**Executive Summary:**

The primary goal of our analysis relates to the determination of key factors that influence bankruptcy in large firms. In order to realize this goal, we utilized various statistical methods on a dataset

collected from the Taiwan Economic Journal, from the years dating from 1999 to 2009. The raw data was extracted from formal financial statements including balance sheets, cash flow statements, and income statements taken from the TSEC financial databases. The dataset is largely comprised of various accounting and finance ratios e.g., cash flow rate, net value per share, etc.

The origins of bankruptcy prediction date back as early as the 1930s, namely with FitzPatrick’s [1] analysis of 13 financial ratios in failing and successful firms. FitzPatrick concluded that, in a majority of instances, the successful firms displayed favorable ratios while the failed firms displayed unfavorable ratios, when compared to the prevailing average ratios and ratio trends. The first multivariate statistical analysis of bankruptcy prediction was performed by Edward Altman in 1968 [2]. Utilizing multivariate discriminant analysis, Altman was able to produce a five-factor model to accurately predict bankruptcy of manufacturing firms. The Z-score model produced by Altman had an 95% ability to accurately predict firm bankruptcy one year before failure. When Altman’s model was tested to predict firm bankruptcy further than one year in the future, the model’s accuracy fell drastically. Neural networks and advanced machine learning have helped to drastically increase the accuracy of model’s when asked to predict firm bankruptcy over a prolonged period of time. Notably, Tsukuda and Baba’s development of a neural network [3] that was able to predict bankruptcy with maximum accuracy, several years into the future, thanks to advanced machine learning.

It goes without saying that the field of bankruptcy prediction has continually evolved, even over the course of the past twenty years. Not only have the ratios become increasingly complex, so have the consequences of mis predicting whether or not a firm is to go bankrupt. In an increasingly globalized world economy, mis predicting bankruptcy or altogether failing to recognize whether or not a firm will go bankrupt can result in untold financial loss. In order to attempt to mitigate said consequences, our analysis has been targeted towards understanding key factors that influence bankruptcy in large firms.

Our original dataset included 6,819 companies and 96 variables. Initial data preprocessing was done in order to be able to analyze useful information. We used multiple methods in order to analyze the dataset. The methods we used were Principal component analysis (PCA), and Logistic Regression. Our research we can conclude that firm bankruptcy is significantly impacted by 3 components, which we garnered from PCA. The 3 components are Earnings Per Share, Inflation and Interest rate expectations, and Return on Assets. Earnings Per Share was found to be the most significant when attempting to predict firm bankruptcy. In addition, the three components derived from the PCA helped us to inform the variable selection of our logistic regression analysis. Upon completion of the logistic regression analysis, it became clear to us that despite not being the most advanced statistical techniques, PCA and logistic regression were still able to predict firm bankruptcy with notably high accuracy. The key factors that were determined thanks to the help of the 3 principal components coupled with the historically identified factors proved that non-advanced statistical techniques were still able to produce accurate predictive results. Ultimately, this conclusion is to the benefit of investors, governments, and entrepreneurs around the globe due to the fact that these results are much more interpretable than the results produced from advanced neural networks and machine learning techniques.

Naturally, there are limitations to the conclusions that we were able to draw regarding bankruptcy prediction. It is our opinion that the largest of these limitations has to deal with the universal application of the conclusions drawn from our analysis. Specifically, we have identified two major issues with the original dataset that would deter our findings from being used in an attempt to holistically assess bankruptcy prediction. For one, the dataset is unique to the Taiwanese market. The implication from this issue is that while our model may be accurate at predicting bankruptcy in Taiwanese firms, that is not to say that our model would lend itself to other international markets. The second issue is that the dataset is only analyzing large firms. The implication from this issue is that our findings are not designed for small firm or individual bankruptcy prediction.

**References:**

[1]. FitzPatrick, P. 1932. A comparison of ratios of successful industrial enterprises with those of failed companies. The Certified Public Accountant (October): 598-605, 656-662, and 727-731, respectively.

[2]. Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, *23*(4), 589-609.

[3]. Tsukuda, J., & Baba, S. I. (1994). Predicting Japanese corporate bankruptcy in terms of financial data using neural network. *Computers & Industrial Engineering*, *27*(1-4), 445-448.

