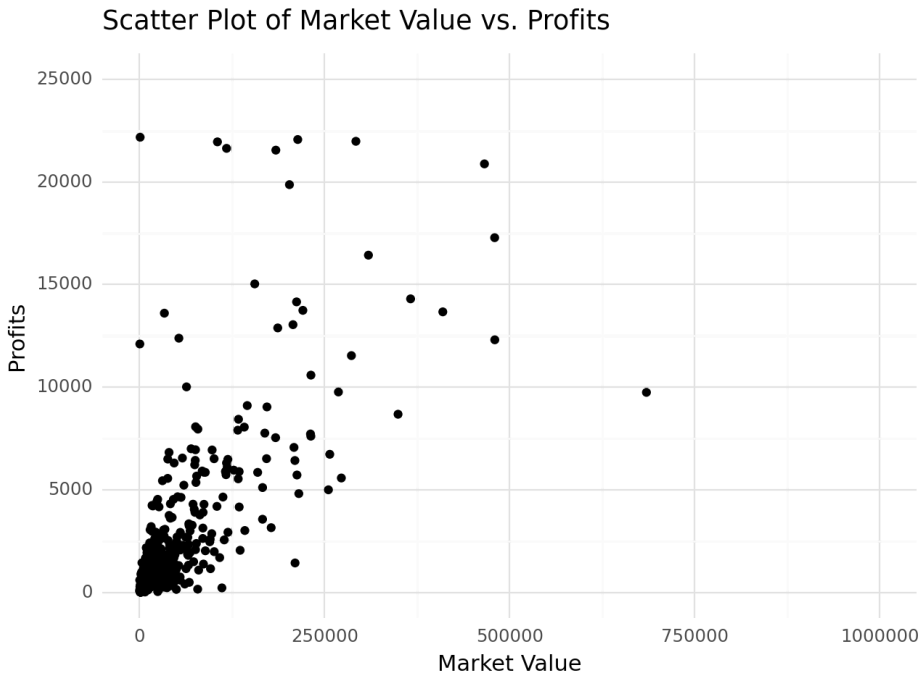

Figure 2: Scatter plot of Market Value vs. Profits in Fortune 1000 Companies

Question #2: Can clustering methods identify groups of companies with similar properties based on revenues, profits, and assets?

Methods

In order to identify groups of companies with similar properties based upon revenues, profits, and assets, a clustering algorithm was utilized, to find the optimal number of clusters, as well as place each company into the most likely cluster. Originally, it was proposed to use K-Means as the clustering algorithm of choice, but was quickly rejected due to the non-spherical nature of the clustering of companies. K-Means is an algorithm that assumes each cluster is roughly spherical in nature, which these clusters evidently are not. Because of this, the method of clustering was altered to utilize a Gaussian Mixture Model(GMM), which does not assume spherical clusters, and rather clusters data points based upon the likelihood of that point appearing in each cluster. In order to find the optimal number of clusters, Bayesian Information Criterion was utilized, to find the number of clusters where this performance value is minimized. Additionally, silhouette score was used to determine the performance of the model, as it recognizes cohesion and separation of clusters (how close points are in each cluster, how far apart clusters are).

Results

Upon analyzing the predictors revenues, profits, and assets, with a GMM, it was found that the optimal number of clusters was 3. When 3 was selected as the number of clusters, the clusters performed very well, achieving an impressive silhouette score of 0.84. Typically, silhouette scores for GMMs tend to be on the lower end, as the points are not all equidistant from the center of the cluster, and rather are placed in clusters based upon probability. With both of these metrics in mind, the model did a successful job of clustering the companies based upon revenues, assets, and profits.

Discussion

This analysis utilizing GMM to cluster companies revealed that three distinct clusters were able to be effectively identified, yielding a high silhouette score of ~0.84. The results indicate that companies within each cluster likely share similar market strategies or operational scales. Upon further investigation, more information could be uncovered about what each cluster of companies shares in common on an underlying basis, rather than just what is apparent in the dataset being analyzed. Each cluster likely has unique characteristics different from the other

clusters, and insights about those clusters could aid in management or investment decisions for any company looking to grow in size and profitability. In each of the clusters, there appear to be three kinds of companies amongst Fortune 1000 companies: ‘Average’ performing, ‘High’ performing, ‘Low’ performing. The red cluster is indicative of the average performance, where assets, revenues, and profits, are all lightly correlated, each graph showing a slow but steady linear increase for these companies. The blue cluster shows companies with massive amounts of assets, but diminishing returns in terms of profits. Finally, the green cluster shows companies that have high profit margins, and less assets than other clusters. With mediocre revenues, but huge profits, this signifies that their costs are very low, and they are able to make a large profit due to this.

