DATA SCIENCE FOR BUSINESS (DATA ANALYTICS)

PREDICTION OF MOVIE SUCCESS USING RANDOM FOREST A MACHINE LEARNING ALGORITHMS AND THEIR COMPARISON

I. INTRODUCTION

A. Aim and objectives

The research's main objective is to determine a major factor for the movie's success. The research work used mathematical tools such as R, Python, Tableau public, SAS, Apache Spark, Excel, and Rapidminer (a financial data mining tool). The movie data will be analyzed by two statistical frameworks and will show through several analytical techniques. The results will be shared with the current journal population. It will be helpful to apply machine learning algorithms like Random Forest to identify interesting insights in the crowdsourcing community.

B. Selection of dataset

The dataset is obtained from the website Kaggle (https://www.kaggle.com/rounakbanik/the-movies-dataset?select=movies_metadata.csv). The dataset includes different variables and features related to the movie.

C. Explanation of Dataset

Our data has more than 5000 records and 28 attributes or variables. The variables that distinguish each record are the columns that build each record. Some other relevant variables are the release date and budget. Each movie that appears on IMDB has a score associated with it.

This research study finds that the relationship between the IMDb score and the movies' success, actor popularity and reviews, and movie success.

D. Business case of the project

The work will also build the model for movie success. Machine algorithms will be used to train machine models. Rating websites like IMDB give ratings for movies based on different factors. People decide to watch the movie by the scores given by the website. Movie ratings can determine the success of a movie. The research uses the IMDB movie dataset to find unknown hidden facts from the dataset. For media, it will be useful to make intelligent decisions [1].

E. Selection of environment

Two statistical tools, R and Python, will be used in this research. Both are open-source software for data analysis. The two tools add new libraries with every release. Python is the tool primarily used for data analysis, and R is a general tool. Python and R are both powerful programming languages used for data science.

F. Selection of the technology

Random forest algorithm is one of the most common machine learning algorithms capable of performing both regression and classification tasks. In the forest of many decision trees, the more trees, the more robust and therefore high precision in the model to model multiple decision trees to build the forefront. The reason Random Forest is more likely to find better insights is that movies cost a lot of money no matter what. This analysis is essential to predicting a more accurate success rate.