**Bankruptcy Prediction in Taiwanese Firms Utilizing Principal Component Analysis and Logistic Regression**

Joe Larson

Nikhil Bhalala

DePaul University, DSC 324, Stephanie Besser

**Abstract:**

The prevalence of extensive financial data has brought us increasingly closer to understanding how to accurately predict bankruptcy in large firms. However, issues still persist regarding the predictive modeling techniques that are deployed in order to make effective bankruptcy forecasts. Utilizing a dataset formed from the Taiwanese Economic Journal, this paper employed a Principal Component Analysis (PCA) coupled with a Logistic Regression Analysis in order to study the important components and variables associated with accurate bankruptcy prediction. Three major components were obtained after the completion of the PCA analysis, the components were defined as follows: Earnings Per Share, Inflation and Interest Rate Expectations, and Return on Assets. Additionally, the Logistic Regression revealed a number of significant variables that serve to explain and predict the bankruptcy dependent variable. The prediction of bankruptcy is a significant financial issue that has concerned regulators, researchers, and investors for years, largely due to the fact that early detection of bankruptcy can result in the reduction of the economic losses that are associated with filing for bankruptcy. Utilizing the exploratory results garnered from the statistical methods employed in this paper, it is clear that specific financial variables and components can be used to accurately predict bankruptcy in large firms.

**Introduction:**

The study of bankruptcy prediction has been a staple in multivariate statistical research for nearly a century. Accurate prediction of firm bankruptcy has become increasingly important as the world economy has shifted towards globalization. The interconnectedness of the global economy has undoubtedly resulted in several positive byproducts, including a general overall increase of economic efficiency [1] (Burlacu, Gutu, Matei, 2018). However, that same interconnectedness has simultaneously raised the level of risk for investors, governments, and entrepreneurs alike. Specifically, one of the primary risks is bankruptcy and the economic fallout that results from bankruptcy.

High predictive accuracy has always been the central focus of bankruptcy prediction research. There are multivariate statistical studies that have been concerned with the accurate prediction of bankruptcy dating back to the 1930s [2] (FitzPatrick, 1932). FitzPatrick was credited as a pioneer in the bankruptcy prediction field becoming one of the first to effectively utilize statistics to analyze financial ratios and consequently predict firm bankruptcy. FitzPatrick and his peers that followed shortly after him provided the framework for which the study of modern bankruptcy prediction lies. Contemporary bankruptcy prediction has ultimately progressed beyond the simple predictive models and statistical methods utilized by early pioneers of the field such as FitzPatrick. As early as the 1990s, researchers began developing advanced predictive models utilizing neural networks and advanced machine learning techniques [3] (Tsukuda, Baba, 1994). Although, these modern neural networks have proven to achieve the primary goal of maximum predictive accuracy, that is not to say that the foundational modeling and statistical methods are not without merit.

In this article, we examine how simple modeling and statistical methods can still accurately determine the key factors that influence firm bankruptcy in large firms. Our aim with this article is not to provide the most accurate and innovative statistical analysis of bankruptcy prediction. Rather, our focus is concerned with the utilization of simple multivariate modeling and statistical methods to determine the key factors of large firm bankruptcy prediction while simultaneously shortening computation time and maximizing interpretability. Two primary techniques were utilized to illustrate our research goals, the first being a Principal Component Analysis (PCA) and the second technique being a logistic regression analysis. These two techniques were subsequently applied to a large real-world set of Taiwanese financial data.

**Literature Review:**

In an effort to gain a holistic understanding of the history of bankruptcy prediction, we studied the work: A Review of Bankruptcy Prediction Studies 1930-Present [4] (Gissel, Giacomino, Akers, 2007). It was important for our study to understand not only recent findings in the field of bankruptcy prediction, but also to grasp the foundations of the field. The three authors of the work successfully illustrate the comprehensive history of bankruptcy prediction studies, ranging from the 1930s to the late 2000s. The author’s summarization of Altman’s [5] (Altman, 1968) massively influential bankruptcy prediction models proved to be crucial in the formation and execution of our central research question. Ultimately, the work of Gissel, Giacomino, and Akers served to be an excellent entry-level piece into the field of bankruptcy prediction. The authors predominately cite primary historical material in a highly objective fashion which ultimately maximized our chronological understanding of the field of bankruptcy prediction.

In this article, the identification of specific financial ratios was a central cornerstone of our bankruptcy research question. One of the foundational works pertaining to identifying ratios that predict firm bankruptcy is entitled: Financial Ratios as Predictors of Failure [6] (Beaver, 1966). Beaver helped to identify several ratios that were noted as statistically significant in prematurely identifying bankruptcy in firms. The significant ratios provided by Beaver proved to be foundational for our variable selection process and conclusions discussed later in the article.

