Regression Based Prediction Model

Project Report

Presented by :

Abdul Majid Lutfi

Presented for :

Sir Rasik Ali

Roll Number :

su92-bsdsm-f23-009

Project Report Artificial Intelligence

Abstract

This project focuses on developing a predictive model using a Gradient Boosting Regressor for regression-based tasks. The dataset from raw text reviews and associated numerical scores, was prepared by cleaning textual data, encoding categorical features, and scaling numerical attributes. Exploratory data analysis revealed critical patterns and relationships between variables.

Introduction

Gradient Boosting Regressor was chosen for its ability to model complex non-linear relationships and its robustness. Key hyperparameters such as the number of estimators and tree depth were tuned using (Grid Search CV) to optimize performance. Feature importance analysis highlighted significant predictors, contributing to interpretability.

Dataset Description

The dataset rotten_tomatoes_movie_reviews.csv taken from kagal the data set take 3 column names included:

• **review Id**: This is the id of members they give the review

• **review Text**: Review text column take review given by the viewers

• **score Sentiment**: This is the target column and the take score I will replace the value of (true & false) with (0,1)

Technologies Used

The project utilized various technologies and libraries for data manipulation, visualization,

machine learning. The following libraries were employed:

Data Manipulation and Visualization:

• **NumPy**: For numerical operations and array manipulations.

• **pandas**: For data manipulation and analysis.

• **matplotlib**: For creating static, interactive, and animated visualizations.

• **Beautiful Soup:** To remove Punctuation and numbers

Data Preprocessing

1. **Loading the Data**: The dataset was loaded into a panda Data Frame for analysis.

2. **Handling Missing Values**: Identified columns with missing values and drop the null values.

3. **Dropping Irrelevant Columns**: The 'publication Name' 'critic Name' 'Is Top Critic' 'Original Score' 'Review State' column was dropped as it does not contribute to the prediction.

4. **Encoding Categorical Variables**: Categorical features were encoded using Label

Encoding to convert them into numerical format

Model Training and Evaluation

- **Training:** Fit the Gradient Boosting Regressor on the training data using optimal parameters.

Evaluation Metrics:

- Mean Squared Error (MSE)
- Mean Average
- R^2 Score

Conclusion:

The project demonstrated the successful application of Gradient Boosting for regression tasks. Preprocessing text data using TF-IDF and optimizing hyperparameters significantly improved model performance.