Performance Comparison of Two Statistical Learning Algorithms

GROUP -E

Introduction and Data Description

In today's competitive restaurant industry, accurately predicting revenue is essential for driving growth and profitability. Revenue serves as a key performance indicator, reflecting a restaurant's ability to attract customers, price menu items effectively, allocate resources, and execute successful marketing strategies. Understanding the factors influencing restaurant revenue can provide valuable insights into consumer behavior, market trends, and operational efficiency.

In this report, we focus on restaurant revenue prediction using machine learning techniques, specifically comparing the performance of two widely used algorithms: Linear Regression and Regression Decision Tree. Both algorithms, while designed to model relationships between variables, approach the task in different ways. Linear regression fits a linear equation to the observed data, making it well-suited for datasets with clear linear relationships between features. Regression Decision Trees, however, are more flexible, handling both linear and non-linear relationships by recursively splitting the dataset into subsets based on feature values.

The dataset used for this analysis consists of 1,000 observations and 8 features, including both numerical and categorical data. The numerical variables include **'Number_of_Customers'**, **'Menu_Price'**, **'Marketing_Spend', 'Average_Customer_Spending', 'Promotions', 'Reviews',** and **'Monthly_Revenue' (the target variable).** The categorical variable, 'Cuisine_Type', represents the type of cuisine offered by the restaurant. With a mix of data types and feature interactions, this dataset provides an ideal scenario for testing and comparing the effectiveness of linear regression and Regression Decision Tree algorithms in predicting restaurant revenue.

Aim

This report aims to evaluate these two models across various performance metrics, exploring their strengths, weaknesses, and suitability for different data structures and prediction scenarios.

DATASET LINK –

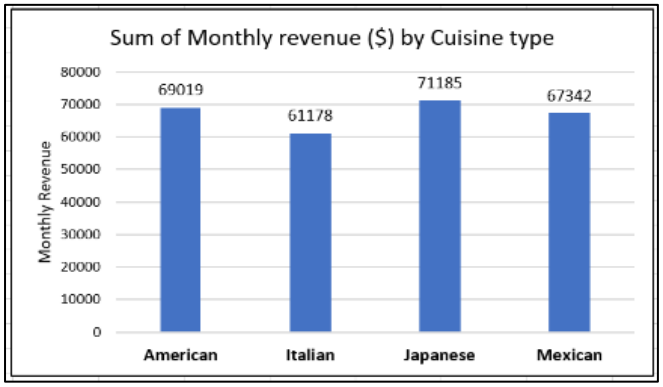
https://drive.google.com/file/d/1u9w6D3ANXfRUSEv5T-HPFAh7hSXxvIZ6/view

Data Information

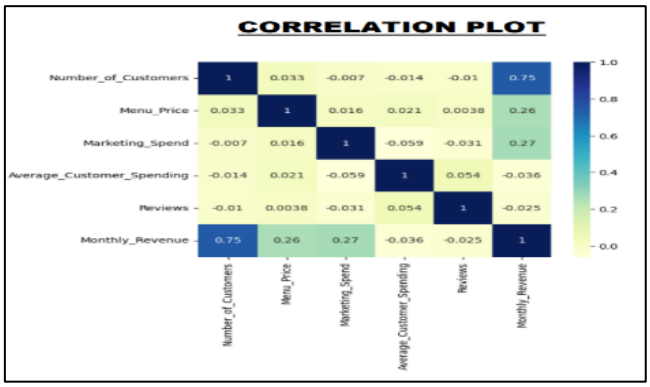
- Null values: There are no null values present in the dataset
- Duplicates: There are no duplicates in dataset

Exploratory Data Analysis

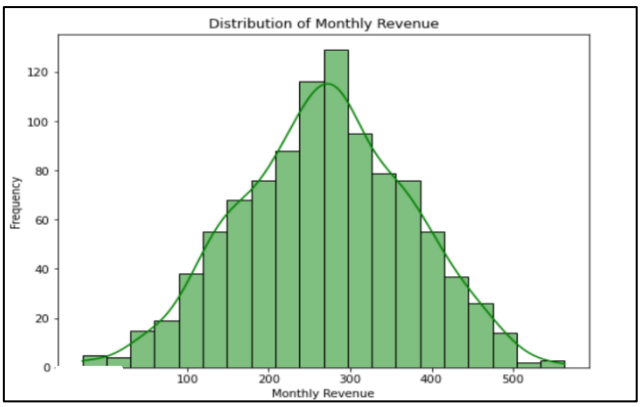
1. Drop column: We removed the "cuisine_type" and "promotions" columns from the dataset, possibly due to their lack of relevance or other considerations.
2. Outliers: Box plots were utilized to identify outliers in the dataset. Outliers, if present, could significantly impact the accuracy of predictive models and thus needed to be examined.
3. Histograms for Normality Check: Histograms were plotted for each remaining feature, allowing us to visualize their distributions and assess whether they approximate a normal distribution.
4. Correlation Plot: To analyse relationship between dependent and independent variables.