In order to gain a better understanding of contemporary statistical methods, we studied a work: Bankruptcy Prediction in Firms with Statistical and Intelligent Techniques and a Comparison of Evolutionary Computation Approaches [7] (Chen, 2011). In addition to a comprehensive analysis of various modern techniques, Chen’s work also helped to cement the continued merit of utilizing PCA in bankruptcy prediction. Ultimately, all of these works were crucial in our efforts to the determination of key factors that influence bankruptcy in large firms utilizing simple statistical methods.

**Methods:**

i. Exploratory Analysis & Pre-processing

Our dataset was sourced from the Taiwan Economic Journal, from the years dating from 1999 to 2009. The raw data was extracted from formal financial statements including balance sheets, cash flow statements, and income statements taken from the TSEC financial databases. The dataset is largely comprised of various accounting and finance ratios e.g., cash flow rate, net value per share, etc. Our original dataset included 6,819 companies and 96 variables. Initial data preprocessing was done in order to be able to analyze useful information. Raw data was preprocessed by deleting the unwanted rows and performing log-transformation. Multicollinearity was checked by using the VIF function on both the models obtained for the PCA and the logistic regression.

Variable selection was chosen based primarily on historical findings revealed in the literature review process. In total, six variables were chosen to be included in our final logistic regression model. Current Ratio and Cash Flow to Total Assets were listed as two of the most commonly utilized factors in predictive bankruptcy models [4] (Gissel, Giacomino, Akers, 2007) according to A Review of Bankruptcy Prediction Studies 1930-Present. The remaining four variables: ROA, Cash Flow to Liability, Gross Profit to Sales, and Persistent EPS were sourced based off of Altman’s 1968 multivariate study of bankruptcy [5] (Altman, 1968) and findings resulting from the PCA analysis. The minimum number of factors across the history of bankruptcy prediction models has been 5 and below [4] (Gissel, Giacomino, Akers, 2007) therefore we deemed six factors to be an adequate number of factors. As mentioned, VIF statistics were ran on the model, utilizing the car:() function in R, The VIF scores for all six factors were 2.7 and below.

i. Principal Component Analysis

Principal Component Analysis was utilized in our analysis primarily due to its ability to tackle the issue of dimensionality that was prevalent in our dataset. PCA is a linear dimensionality reduction algorithm that is used to find a more meaningful basis in a dataset. Our Taiwanese financial data set was clogged with 96 factors, all of which were numerical in nature. When dealing with 96 numerical factors, interpretability becomes a severe issue. PCA was crucial in our ability to reduce the overwhelming scope of our dataset, and ultimately our ability to interpret which of the 96 factors are the best at predicting firm bankruptcy. PCA also helped produced numerous visualizations that helped to us to visually interpret the results of our analysis and contextualize the data.

Principal component analysis (PCA) was conducted using the prcomp function, and principal function from the Psych package in RStudio, varimax rotation was used. As mentioned, the original dataset contained 96 variables and 6819 observations. After cleaning and preprocessing, the resulting dataset contained 93 variables and 6819 observations. The dependent variables related to bankrupt were removed from the dataset as well, leaving 2 total variables. Before starting the PCA process, the following techniques were used to verify assumptions about the dataset: Cronbach’s Alpha, Kaiser-Meyer-Olkin (KMO), and Bartlett’s Test of Sphericity. Cronbach's Alpha assesses the consistency of each summated scale in the dataset. A value of 0.02 was recorded, suggesting good reliability for the dataset. Next, KMO was checked, recording a value of 0.50, indicating a relatively strong relationship and durable sample size. Lastly, Bartlett’s Sphericity Test was checked to evaluate shared variance. A value of p < 2.22e-16 was recorded. The null hypothesis that there are no correlations present in the dataset was rejected, accepting the alternative hypothesis that there are correlations, shared variance between the variables. In order to choose the number of components for analysis to best represent the variance in the dataset, both the eigenvalue and scree methods were considered. Results of the PCA will be discussed further in the discussions and results section below.

ii. Logistic Regression

Logistic regression was utilized in our study due to its ability to produce comprehensive predictive analyses. More specifically, logistic regression is the appropriate analysis to conduct when the dependent variable is dichotomous. In the case of our dataset, our dependent variable “Bankrupt.”, is coded in as a binary value. Additionally, the inordinate amount of predictor variables resulted in difficulties with data interpretation, and logistic regression provided an efficient means in which we would be able to explain the relationship between the predictors variables and the “Bankrupt.” dependent variable. Furthermore, logistic regression was deemed to be appropriate for our analysis due to the lack of outliers and the lack of multicollinearity among the predictor variables.

