Project : Multivariate Analysis of Spotify most streamed Songs

a summary report authored by
Yuyutsu Saini
2020MT60571

supervised by
Prof. Rahul Singh

06/11/2024

MTL766 : Multivariate Statistic

Contents

1 Introduction

The objective of this analysis was to perform a multivariate analysis on two datasets containing attributes related to user purchases and loyalty scores. Various statistical tests, regression models, and dimensionality reduction techniques were employed to uncover insights from the data.

2 Dataset Description

The datasets consist of user-level data with attributes such as 'age', 'annual_income', 'purchase_amount', 'loyalty_score', and 'purchase_frequency'. The two datasets represent different regions: North-East and South-West.

3 Exploratory Data Analysis

Observations:

- The pairwise relationships between attributes show a strong correlation, suggesting interdependency among the variables.

Conclusion:

- All attributes in both datasets are highly correlated, indicating that they are likely to influence each other.

3.1 Comparison of Covariance Matrices

3.1.1 Box M Test

The Box M test is used to test the hypothesis that the covariance matrices of two or more groups are equal. A small p-value indicates that the covariance matrices are significantly different.

Test Statistic:

$$M = 100.776012$$

Degree of Freedom:

$$df = 15$$

P-value:

$$p = 9.297145 \times 10^{-15}$$

Conclusion: The p-value is extremely small, indicating that we reject the null hypothesis that the covariance matrices are equal. Therefore, we conclude that the covariance matrices of the two datasets are significantly different.

3.1.2 Likelihood Ratio Test

The Likelihood Ratio Test is used to compare the fit of two models, typically a restricted and an unrestricted model. In this case, it tests the equality of covariance matrices between two datasets.

Test Statistic:

$$\Lambda = 3.499576610159108 \times 10^{-23}$$

Degree of Freedom:

$$df = 13$$

P-value:

$$p = 3.330669 \times 10^{-16}$$

Conclusion: Given the very small p-value, we reject the null hypothesis that the covariance matrices are equal. This indicates that there is a significant difference in the covariance structures of the two datasets.

3.2 Comparison of Mean

3.2.1 Welch's t-test

Welch's t-test was performed to compare the means of the two datasets for each attribute, with the results provided below:

- **Age:** t-statistic = -0.8109, p-value = 0.4643
- **Annual Income:** t-statistic = -0.0414, p-value = 0.9690
- **Purchase Amount:** t-statistic = 0.0898, p-value = 0.9329
- **Loyalty Score:** t-statistic = 0.1208, p-value = 0.9108
- **Purchase Frequency:** t-statistic = 0.2500, p-value = 0.8194

Conclusion: Since the p-values for all the attributes are greater than 0.05, we fail to reject the null hypothesis for each attribute. This suggests that there is no significant difference between the means of the two datasets for any of the attributes.

3.2.2 James' Test

James' test is used to compare the means of two datasets when the variances of the datasets may differ. It is based on a generalized form of the F-statistic and is used to detect whether the means of the two datasets are significantly different.

Test Statistic:

$$\text{James' Test Statistic} = 78.5758$$

Critical F-value:

$$F_{critical} = 8.7901$$

P-value:

$$p = 0.0020$$

Conclusion: The p-value is less than 0.05, indicating that we reject the null hypothesis. This suggests that the means of the two datasets are significantly different, and there is strong evidence to support this conclusion.

4 Outlier Detection using Mahalanobis Distance

Observations:

- 8 outliers were detected in Dataset 1, and 24 outliers were detected in Dataset 2 using Mahalanobis distance.

Conclusion:

- A significant number of outliers were identified, suggesting anomalies in the datasets.

4.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms the data into a new coordinate system, where the first few components retain most of the variance in the data. In this analysis, PCA was applied to reduce the dimensionality of the dataset.

Eigenvalues:

$$\lambda_1 = 4.9398, \quad \lambda_2 = 0.0256, \quad \lambda_3 = 0.0035, \quad \lambda_4 = 0.0182, \quad \lambda_5 = 0.0129$$

Observations:

- The first eigenvalue ($\lambda_1 = 4.9398$) is significantly larger than the others, indicating that the first principal component captures the most variance in the data.
- The remaining eigenvalues are much smaller, suggesting that the remaining components explain much less variance.
- This supports the finding that the attributes are strongly correlated, as the first few components explain most of the variance in the dataset.

Conclusion:

- PCA was effective in reducing the dimensionality while retaining most of the variance in the dataset. The high eigenvalue for the first component indicates that a single component can capture most of the data's information.

4.2 Linear Regression

Linear regression is used to model the relationship between a dependent variable and one or more independent variables. In this analysis, we performed linear regression for various pairs of attributes in the dataset.

R^2 Values for Regression:

$$R^2(\text{Age vs Purchase Amount}) = 0.9725$$
$$R^2(\text{Age vs Annual Income}) = 0.9503$$
$$R^2(\text{Age vs Loyalty Score}) = 0.9640$$

Observations:

- The R^2 values are very high, indicating that the linear models explain a large proportion of the variance in the data.
- Specifically, the regression model between Age and Purchase Amount has the highest R^2 value ($R^2 = 0.9725$), suggesting a very strong linear relationship between these two variables.
- All the models show a good fit, with R^2 values above 0.95, indicating that the linear regression models can predict these attributes with high accuracy.

Conclusion:

- Linear regression has shown to be an effective model for explaining the relationships between Age and other attributes like Purchase Amount, Annual Income, and Loyalty Score, with high R^2 values.

5 Softmax Regression for Region Classification

Observations:

- The softmax regression model achieved an accuracy of 43.28

Conclusion:

- The softmax model's low accuracy reflects the complexity of the dataset and the lack of linear separability.

6 Conclusion

- The datasets exhibit strong correlations between the attributes, confirming interdependence.
- The Box M test and Likelihood Ratio Test both rejected the null hypothesis, indicating significant differences in covariance and means between the datasets.
- Outlier detection identified several anomalies in both datasets.
- PCA effectively reduced the dimensionality while retaining most of the variance, confirming the high correlation between attributes.
- The linear regression models demonstrated strong predictive power for certain attributes.
- The softmax regression model showed low accuracy, reflecting the complexity and lack of linear separability in the data.