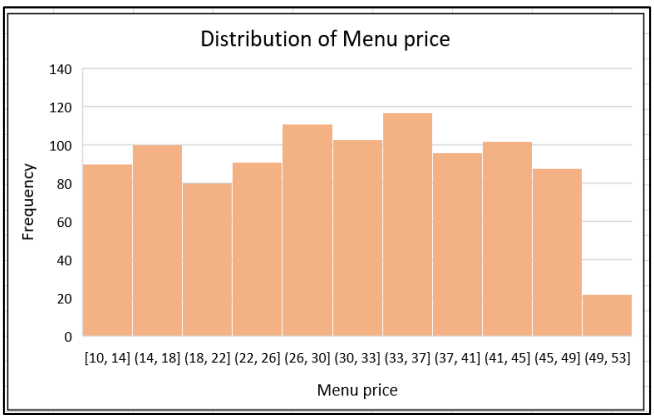
We observe uniformity almost across all cuisine types based on the monthly generated revenue. If taken a deeper look at the numbers, the chart suggests that Italian cuisine generates least while Japanese cuisine generates the most monthly revenue.



The number of customers shows a strong positive correlation with monthly revenue (0.745), while menu price (0.260) and marketing spend (0.270) have moderate positive correlations. Average customer spending (-0.036) and reviews (-0.025) show weak negative correlations with monthly revenue.



Our target variable,'Monthly_Revenue', demonstrates a relatively normal distribution cantered around mean equal to 270.

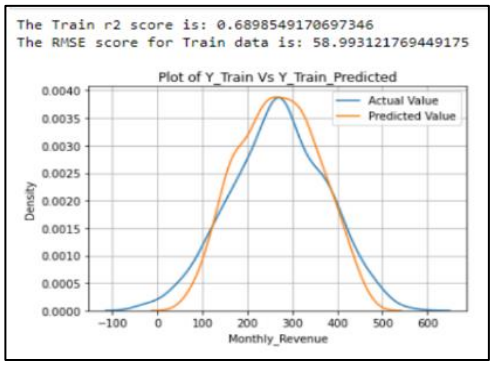


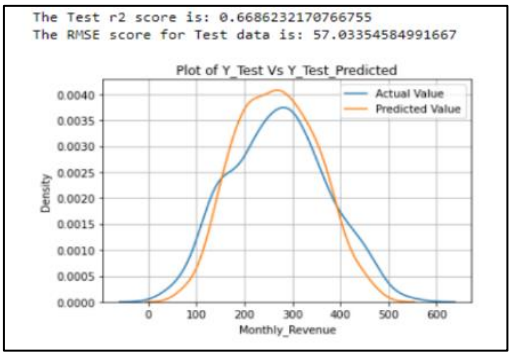
This conclusion is based on visually inspecting the histograms or density plots of all the features, which show differences from the usual bell-shaped curve that represents a normal distribution.

MULTIPLE LINEAR REGRESSION

The goal of MLR is to find a linear equation that best fits the data, allowing us to predict the dependent variable based on the multiple independent variables.

The general form of the MLR model is: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots\ldots + \beta_n X_n + \varepsilon$





❖ **Training Data Performance:**

➢ The linear regression model achieved an R^2 score of approximately 0.690. This indicates that approximately **69.0%** of the variance in the monthly revenue can be explained by the predictor variables incorporated into the model. In other words, the model accounts for a substantial portion of the variability in monthly revenue observed in the training dataset.

➢ Additionally, the RMSE score was approximately 58.99 units which suggests that, on average, the model's predictions deviated from the actual monthly revenue by 58.99 units. While this represents a level of prediction error, it is important to note that the magnitude of the RMSE score is relative to the scale of the target variable.

❖ **Test Data Performance:**

➢ The linear regression model exhibited consistent performance with an R^2 score of approximately 0.669. This indicates that approximately **66.9%** of the variance in the monthly revenue was captured by the predictor variables included in the model.

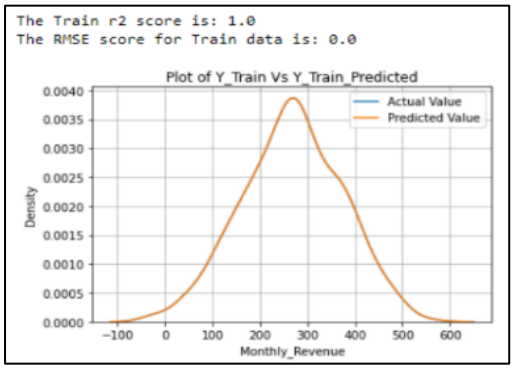
➢ Moreover, the RMSE score for the test data was approximately 57.03 units. This implies that, on average, the model's predictions deviated from the actual monthly revenue by 57.03 units. The similarity in performance metrics between the training and test datasets suggests that the model generalizes well to unseen data and maintains its predictive accuracy outside the training set.