Logistic regression was performed utilizing a set of six predictor variables that were selected based on previous notable studies as well as results garnered from the PCA. The logistic regression, ran in RStudio, was conducted using the glm() function from the glmnet library. Summary statistics were produced after the initial running of the logistic regression model. The VIF statistics checking for multicollinearity illustrated that each predictor variable had a VIF of 2.7 or less, suggesting that there was no issue of multicollinearity. Model fit was tested utilizing McFadden’s pseudo r squared. The resulting McFadden pseudo r-squared value of .2598 suggest appropriate model fit. Results of the logistic regression will be discussed further in the discussions and results section below.

**Discussion and Results:**

Principal Component Analysis:

The scree method **Figure -1** in contrast, suggested all of 10 components. To test the components suggested by the eigenvalue method, 3 components were used, using varimax rotation and a final cutoff value of 0.2, in three separate analyses. Using 4 components, the fourth component was questionable from an interpretability standpoint, as the variables did not give a clear picture of what the component meant in terms of application. Lastly, analysis was conducted and evaluated using 3 components. The main difference here, compared to the analyses with 3 and 4 components, was that all of the components were easily interpretable. Components 3 and 4, respectively, contained variables that made sense together, in terms of the application. Thus, 3 components were determined to be ideal for analysis. After configuring the PCA model with 3 components, the results are shown in **Table (1).** Using 3 components, 46.8% of the variance was accounted for overall. Components 1 and 2, respectively, made up 21.1% and 36.5 % of the variance. Components 2 and 3 added 14.3% and 10.3%, Proportion Var respectively. In additional information about the variance of the independent variables in the dataset that was relevant to the application.

As we see in the give **Figure(3)** that component 1 the minimum score is -91.0578and the maximum score is 170.6308.As we see in the give **Figure(3)** for component 2 the minimum score is -110.7711and the maximum score is 337.5442.Then we see in the give **Figure(3)** for component 3 the minimum score is -21.1098 and the maximum score is 44.8143.After consulting the loadings **Figure(2)**, Component 1 was identified as being as net value per share and also persistent EPS in the last four seasons, it was cash flow per Share and operating profit per share yuan assets, therefore, Component 1 was defined as **Earnings Per Share.** Component 2 was identified as being categorized by strong borrowing dependency, current liabilities of equity and net income to stockholder share equity thus it was appropriately labeled, **Inflation and Interest rate expectations.** Lastly, component 3 was correlated with net worth Assets, fixed assets turnover frequency and cash total Assets, that why it labelled **Return on Assets.**

Logistic Regression:

Logistic regression was after defining the six predictor variables that were to be included in the model. As we can see from the summary statistics in **Figure (4)**, there is a slight issue in the calculation of our linear regression model due to the fact that five of the predictor variables were deemed to have p-values that were not statistically significant. However, the variables were ultimately still included in the final model due to the historical significance of the predictor variables. VIF statistics were then run on the model as seen in **Figure (5)**, showing all predictor values having VIF scores lower than 2.7 which suggests that multicollinearity is not an issue for out model. In order to understand how accurate our model was at predicting the binary dependent variable “Bankrupt.” we ran the logistic model through a fitted results test, that summarized the totally accuracy and misclassification error of the model. As seen in **Figure (6)**, the total model accuracy and predicative ability was clocked in at a very high 96 percent, which suggests that our model is highly accurate at predicting the “Bankrupt.” dependent variable. In order to further capture the accuracy of our logistic regression, an ROC curve and the area under the curve score was calculated. The ROC plot displayed in **Figure (7)**, depicts a promising curve which suggests the test is accurate in nature. A non-accurate test would depict a curve which veered at a 45-degree angle shortly after the 0.0 value on the x-axis. Finally, the area under the curve was calculated in order to illustrate the predictive ability of the model, an auc score of 1 is optimal. As seen in **Figure (8)**, the auc score is .88, providing further evidence that the logistic regression model is accurate in its predictive ability. Overall, despite the lack of significant values displayed in the summary statistics of the regression model, it is clear that the model, utilizing the six predictor variables, was successful in it’s ability to predict the “Bankrupt.” dependent variable.

Limits of Application:

There are limitations to the conclusions that we were able to draw regarding bankruptcy prediction. It is our opinion that the largest of these limitations has to deal with the universal application of the conclusions drawn from our analysis. Specifically, we have identified two major issues with the original dataset that would deter our findings from being used in an attempt to holistically assess bankruptcy prediction. For one, the dataset is unique to the Taiwanese market. The implication from this issue is that while our model may be accurate at predicting bankruptcy in Taiwanese firms, that is not to say that our model would lend itself to other international markets. The second issue is that the dataset is only analyzing large firms. The implication from this issue is that our findings are not designed for small firm or individual bankruptcy prediction. Another limitation to the application of our conclusions is that fact that, our statistical methods and modeling techniques are not the most advanced techniques utilized in the fields. Neural networks and advanced machine learning techniques are superior to our more simplistic methods at being highly accurate in their prediction of firm bankruptcy.

