Indian Institute of Technology Patna

CS564 - Foundation of Machine Learning

Assignment #3 Linear Regression

Submission Date: 21th April 2024

Baskar Natarajan - 2403res19(IITP001799)

Jyotisman Kar – 2403res35(IITP001751)

SEMESTER-1

MTECH AI & DSC

INDIAN INSTITUTE OF TECHNOLOGY PATNA.

1. Problem Description:

Amazon_cloths sell clothes online. Customers come into the store, have meetings with a personal stylist, then they can go home and order either on a mobile app or website for the clothes they want. The company is trying to decide whether to focus its efforts on its mobile app experience or its website. Apply Linear Regression to find out the below-mentioned questions;

1. Find out the intercept
2. Find out the slope
3. Find out the coefficient by using features: Avg. Session Length, Time on App, Time on Website Length of Membership and (y= Yearly Amount Spent)
4. Find out in which feature the company should invest more. 5. Plot the test prediction

2. Linear Regression:

Linear regression is a statistical method used to model the relationship between one or more independent variables (features) and a dependent variable (target) by fitting a linear equation to observed data. The goal of linear regression is to find the best-fitting line (or hyperplane in the case of multiple features) that minimizes the difference between the actual and predicted values of the target variable.

2.1 Types of Linear Regression:

1. **Simple Linear Regression**: Involves a single independent variable.
2. **Multiple Linear Regression**: Involves multiple independent variables.

2.2 Advantages of Linear Regression:

- Simple and easy to understand.
- Provides interpretable coefficients that indicate the relationship between variables.
- Requires less computational resources compared to complex models.
- Works well with linearly related data.

2.3 Disadvantages of Linear Regression:

- Assumes a linear relationship between variables, which may not always hold true.
- Sensitive to outliers and multicollinearity.
- Limited in handling non-linear relationships between variables.
- May not perform well with high-dimensional data.

2.4 Where to Use Linear Regression:

- **Predictive Modeling**: Predicting numerical outcomes such as sales, prices, or scores.
- **Trend Analysis**: Analyzing trends and patterns in data over time.
- **Risk Assessment**: Assessing risk factors and predicting outcomes in various domains.
- **Causal Inference**: Identifying causal relationships between variables when experimental data is available.

2.5 Where Not to Use Linear Regression:

- **Non-linear Data**: When the relationship between variables is non-linear.
- **High-Dimensional Data**: When dealing with high-dimensional data and complex relationships.
- **Non-Normal Data**: When the assumptions of normality, linearity, and homoscedasticity are violated.

3. Assignment Description:

- We are performing linear regression analysis on a dataset containing information about customers of an e-commerce company.
- It aims to determine the relationship between various factors such as session length, time spent on the app and website, length of membership, and the yearly amount spent by the customers.
- Additionally, the code evaluates the model's performance and provides insights into which factors the company should focus on to increase customer spending.

Algorithmic Steps:

1. **Import Libraries**: Import necessary libraries including pandas, scikit-learn, matplotlib, seaborn, and numpy.
2. **Define Function to Compare App and Website Usage**: Define a function **compareAppandWebsite()** to visualize pairwise relationships and correlation between features using pairplot and heatmap.
3. **Load and Preprocess Data**:
   - Load the dataset containing relevant columns ('Avg. Session Length', 'Time on App', 'Time on Website', 'Length of Membership', 'Yearly Amount Spent').
   - Call the **compareAppandWebsite()** function to visualize the data.
4. **Split Data**:
   - Split the dataset into features (X) and target variable (y).
   - Split the data into training and testing sets using train_test_split().

5. **Build Linear Regression Model**:

- Create a LinearRegression model.
- Fit the model to the training data.

6. **Calculate and Print Model Statistics**:
   - Calculate and print the intercept and slope (coefficients) of the linear regression model.
   - Calculate and print the coefficients for each feature.
   - Determine which feature the company should invest more in based on the coefficients.
   - Compare the coefficients for 'Time on App' and 'Time on Website' to decide which feature the company should focus on.
7. **Visualize Test Predictions**:
   - Plot the true values vs. predicted values to visualize the performance of the model.
8. **Evaluate Model Performance**:
   - Calculate and print evaluation metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) score.

