Student Project Report

Student Performance Analysis ? Movie IMDB Review (Regression)

Student Name: Kathir Kavin Kumar M

Roll No: 23AD027

Department: Artificial Intelligence & Data Science

Course Code: U21ADP05

Course Title: Exploratory Data Analysis and Visualization

Course In-Charge: Mr. Rushikesh Kadam

Date of Submission: 20 October 2025

Abstract

This project explores and predicts movie review ratings from IMDB textual reviews using exploratory data analysis and a regression model. The pipeline includes data loading, cleaning, text preprocessing (tokenization, stopword removal, TF-IDF), feature engineering, and model training using an MLP regressor. The deliverables include visualizations, evaluation metrics (RMSE, MAE, R2),
and Python source code to reproduce the results.

1. Introduction & Objective

This project aims to build a model that predicts numeric review ratings (for example 1?10) from IMDB review text. Objectives: perform EDA, preprocess text data, visualize distributions and relationships, train a regression model, and evaluate its performance.

2. Dataset Description

Source: Kaggle / Public IMDB review datasets (expected CSV with columns: 'review', 'rating').
Type: Text dataset (movie reviews) with numeric ratings (regression target).
Approx. Size: User-provided dataset (recommended >= 10,000 rows). If dataset is not provided the code will run a small demo sample.

3. EDA and Preprocessing

- Check missing values and duplicates.
- Basic text cleaning (lowercase, remove HTML tags, punctuation).
- Tokenization and stopword removal (optional lemmatization).
- Convert text to TF-IDF features; optionally try word embeddings.
- Split into train/validation/test sets (70/15/15).

4. Data Visualization (examples to produce)

1. Histogram of Ratings ? shows rating distribution.
2. Review Length Distribution ? words per review.
3. Top N frequent words (bar plot) ? most common tokens.
4. Scatter plot: Review length vs Rating ? correlation check.
5. Error distribution after model prediction (residuals).

5. Model Building

Model: Use a scikit-learn pipeline with TfidfVectorizer + StandardScaler (if needed) +

MLPRegressor.
Hyperparameters: hidden_layer_sizes=(128,64), activation='relu', alpha=1e-4, learning_rate_init=0.001, early stopping.
Training: use early stopping and monitor validation loss.

6. Result Visualization & Interpretation

- Plot training & validation loss curves (if using a Keras model) or learning curves for scikit-learn.
- Evaluate: RMSE, MAE, R2 on test set.
- Plot Predicted vs Actual ratings and residual histogram.

7. Conclusion and Future Scope

Predicting numeric ratings from raw text is challenging but useful for recommendation and analytics.
Future work:

- Use pretrained language models (BERT/transformers) for better features.
- Increase dataset size and incorporate metadata (genre, user info).
- Try ensemble models and interpretability methods (SHAP).

8. References

1. Kaggle ? IMDB datasets and related resources.
2. Scikit-Learn documentation: https://scikit-learn.org/
3. TensorFlow and Keras documentation: https://www.tensorflow.org/
4. Jurafsky & Martin ? Speech and Language Processing (for NLP foundations).

Note: Run the provided Python script 'movie_review_regression_project.py' to reproduce EDA, visualizations, and model results.