**Conclusion:**

Our analysis indicates that more simplistic statistical methods, PCA and logistic regression, are able to accurately predict bankruptcy in large firms. The empirical results obtained from our PCA and logistic regression analyses suggest that the aforementioned methods, despite not being nearly as complex as advanced machine learning techniques, are still able to produce acceptable prediction results. Our findings are in line with previous research involving the use of multivariate statistical analysis of bankruptcy prediction.

We were able to determine that PCA can still be utilized as a useful vehicle in which to classify attributes pertaining to bankruptcy prediction. The three components identified from the PCA (Earnings Per Share, Inflation and Interest rate expectations, Return on Assets) allow researchers, governments, and investors to easily identify the financial components that are most likely to illustrate bankruptcy in large firms. The three components produced help to summarize the most important contributors of bankruptcy while remaining interpretable to the average investor.

Our analysis also revealed that logistic regression proved to be effective at predicting factors that determined bankruptcy while also remaining simplistic in nature. While only two of the six predictor variables revealed to be statistically significant, when tested, the model was still able to be extremely accurate at whether or not a firm was bankrupt or not. The results from this analysis prove that, these non-advanced techniques still have merit within the field of bankruptcy prediction. Logistic regression, while being far less complex than the advanced neural networks and machine learning AI, can still be utilized to produce accurate predictive results. The Logistic regression concluded that: ROA, EPS, Current Ratio, Cash Flow to Total Assets, Cash Flow to Liability, and Gross Profit to Sales can all be utilized to predict accurately. Our results suggest that the historical studies that the final model was built off of, many of which are now over a half-century old, were statistically sound in their methodologies.

Ultimately, while neural networks and AI will undoubtedly continue to dominate the field of bankruptcy prediction moving forward, our analysis revealed that using basic multivariate statistical methods can still be an effective way to determine the factors that determine bankruptcy in firms.

**References:**

[1]. Burlacu, S., Gutu, C., & Matei, F. O. (2018). Globalization–pros and cons. Calitatea, 19(S1), 122-125.

[2]. FitzPatrick, P. 1932. A comparison of ratios of successful industrial enterprises with those of failed companies. The Certified Public Accountant (October): 598-605, 656-662, and 727-731, respectively.

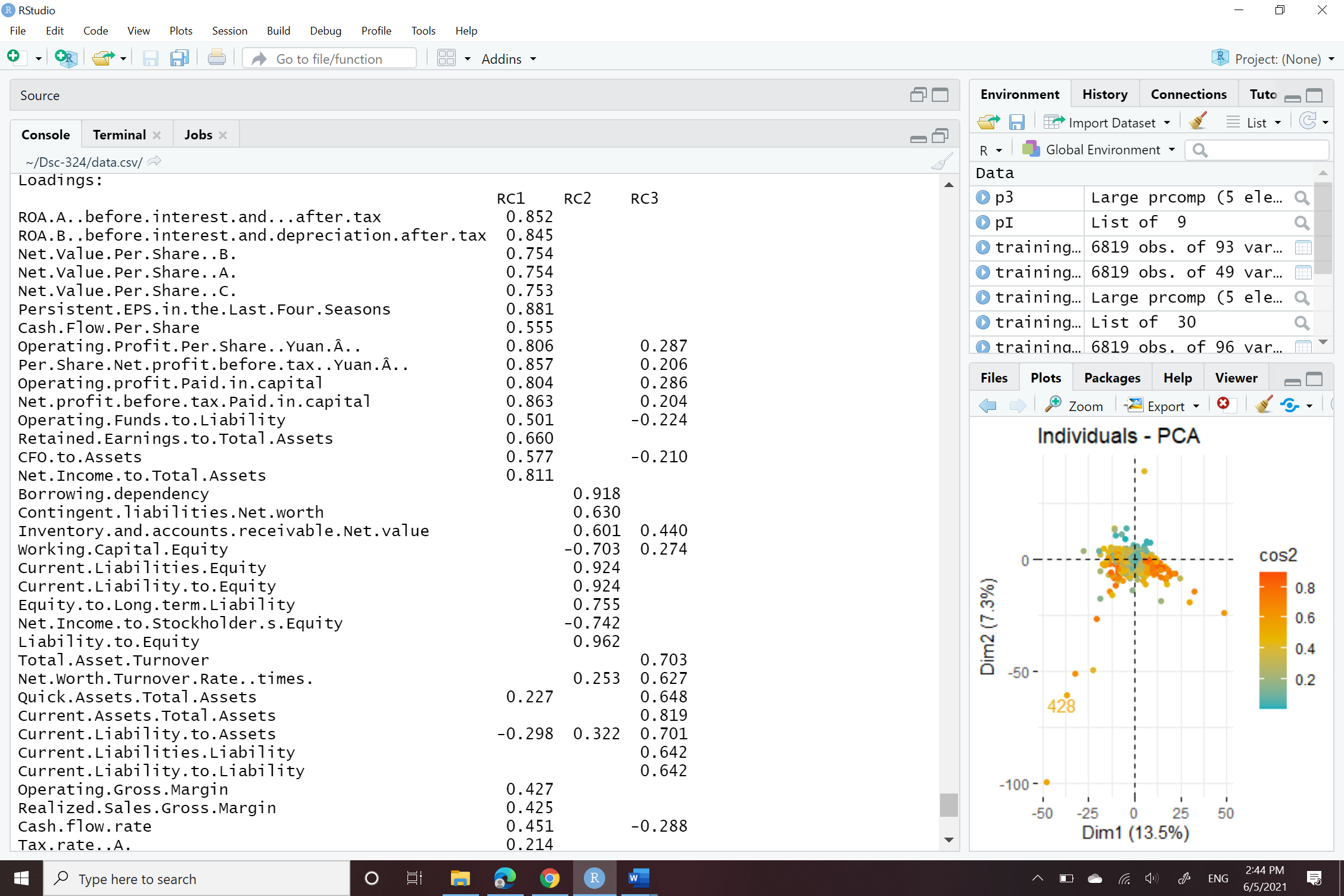
[3]. Tsukuda, J., & Baba, S. I. (1994). Predicting Japanese corporate bankruptcy in terms of financial data using neural network. Computers & Industrial Engineering, 27(1-4), 445-448.