To insert an image use Insert > Image

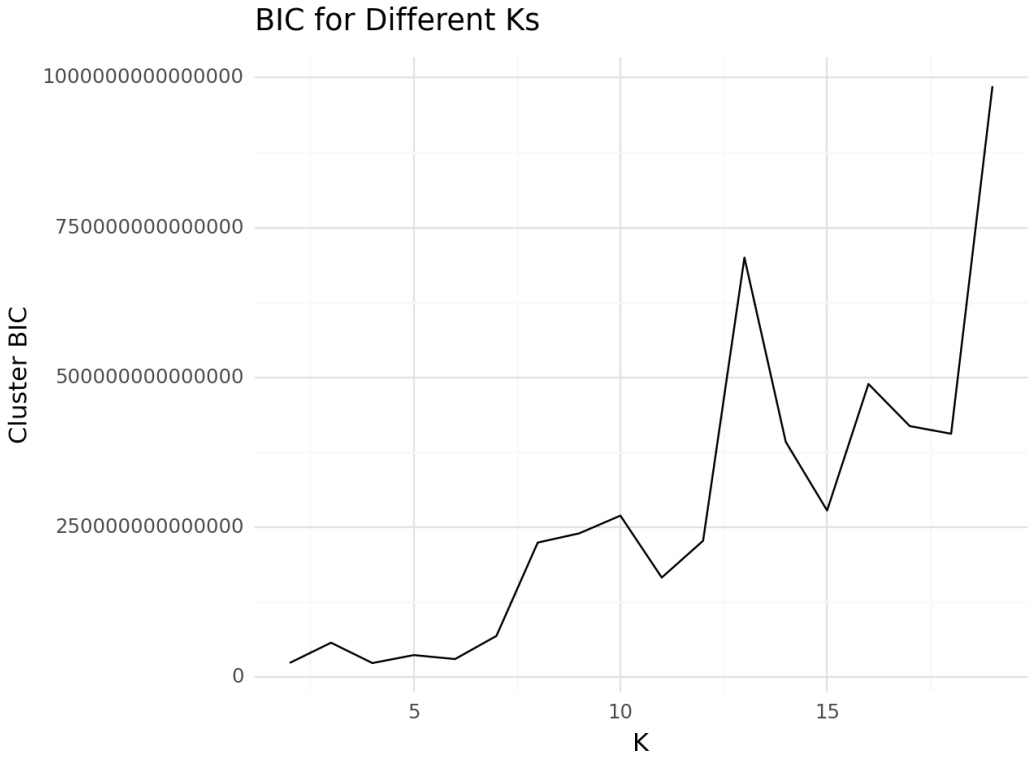

Figure 1: Bayesian Information Criterion to Determine Optimal K-Clusters

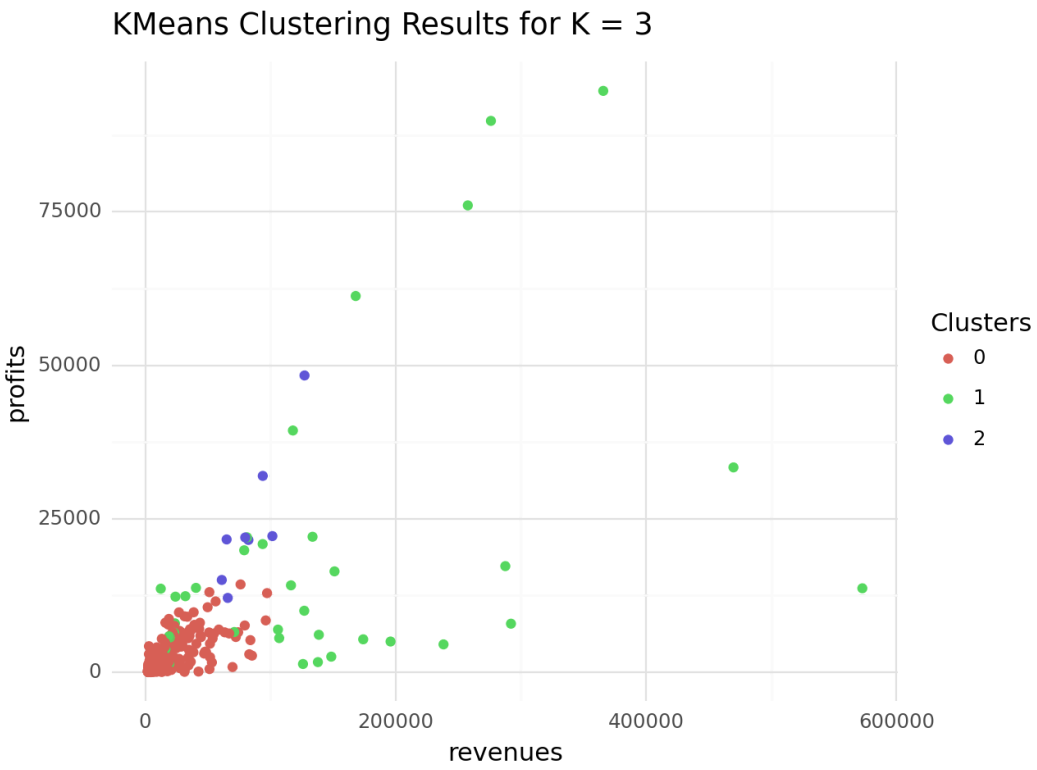

Figure 2: Scatter Plot of Revenues vs. Profits by Cluster

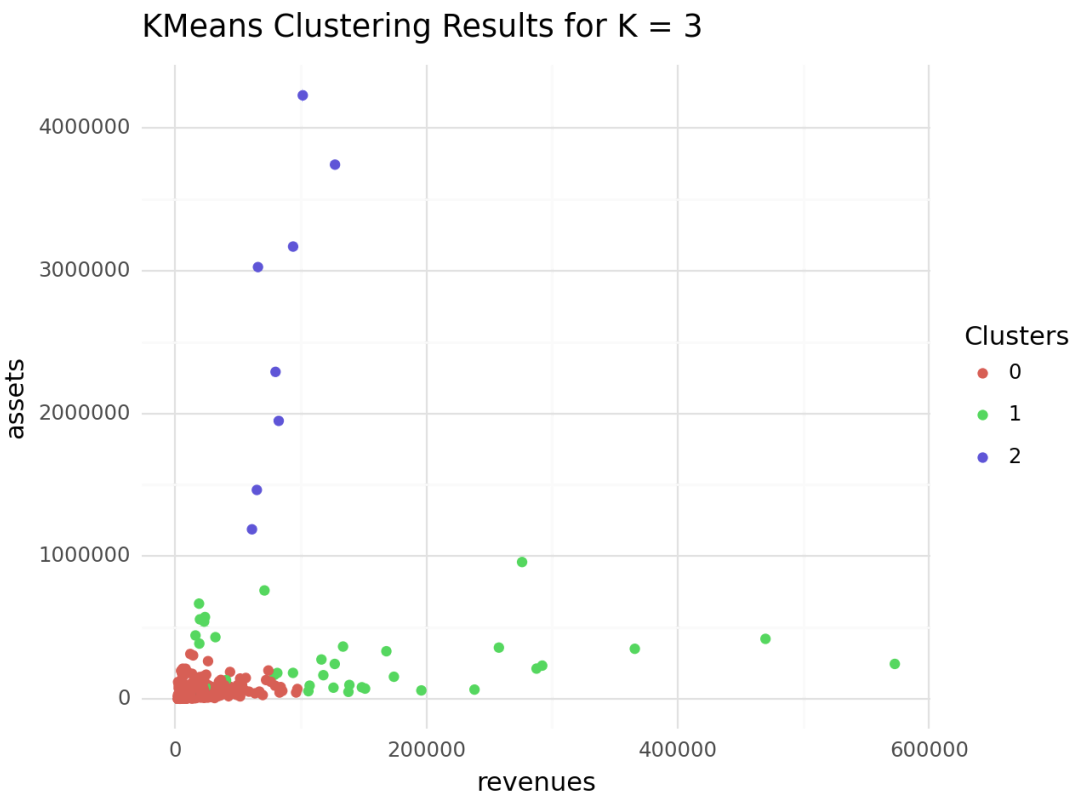

Figure 2: Scatter Plot of Revenues vs. Assets by Cluster

Figure 2: Scatter Plot of Assets vs. Profits by Cluster

Question #3: In predicting profit margins, how effective is dimensionality reduction at reducing the number of variables needed to maintain a high level of accuracy? Which key components or variables are the most important in determining profit margins after reduction?

