BOX OFFICE PREDICTOR

Contributors:

1. Mesbaul Alam Khan - 3035548740
2. Ritvik Singh - 3035553044

Objectives:

The movie entertainment business is growing rapidly all over the globe. Every year, we see multiple blockbusters, be it in Hollywood, Bollywood or in other respective film industries. With a plethora of movies releasing every year, consumers are constantly faced with a critical decision: Should we invest our time and effort in order to purchase a ticket for a particular movie? Is the movie experience gonna be worth it?

Cinema halls also face a complicated dilemma on whether to screen a particular movie or not. This is because the rate at which films are being released is not slowing down. When a cinema theatre decides to screen a movie, they are spending a lot of resources, hoping that the respective film garners enough attention and praise, so that they can earn more revenue from higher ticket sales. This iis the very reason why cinema halls need to make the ultimate decision on which films to screen. Which film will bring in more customers, and which films will generate a higher revenue?

Both customers and cinema halls could utilise a system, where they could see an estimated box office prediction for an upcoming release.

- Customers may be swayed by the box office numbers, and it may factor into their decision on whether to watch a designated film on the big screen. If a box office prediction is considerably low, the customer may feel it does not belong to the general audience, and hence, they may decide not to waste their time and money on the designated release.

- Cinema halls may use the box office predictions to ultimately decide on whether to screen the film or not. A higher box office prediction will indicate a higher revenue generated for the cinema hall by screening the respective film. They will make higher ticket sales, more revenue on snacks, a better reputation among movie buffs for screening popular films etc.

Dataset:

We have decided to use the TMDB Movie Dataset for our Box Office Predictor. It is the default dataset that has been used by a wide array of Box Office Prediction competitions in Kaggle. It consists of 7398 movies, with a variety of data.

Link to dataset:
https://www.kaggle.com/c/tmdb-box-office-prediction/data

Data Engineering:

The dataset contains a wide array of information, in various primitive and non-primitive formats. We have strings, which provide a summary of the movie. Integers which provide the budget and revenue of the film. Floating point numbers providing the runtime and popularity of the film. We have several columns, which have information stored in a list of dictionaries.

As a result, we had to perform a wide array of Data Exploration and Data Engineering, in order to understand our dataset and how it affects the target column, **Revenue**.

First of all, we had to understand the different types of data present, and how the values were structured.

For **belongs_to_collection**, the data was stored in a list of dictionaries, which mainly referenced that the film belonged to a franchise, along with the franchise name. Hence, we extracted another column storing the franchise name, and used a **LabelEncoder** to label the different franchises. We also created a separate column to mark the absence or presence of a franchise using 0 and 1 respectively.

For columns like **budget**, it is difficult to fill up the null values using mean of other budgets, since this parameter has a wide range of values. Hence, we decided to extract the ground truth values of budgets by collecting data from external sources. All the respective sources have been cited in our notebook. This reduced the number of null values to a significant extent.

For columns like **genres, production_company, production_countries, keywords,** we perform similar operations as all the data in these columns are stored in a list of dictionaries. We extract the relevant **genre, production_company etc** and append them to a list. Using a **Counter**, we count the most frequently occurring item in the list, and create separate columns to indicate their absence or presence in our films.

For strings like **overview** and **tagline**, we computed the number of words in each of them, as through **Data Exploration**, we deduced that length of a **title** or **overview** may affect the revenue slightly.

Experimentation within the Model:

- We initially trained the model with default parameters and achieved a Root Mean Square Error of around 2.9, and the relative “importance” was distributed among only 5 features of the model, the rest being 0.

- In the second stage we decided to change the number of estimators from 500 to 5000. This not resulted in a lower root mean square error, but also assigned “importances” to more number of features within the model.In addition, only a handful of features were left with 0 importance.

- Next, we changed the maximum depth of the model from 8 to 11, resulting in the lowest RMSE score achieved of all i.e. around 2.55. Thus, by suitably changing the model parameters, we were able to increase the accuracy and reduce the error.

Improvements:

We would like to add a note that one of our group mates, Manya Agarwal, opted to exercise the late-drop option, leaving me and my groupmate in a rush to formulate a new project idea, and implementing it in less than a week.

We plan to make a much more polished model around the final deadline, and improve on a series of factors, to make our model accurate and predict box office numbers flawlessly.

- **Budget:** This is one of the most complicated issues that we had to face. The budget of a film is an integral part of the film's identity, and hence, having so many null values may put a detrimental effect on our model. We plan to perform a more thorough data engineering, to eliminate much more null budget values, and also consider adding external datasets.

- **Modelling**: We have used a simple **Random Forest Classifier** for our modelling and training. However, in the coming weeks, we plan to use other modelling techniques in order to see improved results. As we referenced a series of examples in **Kaggle**, we have noticed quite a few models being trained using a **LGBM Regressor** and **XGB Regressor**. We plan to test out different models using different model parameters, and adopt the model which will provide us with the most prominent results.

- **Feature Engineering:** We plan to understand our model more precisely, and create new columns based on existing information, that may improve our model further. We will study which parameters are affecting our model most prominently, and for the ones that don’t, we will try to create more columns out of existing information that may improve the model.