Machine Learning 1 - BIA-5302-0LC

Assignment 2

Group 3

Student Names	**Student N**
Riya Choudhary Kolli	N01485460
Nidhi Sharma	N01629511
Yugam Gaba	N01719779
Anubhav Chawla	N01663781
Vishal	N01676081

Index

1. Introduction

1.1. Project Objective

The project aims to analyze and model customer spending behavior using the Mall Customers Dataset. The goal is to gain insights into customer demographics, income, and spending patterns.

1.2. Chosen Dataset

The dataset used is the Mall Customers Dataset.

Link to Dataset:
https://www.kaggle.com/datasets/abdallahwagih/mall-customers-segmentation/data

2. Dataset Selection

2.1. Justification for Dataset Choice

The Mall Customers Dataset provides demographic information, annual income, and spending habits of mall customers, making it suitable for exploratory data analysis, customer segmentation, and clustering tasks.

3. Data Preprocessing

3.1. Handling Missing Values

The dataset had no missing values or duplicates.

3.2. Encoding Categorical Variables

The ‘Genre’ column, representing gender, was encoded as Male = 1 and Female = 0.

3.3. Data Scaling

Numerical features were standardized using StandardScaler to ensure all features contribute equally to the models.

3.4. Outlier Treatment

Boxplots were used to identify outliers in ‘Age, ’ ‘Annual Income (k$), ’ and ‘Spending Score (1-100). ’

4. Exploratory Data Analysis (EDA) and Visualization

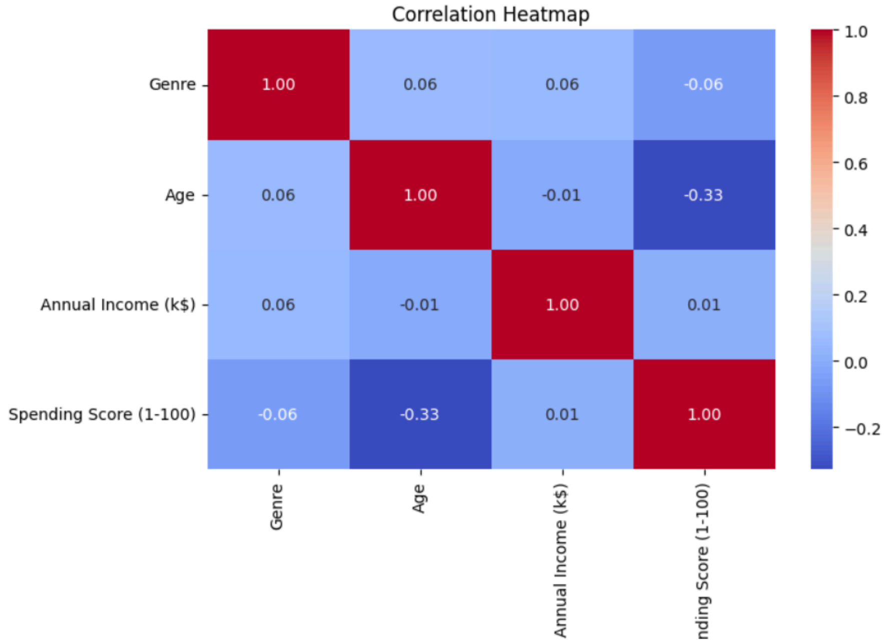
4.1. Distribution of Key Variables

- Histograms showed the distribution of ‘Genre’, ‘Age’, ‘Annual Income (k$)’, and ‘Spending Score (1-100)’.
- Age is skewed towards 20-40 years.
- Annual Income is fairly evenly spread.
- Spending Score is bimodal.

4.2. Correlation Analysis

A heatmap was used to visualize the correlation between features.

- There were weak correlations overall.
- A slight negative correlation exists between Age and Spending Score.
- Gender and Income showed minimal direct correlation.



4.3. Relationships Between Variables

- Scatterplots showed the relationship between ‘Age’, ‘Annual Income (k$)’, ‘Genre’, and ‘Spending Score (1-100)’.
- No strong linear relation between Age and Spending Score.
- Some patterns between Income and Spending Score (clusters).
- Gender doesn't show a strong impact on the Spending Score.
- Pairplots illustrated feature interactions, revealing clustering in Income vs. Spending Score and slight differences in male and female patterns.