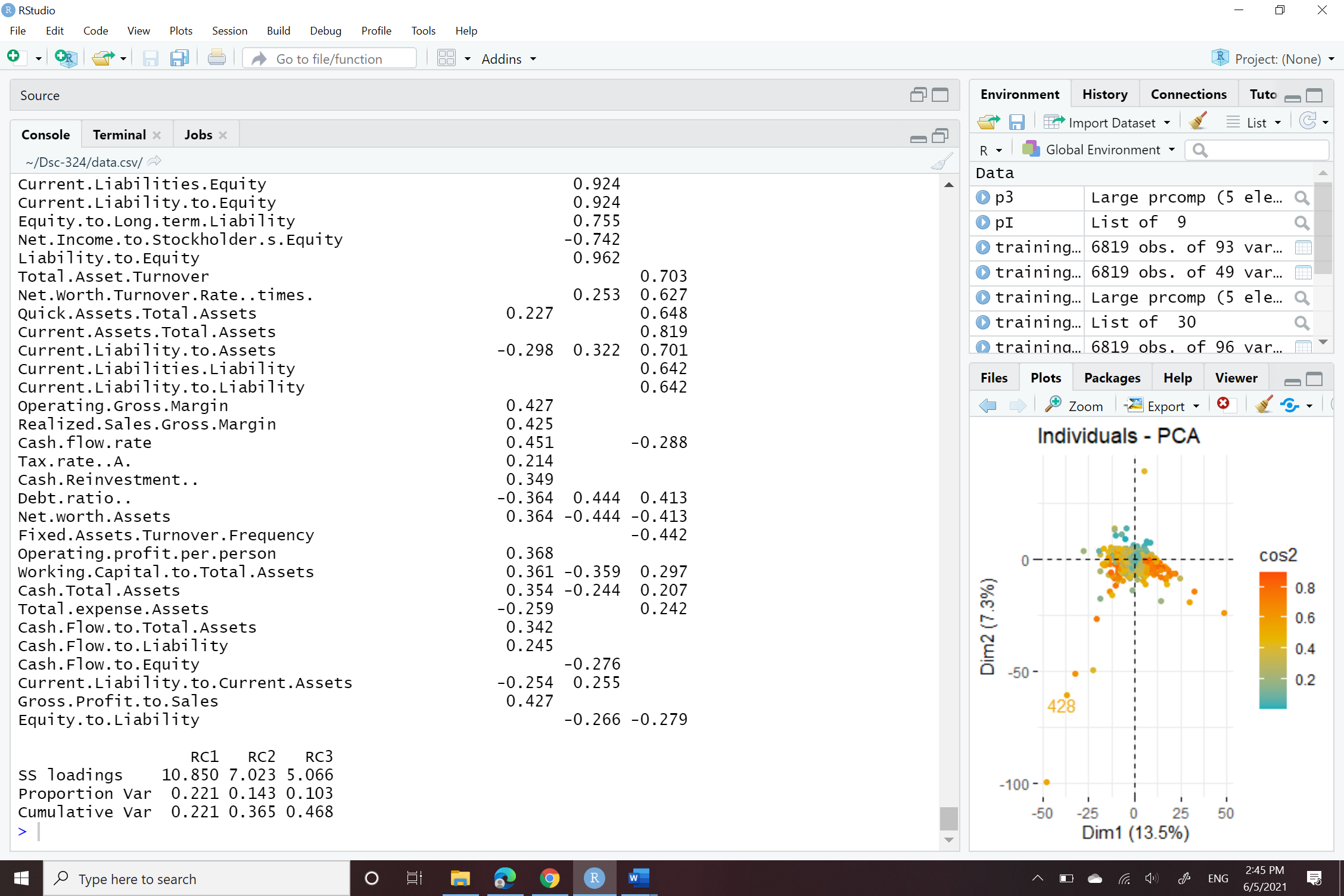
[4]. Bellovary, J. L., Giacomino, D. E., & Akers, M. D. (2007). A review of bankruptcy prediction studies: 1930 to present. Journal of Financial education, 1-42.

