Linear Regression Training Project: Ecommerce Clients

Introduction

This project analyses a dataset of e-commerce customers to understand their spending habits and determine whether the company should focus on improving their mobile app or website. The dataset, available on Kaggle, includes various features such as average session length, time spent on the app, time spent on the website, and length of membership.

Data Overview

The dataset includes the following features:

- Avg. Session Length: Average session length of in-store style advice sessions.

- Time on App: Average time spent on the app in minutes.

- Time on Website: Average time spent on the website in minutes.

- Length of Membership:Number of years the customer has been a member.

- Yearly Amount Spent:Amount of money spent by the customer in a year.

The dataset contains 500 entries with no missing values.

Exploratory Data Analysis

To explore the relationships between different features and the yearly amount spent, various visualizations were performed using Seaborn.

Time on Website vs. Yearly Amount Spent

The joint plot of time on the website versus yearly amount spent shows no significant correlation between these variables.

Time on App vs. Yearly Amount Spent

The joint plot of time on the app versus yearly amount spent suggests a small positive correlation. Customers who spend more time on the app tend to spend more money.

Pairplot of All Features

The pairplot visualizes the relationships between all features. The most notable correlation is between the length of membership and yearly amount spent, indicating that longer membership is associated with higher spending.

Length of Membership vs. Yearly Amount Spent

The regression plot of length of membership versus yearly amount spent confirms a strong positive correlation. Longer membership duration tends to result in higher spending.

Data Splitting

The data was split into training and testing sets to create a model that predicts yearly amount spent based on the input features.

Model Training

Using Scikit-Learn

A linear regression model was trained using the Scikit-Learn library. The coefficients of the model indicate the importance of each feature:

- Avg. Session Length:25.72
- Time on App: 38.60
- Time on Website:0.46
- Length of Membership:61.67

The length of membership has the highest impact on yearly spending, followed by time on the app and average session length. Time on the website has a minimal impact.

Using Statsmodels

A more detailed model was created using Statsmodels. The results confirmed the findings from Scikit-Learn, highlighting the significant impact of length of membership on yearly spending.

Model Evaluation

The model was evaluated using the test data to assess its performance.

Predictions

The scatter plot of actual versus predicted values of yearly amount spent shows that the model's predictions are accurate, with points closely clustered around the diagonal line.

Error Metrics

The model's performance was evaluated using mean absolute error, mean squared error, and root mean squared error:

-Mean Absolute Error:7.23

-Mean Squared Error:79.81

-Root Mean Squared Error:8.93

Residuals

The distribution plot of residuals (differences between actual and predicted values) indicates that the residuals are normally distributed, suggesting a good fit.

Conclusion

The analysis shows that the length of membership has the most significant impact on the yearly amount spent by customers. Between the app and the website, the time spent on the app has a stronger influence on spending. The time spent on the website appears to have minimal impact. This suggests that the company should focus on enhancing the mobile app experience to increase customer spending. Further insights from online marketing experts could provide more actionable recommendations based on these findings.

Reference

Dataset : Ecommerce customers by Saurabh Kolawale on Kaggle