Movie Review Score Prediction

Overview

This project implements machine learning models to predict movie review scores based on financial and popularity metrics from The Movie Database (TMDB). Using features such as budget, revenue, popularity, and return on investment (ROI), we classify review scores into three categories: Low, Medium, and High.

Project Highlights

- **Data Source**: Full TMDB Movies Dataset
- **Models Implemented**: Random Forest Regressor, Random Forest Classifier, XGBoost Classifier
- **Primary Objective**: Predict audience reception through classification of review scores

Getting Started

Requirements

- Python 3.9+
- Dependencies listed in `requirements.txt`

Installation and Setup

1. **Clone Repository**

```
git clone https://git.txstate.edu/zzi1/Machine-Learning-Tabular-Data-Project.git
cd Machine-Learning-Tabular-Data-Project
```

2. **Create and Activate Virtual Environment**

```
# Create virtual environment
python -m venv .venv

# Activate virtual environment
# On Windows
.venv\Scripts\activate
# On macOS/Linux
source .venv/bin/activate
```

3. **Install Dependencies**

```
pip install -r requirements.txt
```

4. **Download Dataset**
    - Download the dataset from Hugging Face
    - Place the downloaded file in the directory of the repository and make sure its named `Dataset.csv`

Running the Project in VS Code

1. **Open Project**

```
code .
```

2. **Configure Jupyter**
    - Install Jupyter extension in VS Code
    - Select Python interpreter from your .venv (bottom status bar)
    - Run "Python: Create Jupyter Kernel" from command palette (Ctrl+Shift+P)
3. **Run the Analysis**
    - Open `ProjectFinalSub.ipynb`
    - Select your environment's kernel from the top-right dropdown
    - Run cells using the play button or Shift+Enter

Project Details

Key Features

The analysis pipeline includes:

- Data preprocessing and feature engineering
- Exploratory data analysis with visualizations
- Model training and hyperparameter tuning
- Performance evaluation and comparison

Results

The project generates comprehensive outputs including:

- Model performance metrics (Accuracy, F1 score)
- Confusion matrices for classification results
- Visualizations of feature importance
- Data distribution and correlation analysis

Dataset

This project uses the Full TMDB Movies Dataset from Hugging Face, which contains comprehensive information about movies including financial metrics, popularity scores, and review ratings.

Contributors

- Maitland Huffman
- James Allen