[5]. Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The journal of finance, 23(4), 589-609.

[6]. Beaver, William H. "Financial ratios as predictors of failure." Journal of accounting research (1966): 71-111.  
[7]. Chen, M. Y. (2011). Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches. Computers & Mathematics with Applications, 62(12), 4514-4524.

**Tables and Figures:**





**Figure -1**

Chart, histogram

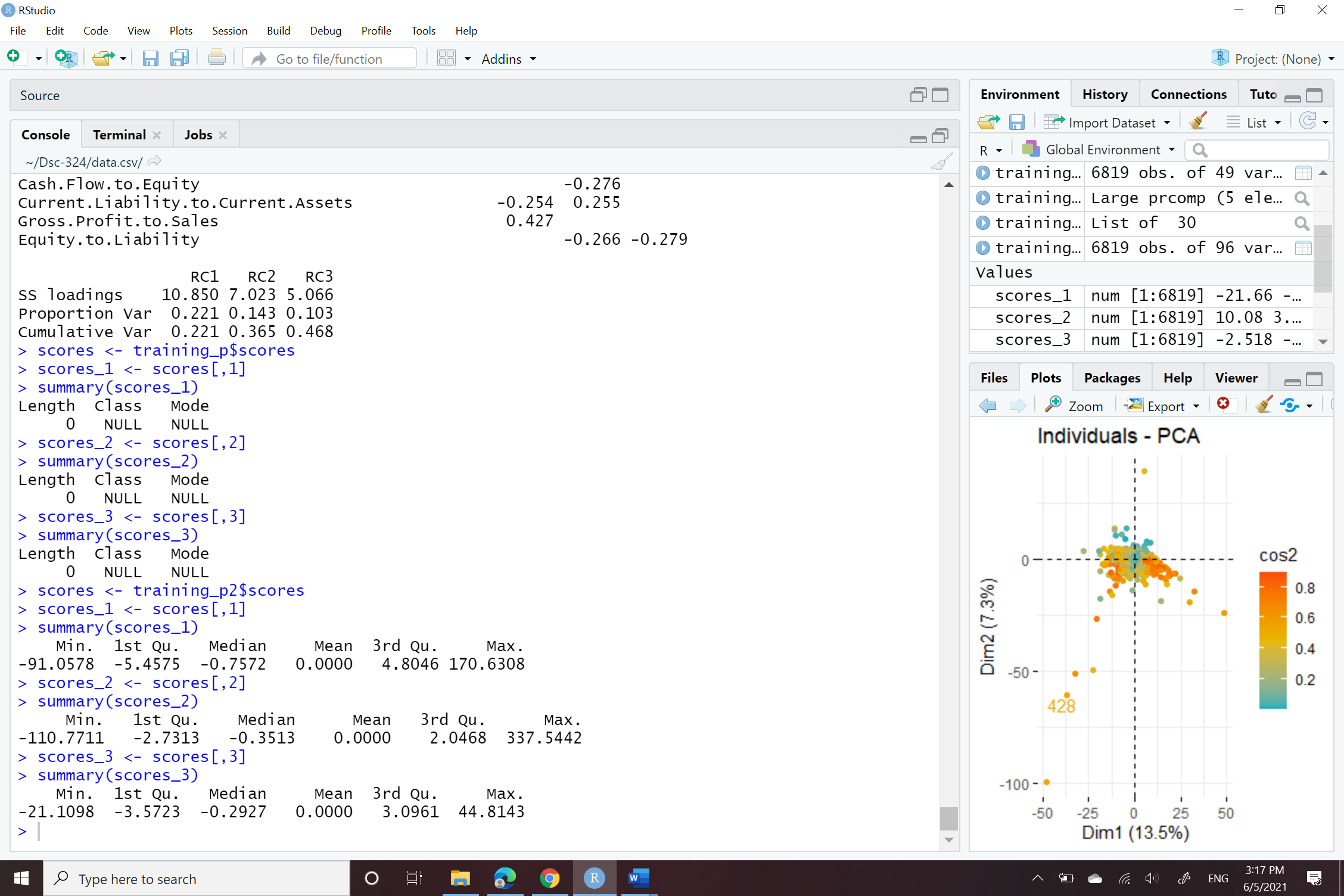
Description automatically generatedChart, histogram

