Feature Selection: A linear regression approach to find the impact of the features of e-commerce sales data

Rafiq Islam

2022-08-30

Table of contents

Notebook

Project Overview

This is a preliminary level linear regression based machine learning project to investigate the feature importance for an e-commerce based company or simply building a predictive model to generate insights on different features.

Dataset

The data is collected from kaggle.com. It contains 500 observations with the following columns

- `Email:` Email address of the customers
- `Address:` Physical mailing address of the customers

- `Avatar:` The fancy avater of the customers
- `Avg. Session Length:` Average session lenght spent either on app or web
- `Length of Membership:` Length of the membership of the customers with the e-commerce company
- `Time on App:` Time spent on the mobile app
- `Time on Website:` Time spent on web based browser
- `Yearly Amount Spent:` This is the dependent variable.

Stakeholders

If the company wants to decide whether to focus their efforts on the mobile app or the website.

Key Performance Indicators (KPIs)

All the quantitative features were considered to find their importance on the `Yearly Amount Spent` variable. However, it was found that `Length of Membership, Time on App, Avg. Session Length` have the highest impact on the dependent variable in decreasing order.

Modeling

$$\begin{aligned}\text{Yearly Amount Spent} = &-1054.215476 + 25.362665 \times (\text{Avg. Session Length}) \\ &+ 38.823679 \times (\text{Time on App}) + 0.803568 \times (\text{Time on Website}) \\ &+ 61.549053 \times (\text{Length of Membership})\end{aligned}$$

Results and Outcome

Model Explanation

Based on the model above, we can sumerize as follows

- If everything else remain unchanged, a 1 unit increase in `Avg. Session Length` is associated with an increase of 25.36 in total `Yearly Amount Spent`

- If everything else remain unchanged, a 1 unit increase in `Time on App` is associated with an increase of 38.82 in total `Yearly Amount Spent`

- If everything else remain unchanged, a 1 unit increase in `Time on Website` is associated with an increase of 0.80 in total `Yearly Amount Spent`

- If everything else remain unchanged, a 1 unit increase in `Length of Membership` is associated with an increase of 61.55 in total `Yearly Amount Spent`

Now the key question, *should the company focus more on `Time on App` more?*

The answer to the question above is a little bit tricky. Based on the modeling approach, appearantly it may seems that time on app has more impact than the time on web. However, the most significant factor seems the `Length of Memberhsip`. So we need further analysis of this two features to properly answer if the company should focus more on app.

Model Accuracy

The model above returns a `MAE` of 7.99, `MSE` of 102.72, `RMSE` of 10.14, and $R^2 = 98.46\%$