Project Summary

Design : This project was designed to develop a linear regression model that predicts revenue of movies using Budget, running_time, genre,rating, genre.

Data: Data used for this project was scrapped from Boxoffice mojo.com and IMDB.com. The data is taken for movies released between 2000 to 2019.

Algorithm:

1. Data scraping

From Boxoffice mojo - movie-title, domestic_gross,Total_gross, budget,runing_time_minutes, genres
From IMBD.com - Director, producer and cinematographer

2. EDA

2.1 Data cleaning - got rid of nan-values
2.2 Analyze the relationship between variables and the target using correlation map
- Features selection and tuning - After checking on the frequency of occurrences of the values in the director, cinematographer and producers, they don't have significant predictive value for the target therefore, they were dropped.
-Adjust the categorical variables distributor, ratings and genre by putting the unpopular or low count values into one category i.e others for each feature.
-Converted the categorical variables Distributor, Rating and genre into dummy variables.

3. Made a baseline model - linear regression model, fit and evaluate
4. Adjust the data further by getting rid of outliers
    - Movies that made more than 1,000,000,000 and less than 10,000,000 were dropped
    - The target variable was adjusted using log transformation to get rid off the high positive skewness it had
    - More variables were used to develop the model further and to increase performance
    - The model was overfitting, regularization was applied to improve it
5. Regularization using lasso and ridge models- the lasso regression model gave the best r2 and smallest MAE than the linear and ridge model

Tools:

For webscarpping - Beautiful soup
For analysis and model making - python