Minor Project Report

Linear Regression Model for Customer Spending Prediction

Submitted by: Jatin Ghaghat
Institution/Shobhit University

Abstract

This project focuses on building a Linear Regression model to predict the yearly amount spent by customers based on behavioral and membership features. The study uses exploratory data analysis (EDA), data preprocessing, and model training with scikit-learn. Results show that Length of Membership and Time on App are strong predictors of customer spending. The model demonstrates reliable performance, offering insights for business strategy.

Introduction

The rapid growth of e-commerce platforms has created a need for understanding customer behavior and predicting spending patterns. Businesses aim to identify the factors influencing customer expenditure to improve marketing strategies and enhance user experience.

This project addresses the problem of predicting the yearly amount spent by customers using regression techniques. The objective is to analyze behavioral features (such as time spent on app, website, and membership duration) and establish relationships that help forecast customer expenditure. The scope of the project is limited to linear regression modeling but lays the foundation for more advanced predictive models.

Literature Review / Related Work

Previous studies have shown that regression models are effective in predicting customer behavior. Marketing research highlights the importance of engagement metrics (session length, time spent on applications, etc.) in driving purchases. Similar academic projects and research papers demonstrate the effectiveness of linear regression in identifying correlations and making continuous predictions.

Methodology

The methodology for this project includes the following steps:

1. **Data Collection** – Customer dataset with features: Avg. Session Length, Time on App, Time on Website, Length of Membership, and Yearly Amount Spent.
2. **Data Preprocessing** – Checking missing values, descriptive statistics, and data cleaning.
3. **Exploratory Data Analysis (EDA)** – Scatterplots, pairplots, and correlation analysis using Seaborn and Matplotlib.
4. **Model Building** – Train-test split (70%-30%), training a Linear Regression model using scikit-learn.
5. **Evaluation** – Predictions on test data and evaluation using MAE, MSE, and RMSE.
6. **Visualization** – Scatterplot of predicted vs. actual values for performance assessment.

Data Analysis / Results

Key findings from the dataset analysis:

- **Length of Membership** showed the strongest linear correlation with yearly spending.
- **Time on App** was also a strong predictor, while **Time on Website** had comparatively less impact.
- Pairplots revealed significant positive trends between predictors and the target variable.

Model coefficients:

- Avg. Session Length → Positive impact
- Time on App → Positive impact
- Time on Website → Minor impact
- Length of Membership → Strongest predictor

Evaluation metrics:

- Mean Absolute Error (MAE): Low (small average prediction error).
- Mean Squared Error (MSE): Acceptable, indicating reliability.
- Root Mean Squared Error (RMSE): Demonstrates good predictive power.

Discussion

The analysis confirms that customer membership duration significantly influences yearly expenditure. Customers engaged with the mobile application also tend to spend more. Website usage showed weaker correlation, suggesting businesses should invest more in app optimization.

Challenges faced included ensuring data consistency, avoiding overfitting, and visualizing multidimensional relationships. These were addressed through preprocessing, train-test splits, and effective plotting.

Conclusion and Future Scope

This project successfully demonstrates how linear regression can predict customer yearly spending based on behavioral data. The strongest predictor was Length of Membership, highlighting the importance of long-term engagement. Findings suggest businesses should focus on customer retention and mobile app experience to maximize revenue.

Future work may involve:

- Implementing advanced models like Random Forest or Gradient Boosting.
- Expanding dataset with demographic and geographic variables.
- Deploying the model as a web application for real-time predictions.

References

1. Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research.
2. Waskom, M. L. (2021). *Seaborn: Statistical Data Visualization*. Journal of Open Source Software.
3. McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. Proceedings of the 9th Python in Science Conference.
4. Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. Elsevier.