TASK: Predicting Revenue of Movies

Requirements:

1. Pandas
2. Numpy
3. Seaborn
4. Matplotlib
5. Scipy
6. Scikit-learn

The above-mentioned Python libraries need to be installed to run and execute the Jupyter Notebook.

Run:

Jupyter Notebook is used to analyze and predict the data. Jupyter Notebook can be run inside the Terminal which opens into the browser or by using an IDE like DataSpell or VS Code

Approach:

1. The given dataset is in the form of CSV.
2. Pandas has been used to import the dataset and create a Dataframe.
3. Exploratory Data Analysis of the data was done to get various insights about the data such as total number of features and samples in the dataset, NULL values etc.
4. The training dataset has 2400 samples and 21 features
5. The feature "belongs_to_collection" has roughly 1908 missing values, which is approx 80% data, hence it is unusable.
6. Initial Intuition involves the budget of the Movie is correlated with the revenue.
7. The Correlation Matrix reveals that the Budget and Popularity of a movie have high correlation with the revenue of the movie
8. The features are also scaled according using Min-Max scaler and Categorical features are encoded using Label Encoding
9. Since most of the variables have a skewed distribution, to normalize the data, Square Root transformation has been used.
10. The training and test dataset is preprocessed with all the above steps and train data is split in 80-20 ratio as to check model accuracy.
11. ExtraTreesRegressor is used for regression analysis, which is finally used as a model as it has outputs with very low RMSLE and RMSE.
12. The model is fully trained with the training dataset and is used to predict the Movie Revenue.
13. The predicted values are then exported as CSV.

Name: Abhinav Anthiyur Aravindan **Roll No: ESD19I013**