CLUSTERING AND FITTING IN AIRLINE DATASET

Introduction

The goals of this assignment are to examine patterns and patterns in the supplied data through data science approaches such as clustering and regression analysis. The dataset was obtained from secondary data sources about the date, number and revenue. The raw materials include numerical records and revenue collected on a daily basis The data set confirms business performance for a particular time period. Data pre-processing and understanding, application of the K-means clustering technique, and application of the linear regression technique are part of the analysis process in the present study. The relevance of clustering in segmentation and of regression in outcome analysis is highlighted.

Data Preparation

The data were cleaned and preprocessed in order to prepare it for analysis, removal of any unnecessary text and periods were also done. The Date column was also an object that was converted to a DateTime format for time-based data representations and data analysis. To check for any deviations, for example, data entry errors, or outliers, summary statistics were computed. The missing value handling guidelines can be discussed with reference to the dataset, which was deemed mostly complete, although options such as imputation or removal can be considered.

Feature scaling was also performed on the Number and Revenue attributes, to bring them to equal scale for better clustering accuracy. This step was important to prevent the k-means algorithm from being influenced by the some large variables while leaving out more relevant small variables. Other exploratory features like 'Month of the week', which are a month and the day of the week in which the incidents occur were derived to capture the seasonal trends. The cleaned dataset was checked for assessment of validity in order to create a more suitable background for making qualitative designs, clusters and regression types.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) was first done in order to get basic intuition and find preliminary patterns in given datasets. Descriptive analysis of the data showed the differences between Number and Revenue and where more detailed exploration could be utilised. The use of visuals was made with a view of increasing appreciation.

The distribution of Revenue was analysed to identify its distribution by constructing a histogram. This means that the additional data represented here manifested a kind of bias concerning the revenue between days. A line plot of the daily trend of Revenue was also constructed to identify temporal factors, which possibly influence the Revenue. In the same manner, autoregressive features involving the Number were subjected to testing against time in order to pick out significant patterns or periodicity.

In the case of the present study, a Scatter plot was used to represent the relationships among the variables. Positive levels of correlation coefficients ranging from moderate to high were established between Number and Revenue implying linear relationships. Based on this insight, the choice of these variables for regression analysis was done respectively. Another scatter plot was created to highlight revenue over time of 3 years. It was observed that the highest revenue was secured from May 2023 to September 2023 which is 230 units.

Cluster Analysis

Using the Number and Revenue as decision variables, k-means clustering was used to partition the dataset into different clusters. Results from the Elbow Method were used to ascertain the optimal number of clusters, which was found to be three. The clustering results revealed three distinct groups which are,

- **Cluster 0:** This group comprises the biggest part of the data and has an average number of 9.26 and an average revenue of 45.37. It depicts slow to moderate levels of business activity, meaning normal business activities for which organizations are likely to generate moderate levels of revenue.
- **Cluster 1:** This group shows the highest coefficient, having an average Number of 26.61 and an average Revenue of 139.92. It represents bursts of greatly increased traffic, when there is much more than average sales, probably days with much traffic or any event.
- **Cluster 2:** This group has an average score of moderate values, with an average number score of 17.22 and an average Revenue score of 87.08. It accounts for a middle level of business activity, signifying changes in demand and levels of revenue generation.

The above pair plot is presented to visualize the correlation of numbers and revenue with respect to each cluster. Based on the clustering analysis it was demonstrated that these three clusters can be attributed to varying levels of business performance, with Cluster 1 indicating peak periods and Clusters 0 and 2 corresponding to more stable and moderate activity, respectively.

Regression Analysis

A linear regression model was applied to predict Revenue based on the Number variable. The results of the regression analysis are,

- Intercept: -5.6101

- Coefficient (Slope): 5.4404
- Mean Squared Error (MSE): 1.7203
- R-squared Value: 0.9989

The coefficient estimate for the intercept term is negative – this means that the predicted revenue if the Number is zero is about -5.61 but this does not hold much simple relevance in this case. From the coefficient of 5.4404, it can be interpreted that for every unit change in number, an equivalent of 5.44 units of change is expected in revenue, implying a positive linear relationship.

