Analysis of Dividend Stocks Using PCA and LDA Methods

Presented by: *Ravish Kamath, Vi Nguyen, George Zhu & Yutong Pan*

Introduction

Objective and Data Set

- **Data Set:** 200 stocks with **6 continuous features** and **1 binary** variable of interest
- **Objective:** We want to **classify** whether a stock **pays dividend** or **not**.

Feature Details:

- ***Dividend:*** 0 for dividend; 1 for no dividend
- ***Fcfps (Free cash flow per share $):*** Useful to gauge the return a shareholder receives after buying a stock.
- ***Earnings Growth (year %):*** Change in an entity's reported net income.
- ***Debt to Equity Ratio:*** Measure of the extent to which a co5mpany covers its debt.
- ***Market Cap:*** Measure worthiness of a company in stock market.
- ***Current Ratio:*** Measures company's ability to pay its short-term obligation.

Method

Statistical Software: R

Principal Component

- **Goal: Reduce the dimension** of our explainable features.
- **Process: Eigendecomposition** on the **scaled** explanatory variables. The **eigen vectors** will be our **principal components**, which is a **linear combination** of our explanatory variables. We will **choose the component** that best explains the variation in our data through a **Scree plot** and the **cumulative variation proportion**.
- **Equations:**

$$Y_1 = PC_{11}Fcfps + PC_{21}Earn + PC_{31}DE + PC_{41}Mrkt + PC_{51}CR$$

$$Y_2 = PC_{12}Fcfps + PC_{22}Earn + PC_{32}DE + PC_{42}MC + PC_{52}CR$$

$$Y_3 = PC_{13}Fcfps + PC_{23}Earn + PC_{33}DE + PC_{43}MC + PC_{53}CR$$

$$Y_4 = PC_{14}Fcfps + PC_{24}Earn + PC_{34}DE + PC_{44}MC + PC_{54}CR$$

$$Y_5 = PC_{15}Fcfps + PC_{25}Earn + PC_{35}DE + PC_{45}MC + PC_{55}CR$$

Linear Discriminant

- **Purpose: Classification and dimension reduction**. Calculates a linear combination of independent features to classify data into classes by **maximizing separation between projected samples**.
- **Requirement:**

1) Data set must be **continuous**.
2) Distance between 2 projected classes must be **large**. Projection **variance** is **small**.

- **Process:** Given the requirement above, it projects the observations into 2 classes on a **line**.

1) Calculate the **means** of the 2 classes' projections μ_1 and μ_2
2) Calculate **average of** μ_1 and μ_2 which is the **cut-off value**. This is the linear boundary which is perpendicular to the projection line, separating the classes.
3) Projected sample can be classified by **observing which side of the line it falls in**.
4) Perform **1000-fold cross-validation** with 80% data in training. Want to identify misclassification rate.
5) Perform real life data classification

Results

Principal Component Analysis

	PC1	PC2	PC3	PC4	PC5
Fcfps	-0.4910	0.0514	-0.1035	-0.7707	0.3894
Earn	-0.2747	-0.9110	0.2843	0.1038	0.0549
DE	0.4075	-0.3970	-0.7960	-0.2017	-0.0447
MC	-0.5342	0.0345	-0.2363	-0.0349	-0.8102
CR	-0.4817	0.0930	-0.4679	0.5944	0.4324
Cum. proportion	0.4273	0.6129	0.7615	0.8943	1.0000

Figure 1: Principal Components

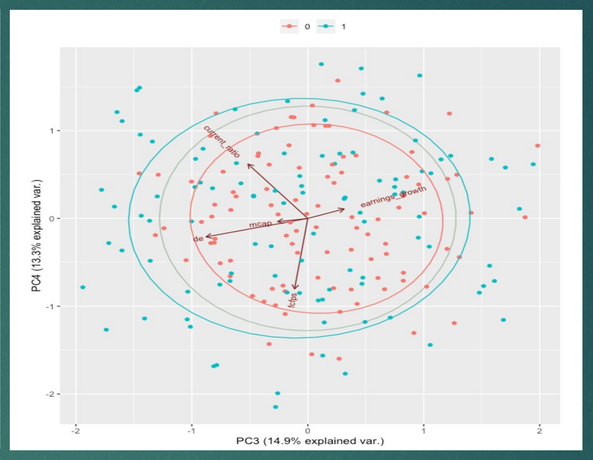

Figure 2: Biplot

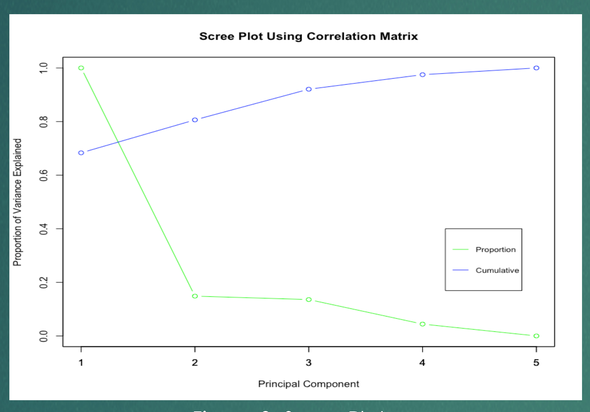

Figure 3: Scree Plot

Interpretation

- **PC1:** Primarily a measure of DE. Low value stock tends to have high DE ratio

Decision

- Use PC1, ... , PC4 for data reduction.
- We chose this based on the cumulative proportion value.
- It is the best PC that explains most of the variation.

