Commute Time Based on Living and Academic Factors

by
Audwit Nafi Anam

Analysis Objective

Draw Insights on impact on students' daily commute time for different factors

Dataset Summary

- **1107 Samples, 7 Columns**
  - **Age (numerical)**
  - **Gender (categorical: Male, Female, Other)**
  - **Major (categorical: Computer Science, Business, Engineering, etc.)**
  - **Distance from office (numerical, in km)**
  - **Mode of Transportation (categorical: Car, Bus, Bike, Walk)**
  - **Traffic Conditions (categorical: Low, Medium, High)**

Dataset Preprocessing

- Deleting Rows with Null Values (13) and Duplicates (50) - 1045 after removal
- One-hot Encoding Categorical Features
  - Gender, Major, Mode,
- Label Encoding Ordinal Features
  - Traffic
  - 19 Columns after Encoding
- Outlier Removal
  - Keeping Numerical Values within 1.5*IQR
  - Keeping Categorical Values having count greater than threshold
  - 1000 Samples after removal
- Challenges
  - Duplicate values with slightly different names
  - Trailing/Leading Whitespaces
- Solutions
  - Used lambda exprsssions, regular expressions, etc.

Exploratory Data Analysis (EDA)

- Average commute distance 8.4 km, median 7 km, majority within 0–15 km
- Mean commute time ~95 min, median 84 min
- Majority studied in CSE, then EEE, BBA and Arts in order
- Most common transportation is Bus, then walking, car, bike and rickshaw is least common
- Low, medium and high traffic equally faced – mostly medium
- Majority take 0–200 min commute time, mean ~95 min, median 84 min
- Commute time has positive correlation with traffic and distance (0.44)
- Greater traffic correspond to higher commute times as well
- Bus travellers take the most time to commute, followed by car, bike

Model Development and Evaluation

- Linear Regression
  - R_2 score: 52%, MSE: 1576.85, RMSE: 39.71, MAE: 27.73
- Lasso
  - Weak (strength: 0.01):
    - R_2 score: 51%, MSE: 1582.33, RMSE: 39.78, MAE: 27.83
  - Medium (0.5)
    - R_2 score: 49%, MSE: 1666.47, RMSE: 40.82, MAE: 29.17
  - High (1)
    - R_2 score: 47%, MSE: 1739.72, RMSE: 41.71, MAE: 30.52
- Ridge
  - Weak (strength: 0.01):
    - R_2 score: 52%
  - Medium (0.5)
    - R_2 score: 51%
  - High (5)
    - R_2 score: 50%
- Polynomial Regression is worse than all models above (best R_2 = 47%)

Key Insights and Recommendations

Insights

- All models perform similarly
- Higher Regularization leads to feature selection in Lasso and shrinking weights in Ridge
- Features are Uncorrelated, hence similar performance across linear models

Recommendations

- Gather more data
- Balanced data
- Modularize, Containerize for scalable deployment

Thank You