Methods

For this analysis, principal components are utilized in order to reduce the dimensions being used in a linear regression model with a high level of accuracy. This is similar to the first model, but instead aims to leverage dimensionality reduction and learn more about the weights of each variable in determining profit margins after reduction. To perform this analysis, a Cumulative Variance Plot and Feature Importance plot are used to determine the optimal number of principal components needed to retain 93% of the original data, as well as determine the loadings of each variable in the first principal component. Finally, the optimal number of PCs is used in a linear regression model, which is tested with R2 score to measure its accuracy in determining profits after dimensionality reduction.

Results

What were the results of your model(s) and analysis?
In this analysis, it was found via the cumulative variance plot that the optimal number of PCs to retain 93% of data is 5. With there being 7 original variables, this represents a 28% reduction in the amount of data being used to model the predictions. Once the model was trained and tested, it was determined that there was an impressive 0.74 training R2, and 0.96 testing R2. This is a much higher R2 than was found in the first question, and is demonstrative that PCA does an excellent job of retaining and representing the original data. Finally, the feature importance plot showed that employees and revenues have the largest impact on the model's success rate.

Discussion

The use of Principal Component Analysis in predicting profit margins and determining the most influential variables has proven to be highly successful, with a very accurate model being generated while simultaneously reducing the data by 28%. By reducing the number of variables from 7 to 5, PCA retained 93% of the data and gained a 0.96 R2 score in testing. This demonstrates that PCA not only simplifies the model, but also has excellent predictive power. Additionally, the analysis found that employees and revenues are very important when considering profit margins, which could be very beneficial to research further as to why this correlation occurs. This successfully shows how dimensionality reduction can be useful to increase efficiency and effectiveness of a model.

