Strata.ai

Full Stack AI Accelerated Program

Module 2: Applied Machine Learning

Assignment 1

This assignment is to learn the "End-to-end Machine Learning Pipeline". The goal is to select and train a regression model to predict the "average rating/vote" for each movie based on the other relevant features. The data is taken from "The Movie Database (TMDb)" (attached) and includes the info for around 5000 movies. You need to do the following:

1- Follow all the steps that we did for the housing data during the class (e.g., data load, handling missing data, pre-processing, feature selection, model selections, cross validation, and hyperparameter selection)

2- Perform exploratory data analysis to understand data

3- Handle missing data (if any) and explain the reason behind your strategy

4- Select the relevant features. Explain the reasons behind any feature elimination

5- Encode categorical values. Note that there are categorical features with more than one value (e.g., genres). How do you represent them?

6- Create a Scikit-learn pipeline to do all pre-processing steps

7- Split data randomly. 80% training data and 20% test data

8- Use 5-fold cross validation for model selection. Compare KNN Regression and Linear regression. Use grid search to fine-tune the models (select hyperparameters)

9- Report and analyze the error (RMSE) on test data for the final selected model.