Meaning of Variables:

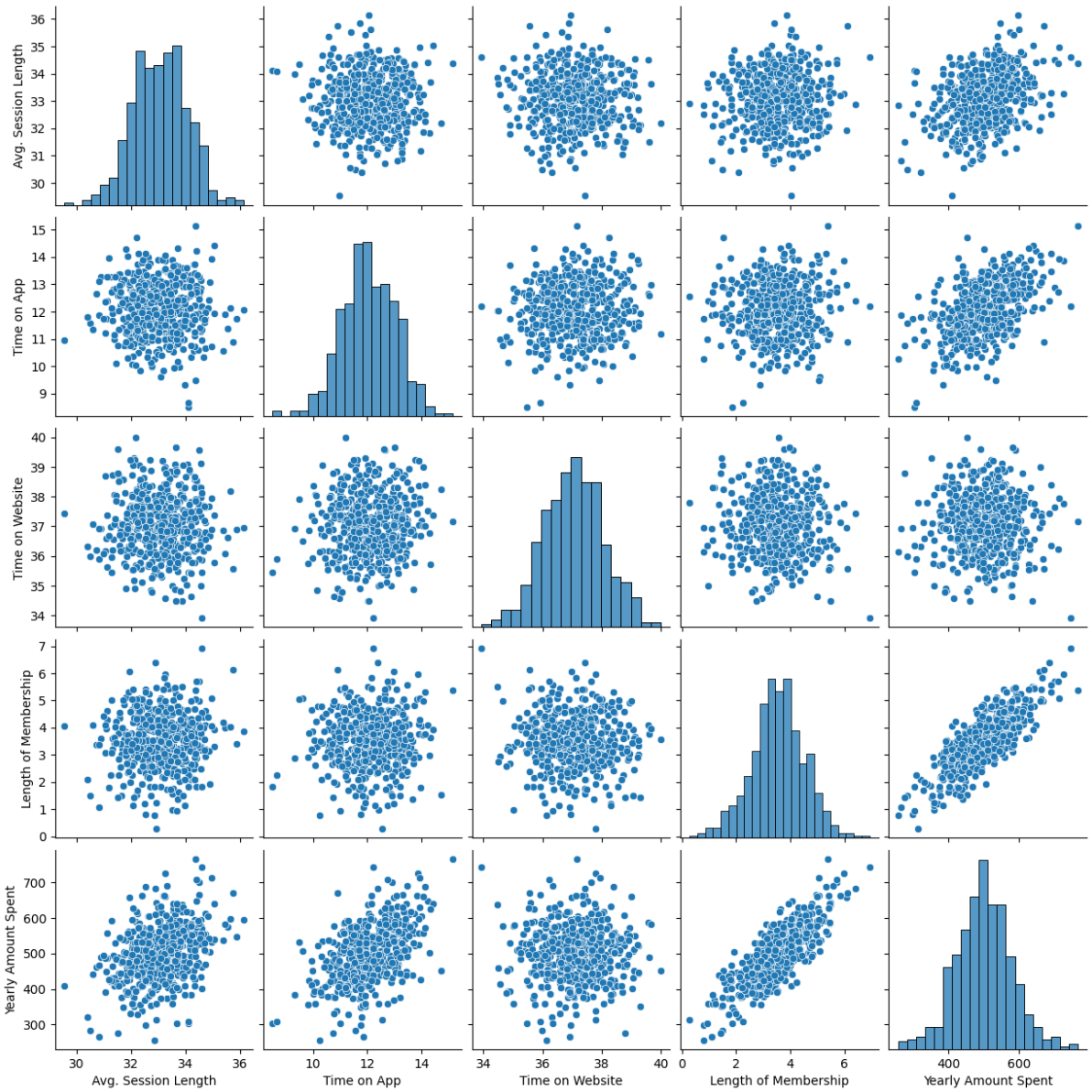
- **data**: DataFrame containing the dataset with customer information.
- **X**: DataFrame containing features (independent variables) used for prediction.
- **y**: Series containing the target variable (dependent variable).
- **X_train**, **X_test**: Training and testing sets of features.
- **y_train**, **y_test**: Training and testing sets of target variable.
- **model**: Linear regression model object.
- **intercept**: Intercept (bias) term of the linear regression model.
- **slope**: Coefficients (weights) of the features in the linear regression model.
- **coefficients**: DataFrame contains coefficients of each feature in the linear regression model.
- **mae**, **mse**, **rmse**, **r2**: Evaluation metrics for the model performance.