5. Dimensionality Reduction

5.1. Techniques Used

Principal Component Analysis (PCA) was applied, although it was deemed unnecessary due to the small number of features.

5.2. Performance Comparison

PCA did not significantly improve model performance.

6. Modeling

6.1. Multiple Linear Regression

- **6.1.1. Model Training:** A Linear Regression model was trained.
- **6.1.2. Hyperparameter Tuning:** Grid search was used to find the best parameters. The best parameter for ‘fit_intercept’ was found to be ‘True. ’

6.2. K-Nearest Neighbors (KNN)

- **6.2.1. Model Training:** A KNN Regressor model was trained.
- **6.2.2. Hyperparameter Tuning:** Grid search was used to find the optimal number of neighbors. The best parameter for ’n_neighbors’ was 9.

6.3. Train-Test Split

The data was split into training (70%) and testing (30%) sets.

7. Model Evaluation

7.1. Multiple Linear Regression Performance

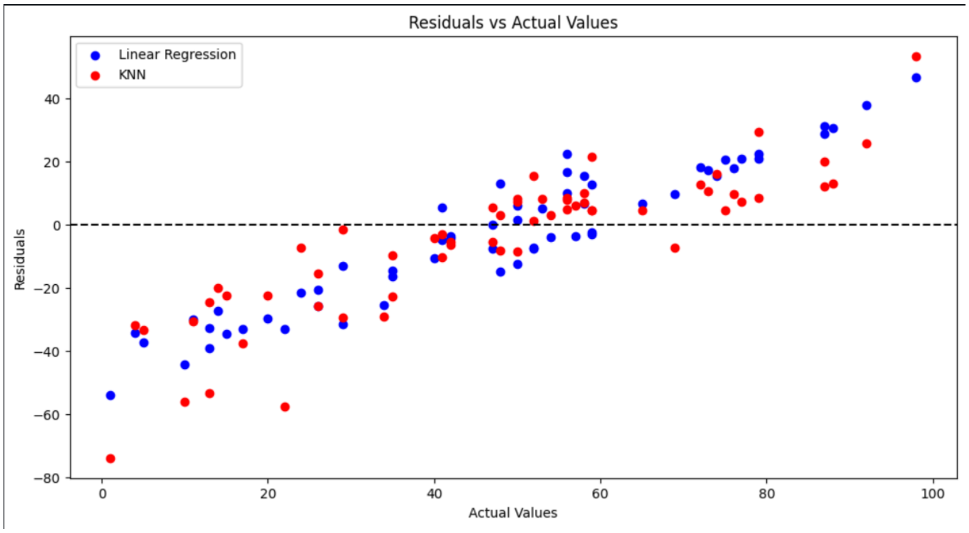
- MAE: 19.06
- MSE: 525.22
- RMSE: 22.92
- R^2 Score: 0.12

7.2. KNN Performance

- MAE: 16.94
- MSE: 534.92
- RMSE: 23.13
- R^2 Score: 0.10

7.3. Model Comparison

- Both models had similar results, with R^2 scores close to 0, indicating they didn’t explain the variance in the data effectively.
- Linear Regression had a slightly better R^2 score.
- KNN residuals showed more variance, suggesting potential overfitting.



8. Research Questions

8.1. Most Influential Variables

Annual Income and Age had the most significant impact on the Spending Score. Gender had less influence.

8.2. Comparison of Model Performance

Linear Regression performed slightly better than KNN, with lower error values and a higher R² score, but both models had R2 scores close to 0.

8.3. Impact of Data Preprocessing

Preprocessing and scaling numerical features were crucial for model performance. Scaling prevented income from dominating the results.

8.4. Insights from EDA

- Customers show two spending behaviors: low and high spenders.
- Most customers are aged between 20 and 40.
- There's no strong linear trend, but patterns are visible.
- Income and age are more important than gender in predicting spending.

9. Conclusion

9.1. Summary of Findings

The analysis revealed that Annual Income and Age are key factors influencing spending behavior. Both the Linear Regression and KNN models had limited success in predicting spending score, with R^2 scores close to 0.

9.2. Best Performing Model

Linear Regression performed marginally better than KNN.

9.3. Limitations and Future Work

The models' low R^2 scores suggest that other factors or more complex models may be needed to better predict customer spending. Future work could explore additional variables or different modeling techniques.