Ha Noi University of Science and Technology
School of Community and Information Technology
------000------

Introduction to Business Analytics

Topic:

Predict the success of an upcoming movie

Supervisor: Nguyen Binh Minh
Members of group:
Le Duc Anh - 20194416 - anh.ld194416@sis.hust.edu.vn
Kieu Trong Thanh - 20194453 - thanh.kt194453@sis.hust.edu.vn
Nguyen Thi Trang - 20194458 - trang.nt194458@sis.hust.edu.vn
Le Thanh Thang - 20194451 - thang.lt194451@sis.hust.edu.vn

January, 2023

1. Introduction

The entertainment industry has been growing in every scope. Be it Netflix, Amazon, or Hotstar, there is a lot of content out here. Now, the challenge these streaming face is what to buy, in the sense that which content will get them more viewers and also satisfy the existing customer base.

For this project, you need to predict the success of an upcoming movie so that whether or not a company should go for buying it based on ROI. To do this, you need to come up with a model and use the historical data of each element involved, such as the actors, the director of the movie, the production company, the genre of the movie, etc.

The main idea of doing this business analytics project is to predict the market for upcoming media content based on some preset parameters as this is one of the most unpredictable industries: big stars might not always shine, while the newcomers might actually do a great job! You will need to keep all of that in mind.

2. Data

2.1. Data collection

The first step was to collect data. There was no pre-existing dataset with the movies and features we wanted, so we collected data from multiple sources and merged it. Our main sources of data are IMDB, Rotten Tomatoes.

From IMDB, we obtained for each movie: movie title, IMDB rating, plot description, budget, box office gross, opening weekend gross, runtime, genres and release date. Also from IMDB, we obtained the number of Academy Awards that actors and directors in each movie had won prior to that movie, and also the number of Best Picture films that actors and directors in each movie had been involved in, also prior to that movie. In order to stay true to our goal of only considering factors known before a movie's release, we considered only awards that had been received prior.

From Rotten Tomatoes, we obtained critic score, audience score, MPAA rating, actors and director.

Using a data processing tool (Python) to parse, and clean up what we retrieved, we had our data set of movies across IMDB and Rotten Tomatoes.

2.2. Data cleaning

Many features are missing from many movies and it must be handled. All these features are necessary for analysis, so the movie is dropped if any of these features are missing. We have three monetary attributes: budget, global gross profit, and opening weekend returns. In order to increase the utility of these fields, we accounted for inflation. For instance: budget = budget * (USD in 2023 / 1 USD in each year of the movie.

After data collection and cleaning, we have the dataset which contains 16 variables for 5584 movies.

Variable name	Description
movie	the name of movie
year	the year in which the movie is released
runtime	the runtime of the movie
imdb_rating	IMDB rating
release_date	Release date
plot_summary	Plot summary of the movie
genres	some main genres of the movie
budget	budget of the movie in USD
gross_earning	global gross earning of the movie (USD)
opening_weekend_gross	opening weekend gross of the movie (USD)
audience_score	audience score in rotten tomatoes
tomatometer_score	tomatometer score of the movie
mpaa	MPAA rating
rotten_info	name of director
cast	name of the actor
award_involve	the number of academy awards that the actor and the director achieved and the number of best picture films they are involved prior to that movie

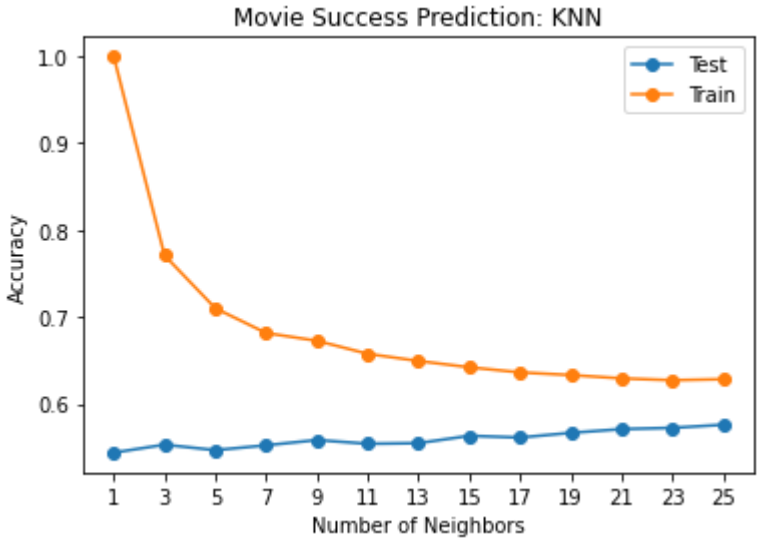
3. Machine learning techniques

In the dataset, we have 4 output variables: gross_earning, imdb_rating, audience_score, tomatometer_score. Instead of values, with the help of other parameters I want to predict whether a movie is a success or failure. Box office gross of a movie is a success when gross earning of the movie is two times bigger than the budget of the movie. IMDB rating of the movie is success when IMDB rating score larger than 7. Audience score of the movie is success when audience score larger than 60 and critic score of the movie is success when tomatometer score larger than 75.