Linear Discriminant Analysis

Without Principal Components:

Predictor	Estimate	Mean 0	Mean 1
Fcfps	0.4338	1.5001	3.0017
Earn	0.0280	5.9086	17.0536
DE	-0.2745	2.5859	1.7084
MC	0.0051	280.0408	550.4118
CR	0.8640	1.0632	1.9241

Figure 4: LDA information without PC

With Principal Components:

Predictor	Estimate	Mean 0	Mean 1
PC1	-1.3052	1.2617	-1.2123
PC2	-0.1479	0.0621	-0.0597
PC3	-0.2240	0.0753	-0.0724
PC4	0.0084	-0.0724	0.0024

Figure 5: LDA information with PC

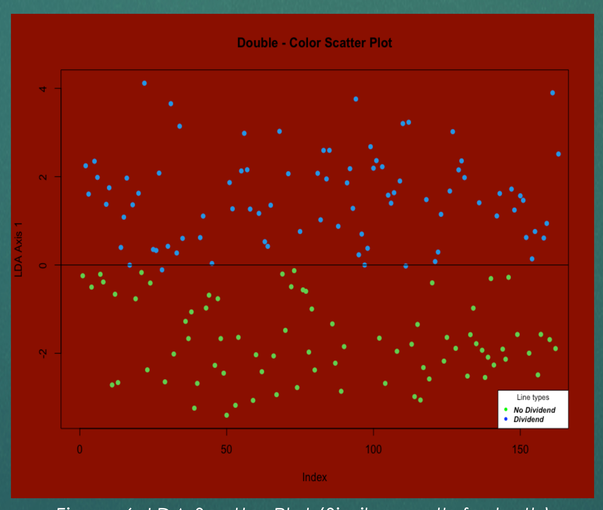

Figure 6: LDA Scatter Plot (Similar results for both)

LDA Method	Accuracy Rate	Misclassification Rate
Without PC	97.3%	6.53%
With PC	97.3%	6.57%

Figure 7: Testing of Model Results

Decision

- Based on Figure 7, since the model with the principal components gives us almost the same accuracy as the full model, we would choose that one as the best model to classify dividend stocks.

Discussion

- We can see that doing PCA has made the model simpler
- We still get the same testing results when we try LDA on all the features.
- Unfortunately, we could only reduce it by 1 dimension

Real Life Data Classification Test

- We can test our data with a real-life stock: Jumia Technologies
- Jumia is a publicly traded stock and does not issue dividend
- Jumia: Fcfps = -1.84 ; Earn = 11.87; DE = 0.0629 ; MC= 474.42 & CR = 2.62.
- Unfortunately, when we try to classify this stock, we get a misclassification
- Clearly shows that data may be too old. Inflation and other factors may affect the accuracy of the model.

Conclusion

- By using PCA and LDA, we were able to get a simpler and accurate model, based on the given data presented.
- In the future, it is best to use the most recent data, in order to better classify current stocks that pay or don't pay dividend.

Strengths & Limitations

Strengths

- Simple model and easy to understand whether a company issues dividend or not.
- This data had no missing values, so we did not need to do an extensive exploratory analysis
- Reduced the features and still managed to get the same result, without loss of accuracy

Limitations

- Unable to identify the industry category for each stock. If we had this information, we can better identify whether the ratios (DE & current) are good or bad.
- Reduction in dimension using PCA was little effect due to the small level of correlations between each feature.
- LDA is more suited for multi-class variable. Better visualization through scatter plot.
- Data set is quite old, since when we tried to predict with a real life data, it misclassified it.

References

- Wikimedia Foundation. (2022, December 5). *Linear discriminant analysis*. Wikipedia. Retrieved December 5, 2022, from https://en.wikipedia.org/wiki/Linear_discriminant_analysis
- Dash, S. K. (2022, August 5). *Linear discriminant analysis: What is linear discriminant analysis*. Analytics Vidhya. Retrieved December 5, 2022, from https://www.analyticsvidhya.com/blog/2021/08/a-brief-introduction-to-linear-discriminant-analysis/
- *Neuralnet: Train and Test Neural Networks using R*. DataScience+. (n.d.). Retrieved December 5, 2022, from https://datascienceplus.com/neuralnet-train-and-test-neural-networks-using-r/
- Zach. (2020, October 30). *Linear discriminant analysis in R (step-by-step)*. Statology. Retrieved December 5, 2022, from https://www.statology.org/linear-discriminant-analysis-in-r/
- *What is the difference between PCA and Lda?* 365 Data Science. (2022, July 15). Retrieved December 5, 2022, from https://365datascience.com/tutorials/python-tutorials/lda-vs-pca/
- Wikimedia Foundation. (2022, November 10). *Principal component analysis*. Wikipedia. Retrieved December 5, 2022, from https://en.wikipedia.org/wiki/Principal_component_analysis
- Team, T. A. I. (2022, January 26). *LDA vs. PCA*. Towards AI. Retrieved December 5, 2022, from https://towardsai.net/p/data-science/lda-vs-pca