Description automatically generated

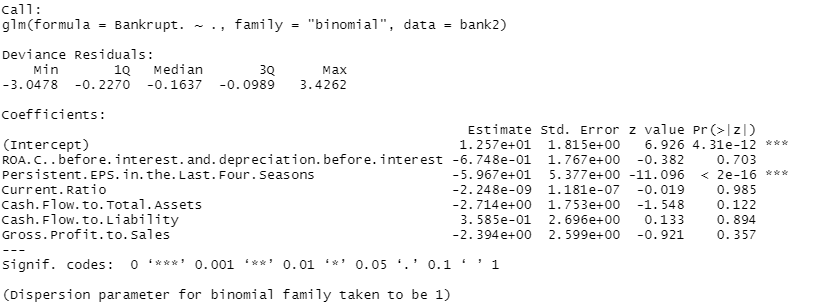
**Figure -2**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Earnings Per Share** | **Inflation and Interest rate expectations** | **Return on Assets** |
| **loadings** | 10.85 | 7.02 | 5.06 |
| **Proportion Var** | 22.1% | 14.3% | 10.3% |
| **Cumulative Var** | 22.1% | 36.5% | 46.8% |

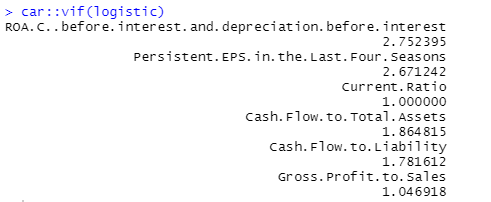
**Table -1**



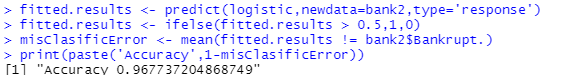
**Figure -3**



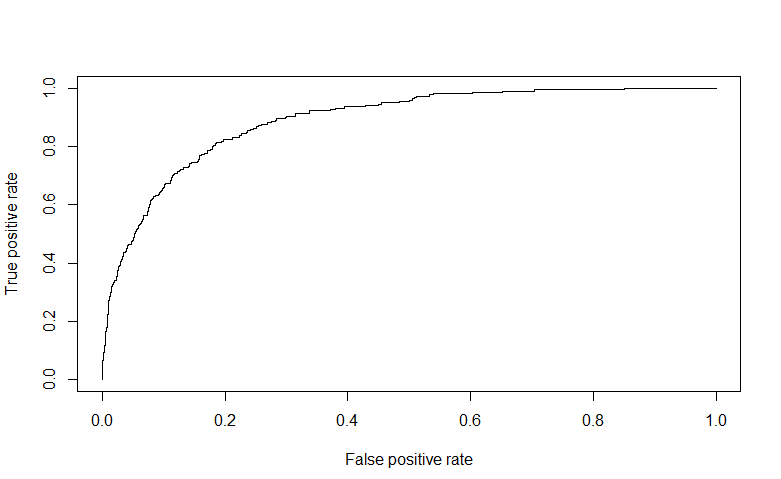
**Figure -4**



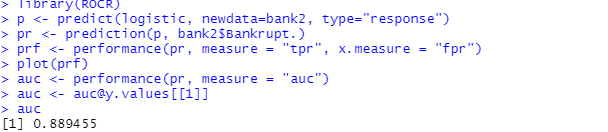
**Figure -5**



**Figure -6**



**Figure -7**



**Figure -8**