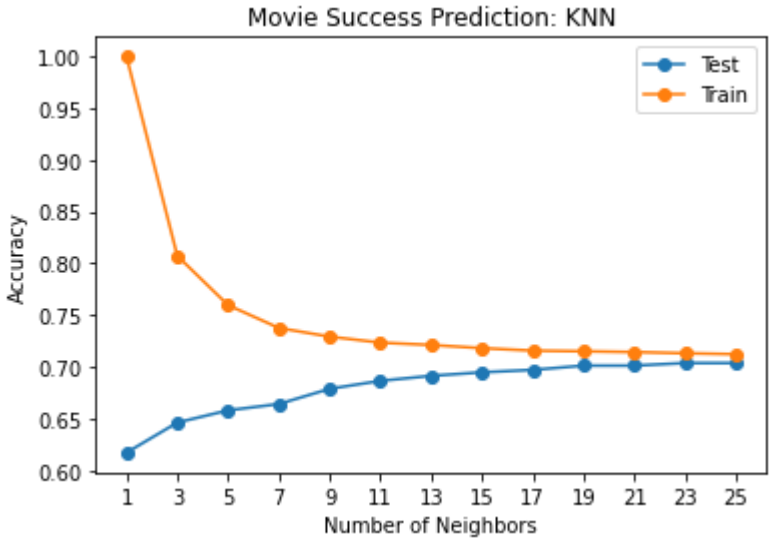
In this project, we use regression machine learning models such as Logistic Regression, KNN, Decision Tree, Random Forest to predict the success of a movie and the result is presented below.

a. Logistic Regression
   For gross: accuracy: 0.5768
   For imdb_rating: accuracy: 0.7294
   for critic_score: accuracy: 0.6705
   For audience_score: accuracy: 0.6153

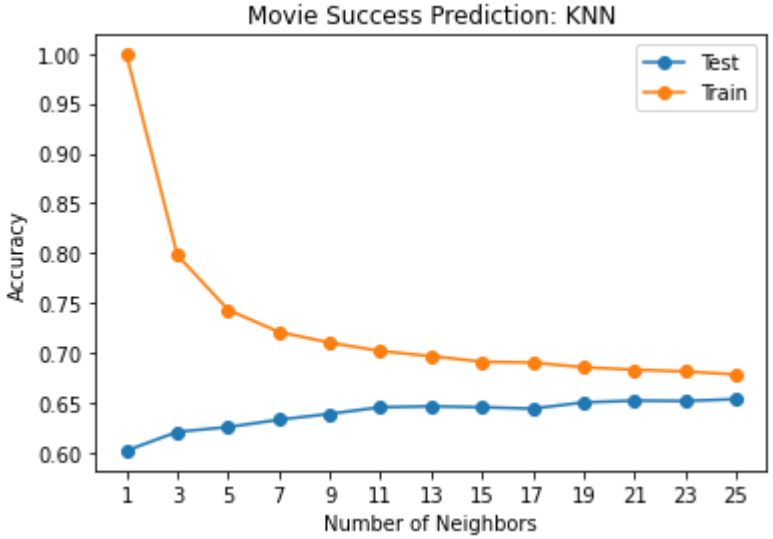
b. KNN
   - For gross:
     best accuracy: 0.5761



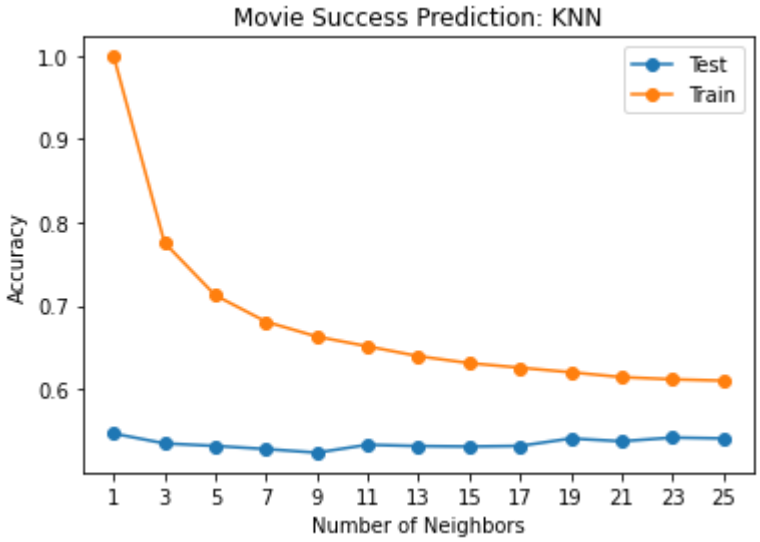
   - For imdb_rating:
     best accuracy: 0.7038



