MOVIE INVESTMENTS

REEM ABANMI - NORAH TAMSAN

ABSTRACT

the goal of this project was to use regression model to predict the worldwide gross of movies to help talent agency determine whether the film will succeed or not and achieve the best profits, which contributes to the success of their clients. We worked on the data scraped from box-office mojo leveraging basic information of movies. To achieve promising results for this project

DESIGN

Using the data scraped from box-office mojo we must predict worldwide gross of movies to help talent agency determine whether the film will succeed or not and achieve the best profit. predicting the worldwide gross using machine learning regression model would enables the agency to make better decisions towards its customers, which contributes to increasing profits and the company's growth

DATA

Data is scraped from box-office mojo using Beautifulsoup from 2011 to 2021 with more than 1000 movie record after cleaning the data. The data basic features are Worldwide Gross, Distributor, Budget, MPAA, Running Time, Genres, Number of Theaters. After creating dummy variable, we ended up with 59 features and uses 20 of them.

ALGORITHM

We started by Scraping the data from box office mojo from 2011 to 2021 second we start data combining and cleaning by dropping some of the unnecessary columns and duplicates and NA values. third is the EDA step where we transform dependent variable and add dummy variables for each unique Distributor, MPAA, Genre, Month, year.

Then we start working on the fourth step the linear regression, we start with Ordinary Least Squares to get a look at the Linear Regression model and calculated the Mean Squared Error: 15461534.13, the Mean Absolute Error: 2993.81 finally we find the average R^2 training set : 0.7416 and the average R^2 validation set : 0.7271, based on cross validation. And the Ridge Regression with R^2 training set: 0.7348 and R^2 test set: 0.7158 and R^2 training set : 0.7405 and R^2 test set : 0.7072 using lasso .

Finally, we reported based on test data set R^2 training set: 0.7409 and R^2 test set: 0.7069, the Mean Squared Error: 15547828.39 and Mean Absolute Error: 3009.98

TOOLS

- seaborn
- Pandas library
- NumPy
- SciPy
- Matplotlib
- Scikit-learn
- Beautifulsoup
- pickle