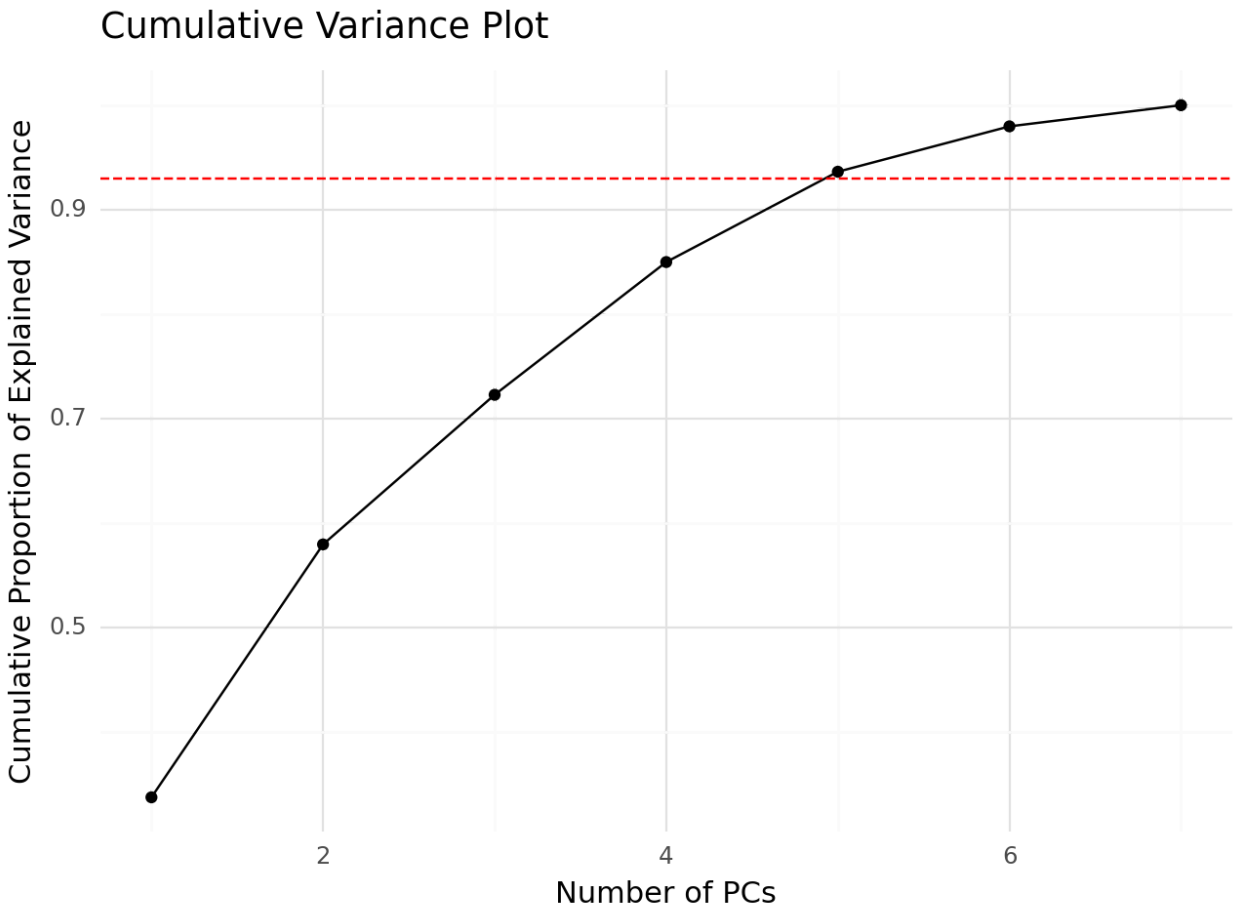

Figure 1: Number of Principal Components vs. Proportion of Variance Explained

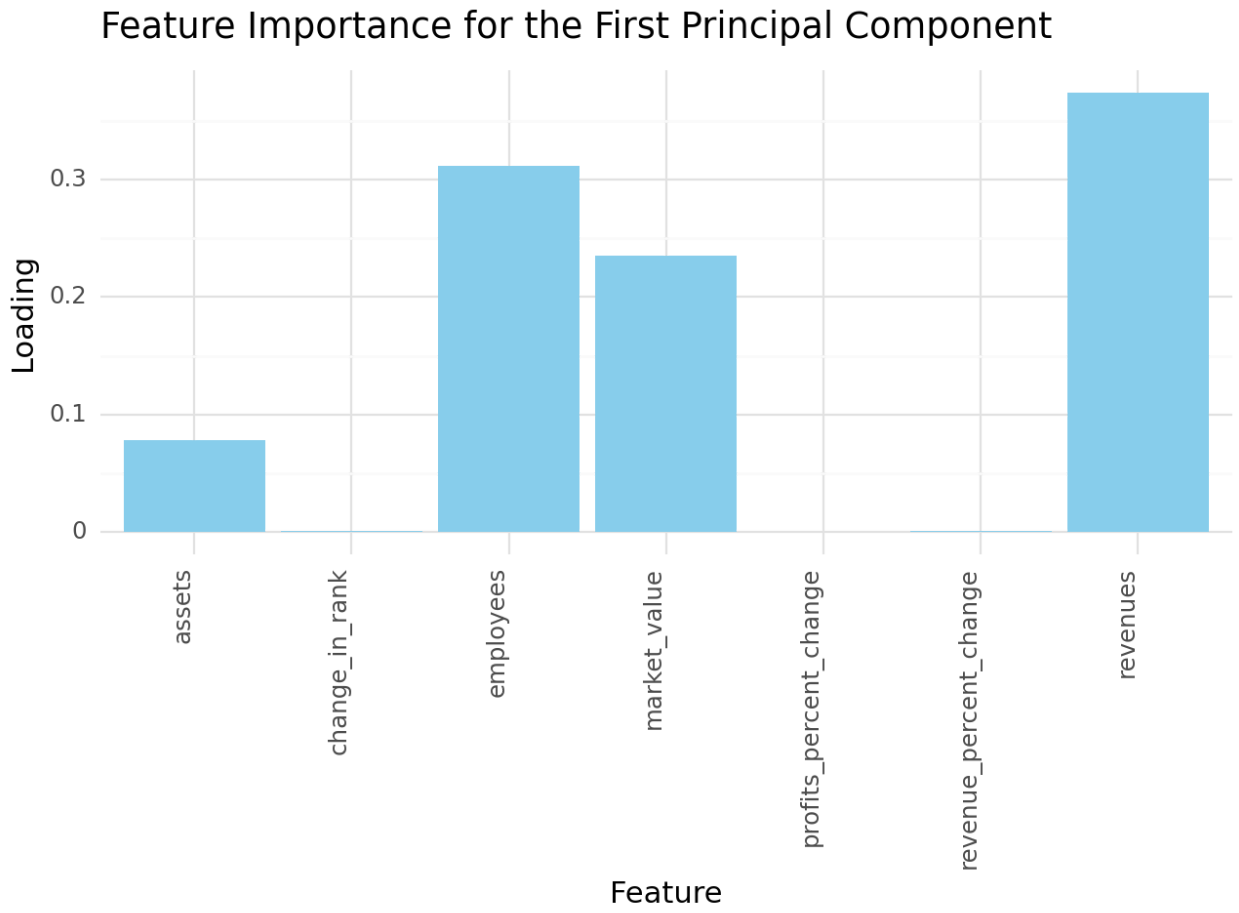

Figure 2: Bar Chart of Feature Importance Loadings for each Predictor