The low MSE value (1.7203) confirms a tight fit of the model, that is, there are small differences between the actual and the predicted Revenue. The possibility to explain 99,89% of the model is high, and proves that R-squared of 0,9989 is high for the relationship between the two models – so it can be described as a very good match for the regression model while using Number as the explaining variable. The regression equation provided and the value of R-squared corroborate this suggestion as the Number variable is shown to have a very high significance for the prediction of Revenue.

Discussion and Performance Assessment

The actual implementation of the k-means clustering analysis produced three clusters to represent the dissimilar intensity of business operations. Cluster 0 which had moderate Numbers and Revenue was able to capture routine Activity. In Cluster 1, Number and Revenue were high; these participants identified specific business hours with a higher demand. Cluster 2 was characterized by intermediate activity levels. The element of clustering was useful, as it introduced sets of outcomes that reported various working conditions. However, the problem with using the k-means clustering approach is the fact that they do not presuppose the fact that clusters may unequal and spherical as may be the case with actual distributions. Outliers or even noise in the data could add worse to the quality of clusters.

When using the same two variables, Number and Revenue, the regression model had a very high value of determination squared of 0.9989 and this was evidence of the fact that the model used explained the variation of Revenue through Number appropriately. From the reported results, the direct line passing through the two variables was obvious with the slope = 5.44. However the regression model used supposes that there is a direct interaction between Number and Revenue variables only, which may be insufficient to reveal other forms of connection.

In comparing the two approaches, clustering brought out various periods of business activities while the regression model gave a quantitative robust prediction of Revenue given Number. Thus, the two techniques were both beneficial, clustering provided a set of interest groups, while regression provided prediction capability. However, suggestions can be made on their respective restrictions for further elaboration.

Conclusion

Therefore, from the analysis it was obtained important patterns that help in the understanding of the dataset through the use of clustering and regression analysis. Analysis using the K-means clustering algorithm was successful in capturing three different business activity levels, each associated with a different intensity of operational load. A similar positive linear regression of `Number` to `Revenue` emerged; the line fit fairly well with an R-squared value of 0.9989. Both approaches proved to be informative but clustering provided information concerning the temporal distribution of the business activities while regression aimed at estimating the revenue given the `Number' variable. These techniques were useful overall for getting a sense of business performance although some concerns were made about the assumptions made by the methods.

Bibliography

Arachchi, T.G., Dahanayaka, M. and Perera, H.N., 2024, April. Analyzing Sustainability Initiatives of the Airline Industry Through Random Forest Classification and K-Means Clustering Techniques. In *2024 New Trends in Civil Aviation (NTCA)* (pp. 179-184). IEEE.

Tian, H., Presa-Reyes, M., Tao, Y., Wang, T., Pouyanfar, S., Miguel, A., Luis, S., Shyu, M.L., Chen, S.C. and Iyengar, S.S., 2021. Data analytics for air travel data: a survey and new perspectives. *ACM Computing Surveys (CSUR)*, *54*(8), pp.1-35.

Vock, S., Garrow, L.A. and Cleophas, C., 2021. Clustering as an approach for creating data-driven perspectives on air travel itineraries. *Journal of Revenue and Pricing Management*, pp.1-16.

Yaakoubi, Y., Soumis, F. and Lacoste-Julien, S., 2020. Flight-connection prediction for airline crew scheduling to construct initial clusters for OR optimizer. *arXiv preprint arXiv:2009.12501*.

Yang, T., Pasquier, N. and Precioso, F., 2020, July. Ensemble Clustering based Semi-supervised Learning for Revenue Accounting Workflow Management. In *DATA* (pp. 283-293).