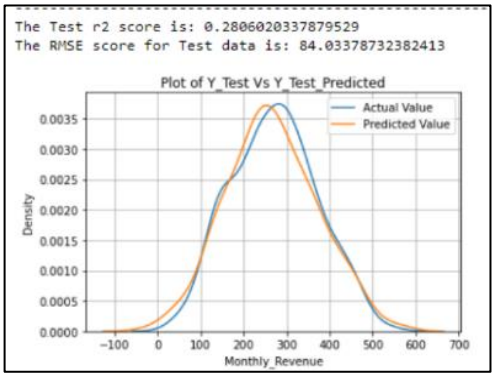
Advantages: MLR captures the effect of multiple variables on a dependent variable simultaneously.
Disadvantages: - Sensitive to outliers, Assumes linear relationships,Prone to multicollinearity.

Regression Decision Tree

A Regression Decision Tree works by splitting the dataset into smaller subsets based on feature values, creating a tree-like structure. At each split, the model selects the feature and value that minimize the prediction error (often measured using metrics like mean squared error).

The process continues until the tree reaches a specified depth or other stopping criteria. The final predictions are made by averaging the target values in the leaf nodes.





❖ Training Set:

➢ The coefficient of determination (R^2) score, which measures the proportion of variance explained by the model, is exceptionally high at 1.0. This indicates that the Regression Decision Tree model perfectly fits the training data, capturing all the variation in the target variable.

➢ The root mean squared error (RMSE) score, which quantifies the average deviation of the predicted values from the actual values, is remarkably low at 0.0. This suggests that the model's predictions align perfectly with the actual values in the training dataset, indicating an overfit.

❖ Test Set:

➢ The R^2 score for the test dataset is 0.281, indicating that the Regression Decision Tree model explains approximately 28.1% of the variance in the test data. While this is a positive indication of the model's performance, it suggests that there is room for improvement in capturing the variability of the target variable compared to the training set.

➢ The RMSE score for the test dataset is 84.03, indicating the average prediction error of the model on the test data. This value represents the average discrepancy between the predicted and actual values of the target variable, with lower values indicating better model performance.

Advantages: No need for feature scaling and Works with both categorical and numerical data.

Disadvatages: Prone to overfitting, especially with deep trees and Unstable, sensitive to small data changes.

Model comparison

Here is the comparison on model performance including R^2 scores for both training and test sets:

1. Linear Regression:

- Training Set R^2: 0.6899
- Test Set R^2: 0.6686
- Interpretation: Linear Regression shows a moderate R^2 score on both the training and test sets, indicating a decent fit of the linear regression model to the data.

2. Regression Decision Tree:

- Training Set R^2: 1.0000
- Test Set R^2: 0.2283
- Interpretation: Regression Decision Tree achieves a perfect R^2 score of 1.0000 on the training set, suggesting overfitting as it perfectly fits the training data. However, its performance drops significantly on the test set (R^2 = 0.2283), indicating poor generalization to unseen data.

Conclusion

The results demonstrate a notable contrast in the performance of Linear Regression (SLR) and Regression Decision Tree.

- Linear Regression provides a moderate R^2 score on both the training set (0.6899) and test set (0.6686). This suggests that the model fits the data well, generalizing reasonably well to unseen data. The consistent performance across both datasets indicates that the model is not overfitting and is able to capture the underlying trend in the data without being overly sensitive to noise.
- Regression Decision Tree, on the other hand, shows a perfect R^2 of 1.0000 on the training set, indicating overfitting. The model has perfectly learned the patterns in the training data, but its performance drastically drops on the test set (R^2 = 0.2283), indicating poor generalization. This sharp decline in performance demonstrates that the Regression Decision Tree is too specific to the training data and does not adapt well to new, unseen data.

The most important advantages of Linear Regression over Regression Decision Tree are:

1. **Better Generalization:**
   Linear Regression performs consistently on both training and test sets, ensuring the model generalizes well to unseen data, while Regression Decision Trees are prone to overfitting and have poor performance on new data.
2. **Less Overfitting:**
   Linear Regression is less complex and less likely to overfit the training data, unlike Regression Decision Trees, which can become overly specific to the training set, leading to poor generalization.

Thus, while Regression Decision Trees may excel in capturing complex relationships, their susceptibility to overfitting and instability makes **Multiple Linear Regression** a more reliable choice when generalization to unseen data is crucial.