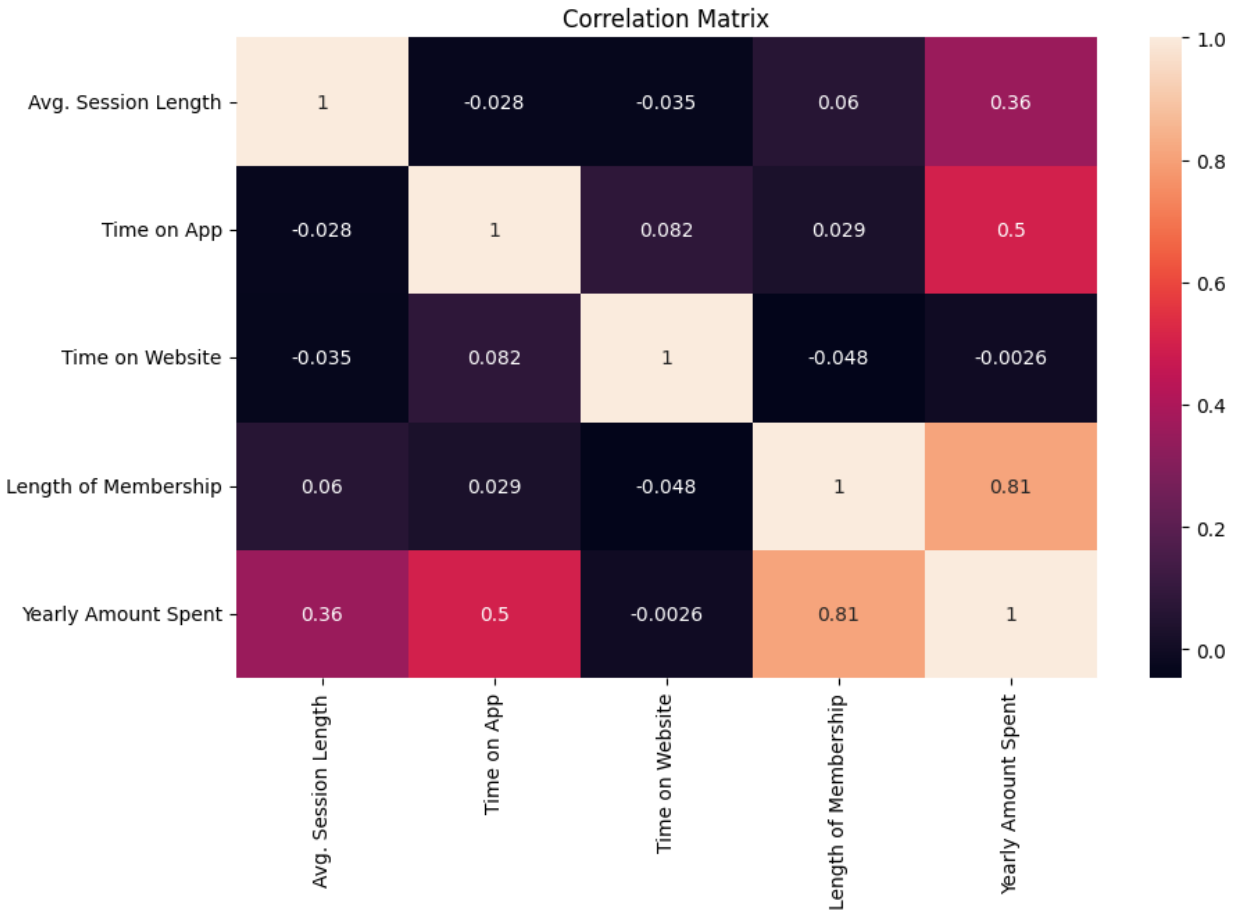
Output Brief:

- **Visualization**:
  - Pairplot: Visual representation of pairwise relationships between features.
  - Heatmap: Correlation matrix showing the correlation between features.
  - Scatter plot: Visualization of true values vs. predicted values to assess model performance.
- **Statistics**:
  - Intercept: Value of the intercept term in the linear regression equation.

    - Slope/Coefficients: Coefficients of each feature in the linear regression model.
    - Coefficients: Table displaying the coefficients of each feature.
- **Recommendations**:
    - Feature to Invest More: Determination of which feature the company should invest more in based on coefficients.
    - Focus of Efforts: Decision on whether to focus efforts on the app or website based on coefficients comparison.
- **Evaluation Metrics**:
    - Mean Absolute Error (MAE): Measure of the average absolute error between true and predicted values.
    - Mean Squared Error (MSE): Measure of the average squared difference between true and predicted values.
    - Root Mean Squared Error (RMSE): Square root of MSE, indicating the average magnitude of error.
    - R-squared (R^2) Score: Measure of the proportion of variance in the dependent variable explained by the independent variables.

