Halah Salem Banuqaytah

T5 Boot Camp Project

Recommendations for movies production companies and predicting the profits of the movies

❖ Problem Background:

The movie industry that includes showing the movies only in cinema is facing a tough time, due to the massive impact caused by the new way of delivering movies to the public by streaming services. Streaming platforms such as Netflix and apple tv are starting to dominate the field with movies from their production and streamed on their own platforms which provide the audience with many features that cinematic production companies cannot afford. That put the big cinematic production companies in a difficult situation for losing their audience to the new direction of movies streaming, which will eventually lead to losing more profits. In addition, the impact that caused by the pandemic which led to global lockdown, almost half year without movies in cinema.

❖ Abstract:

In this project, I aim to predict the profits of the movies before they are released in cinemas to help the production companies know if the movies will be a success or not. I used a dataset from Kaggle under the name "movie industry", which was from the IMDb website. I used two regression models which are: 1-Linear Regression 2-Random Forst Regressor, to predict the profits of the movies and compare the models' accuracy and choose the best model.

❖ Design:

This is my first ever data science project, so I choose a topic the I personally interested in which is movies industry. The dataset includes movies from 1980 to 2020 that only streamed in cinema. Using the features in the dataset, I will predict the profits of a movie before releasing it in the cinema and check the following personal assumptions which will help the production companies to make more profits:

1. Most of the movies under the fantasy genre.
2. The action genre is the most profitable.
3. Main stars and directors with more movies have the highest profits.
4. Marvel production company is the most profitable.
5. Movies number and profits are increasing with time.
6. Movies in December has the highest profits compared to other months.

❖ Data:

The dataset contains 7668 movies with 15 features for every movie. The data types for the features include 9 objects, 5 float and one int. The highly important features for the analysis are profits, budget, star, director, genre and production company and releasing date. Some features could be dropped because they would be irrelevant to the analysis, and one feature could be divided into two features. For the Machine learning algorithm, I will need to convert the object data to a numerical type to be used in the model.

❖ Algorithms:

1- Feature selection:

To choose the independent variables that have a high correlation with the target variable (profits), I convert the data type for the object feature into categorical then to number using Panda's cat.Code. Then plot the heat map for the features and chose the highly correlated feature with our target, which was the (budget).

2- Models

In this project, I used two regression models to compare between them and then chose the most accurate one. They are: Linear Regression and Random Forst Regression. For accuracy, I will compare the R square value for each model for the tarin set (80%) and test set (20%).

Something to note here is that I fit the two models in two different ways:

1- The two models were fitted to the dataset after dropping all the null values, which included more than 2,000 rows from the budget feature.

2- The two models were fitted to the dataset where the null values in the budget feature were filled in by the mean budget for every genre.

R squared value for Linear Regression Table.

	Dropping budget null values	Fillin budget null values
Training set	**0.54**	**0.51**
Test set	**0.56**	**0.51**

R squared value for Random Forest Regression Table.

	Dropping budget null values	Fillin budget null values
Training set	**0.63**	**0.67**
Test set	**0.60**	**0.55**

Since the Random Forest has higher accuracy, it will be the model to be used for farther analysis in the future.

❖ Tools:

1- NumPy and Pandas to deal with the data
2- Seaborn and Matplotlib for data visualizations
3- Scikit-learning for models

❖ Communication:

This heat map shows the correlation between the features of the dataset.