Documentation

Introduction

This documentation outlines the progress and key findings of a project aimed at predicting 12th standard math scores based on various features. The analysis involves exploring a dataset containing 6571 features and 16197 rows.

Week 1: Exploratory Data Analysis and Linear Regression

1.1 Data Overview

- The dataset comprises 6571 features and 16197 rows.
- Initial exploration involved converting numerical values to categorical for Power BI analysis.

1.2 Data Cleaning and Visualization

- Null values were dropped to observe their impact on the dataset size.
- Visualization of gender, race, ethnicity, and English proficiency distribution using Matplotlib and Seaborn.

1.3 Linear Regression

- Linear regression was performed using features such as student ID, stratum ID, PSU, sex, race, language, date of birth, mother's education level, father's education level, socioeconomic status, and 10th-grade math test scores.
- Initial Mean Squared Error (MSE): 418.
- Hyperparameter tuning reduced MSE to 382.90.

1.4 Feature Selection

- Attempted to improve model accuracy by dropping stratum ID and PSU, but no significant change observed (MSE: 382.07).

Week 2: Variable Selection and Model Comparison

2.1 Variable Selection

- Importance of selecting the right set of variables for matching emphasized.

2.2 Model Tuning

- Manual parameter tuning, filling missing values, and dropping null values.
- Employed Random Forest Regressor and Neural Network.
- Random Forest Regressor MSE: 41.10.
- Neural Network MSE: 40.

Week 3: Correlation Analysis and Advanced Modelling

3.1 Correlation Analysis

- Attempted correlation analysis through a heatmap, but due to system limitations, the process was aborted.

3.2 Advanced Modelling

- Explored linear regression, multiple regression, and neural networks.
- Achieved MSE of 20.16 and R2 score of 0.802.
- Incorporated 12th standard math scores as a training feature.
- Visualized the differences between actual and predicted 12th-grade math scores.

Week 4: Model Comparison and Additional Features

4.1 Model Comparison

- Explored KNN, SVM, and Random Forest Regressor without using 12^{th} standard math scores for training.
- KNN and multiple regression provided MSE of 48.
- SVM's MSE was 92.
- Random Forest Regressor provided the best performance with MSE of 19.95 and an R2 score of 0.794.

4.2 Additional Features

- Included a new feature, BYOCCHS (occupation right after high school).

4.3 Visualization

- Plotted actual vs. predicted 12th-grade math scores for the best model.
- Analysed the count of points in different categories based on the difference between actual and predicted scores.

Conclusion

The project successfully explored various models and techniques for predicting 12th-grade math scores. The Random Forest Regressor emerged as the most effective model, achieving an MSE of 19.95. Further exploration and fine-tuning can be performed to enhance the model's accuracy and robustness.