4. Output:


Avg. Session Length
Time on App
Time on Website
Length of Membership
Yearly Amount Spent



Assignment Question Answers:

Intercept: -1044.2574146365562 Slope: [25.5962591 38.78534598 0.31038593 61.89682859]

****

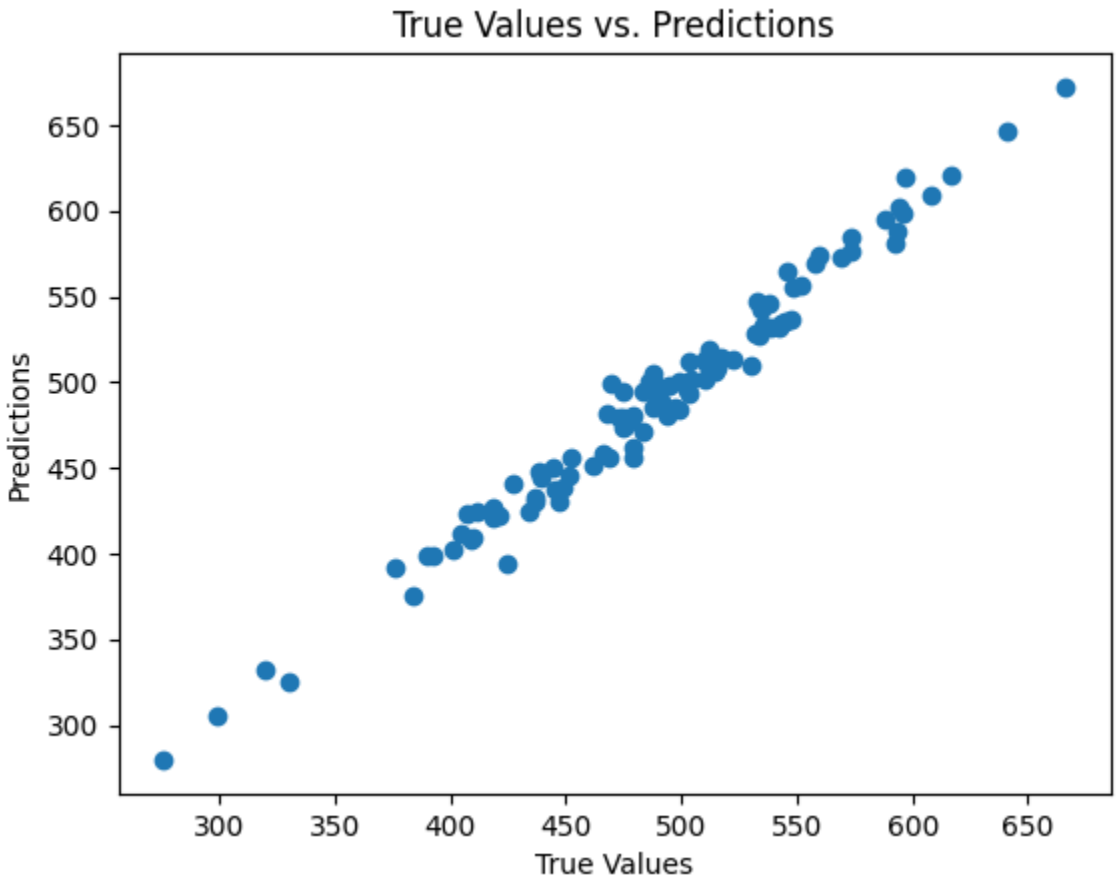
Coefficients: Coefficient Avg. Session Length 25.596259 Time on App 38.785346 Time on Website 0.310386 Length of Membership 61.896829

****

Feature the company should invest more is: Length of Membership The coefficient of (App) is 124.96 times higher than the coefficient of (Website). So the company should focus thier efforts on :App

****

Predictions Plot



Model Performance Metrics:

Mean Absolute Error (MAE): 8.558441885315217

Mean Squared Error (MSE): 109.86374118393982

Root Mean Squared Error (RMSE): 10.481590584636466

R^2 Score: 0.9778130629184127