- For critic_score: 0.6537
  best accuracy: 0.6537

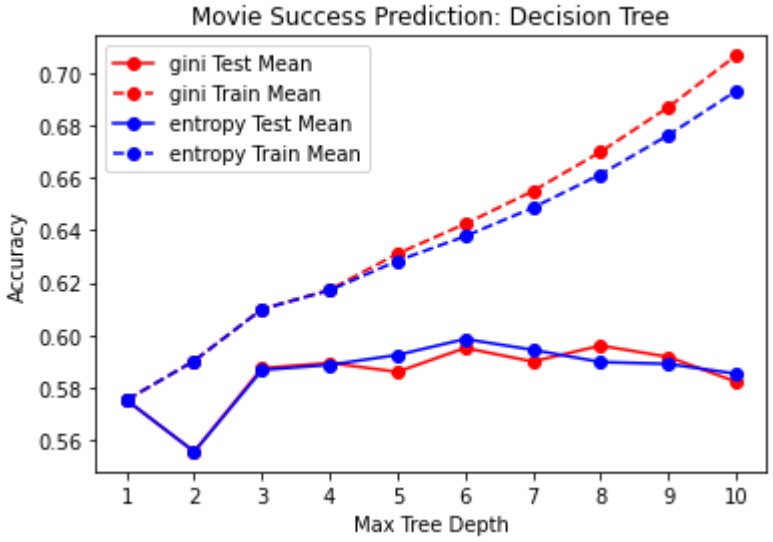


- For audience_score:
  Best accuracy: 0.5467

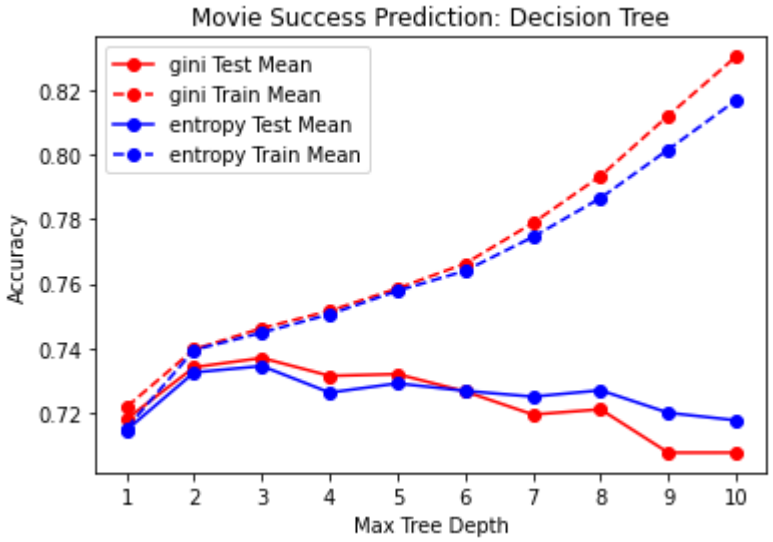


c. Decision Tress

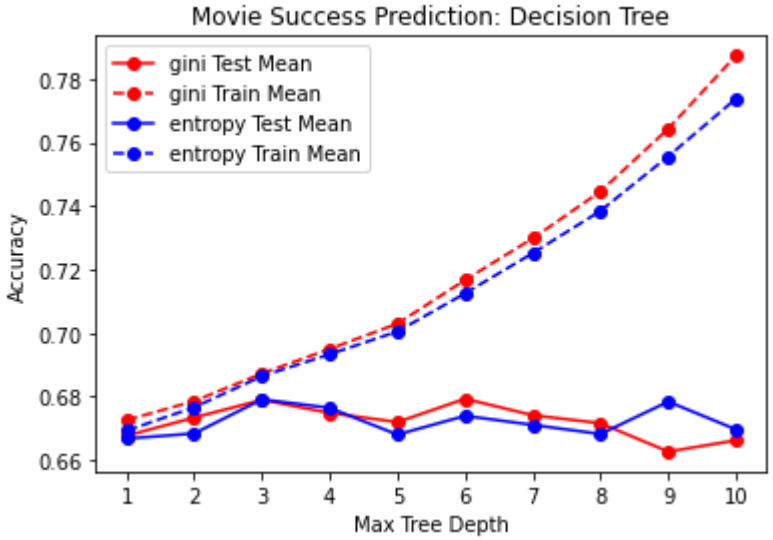
For gross: 0.5985



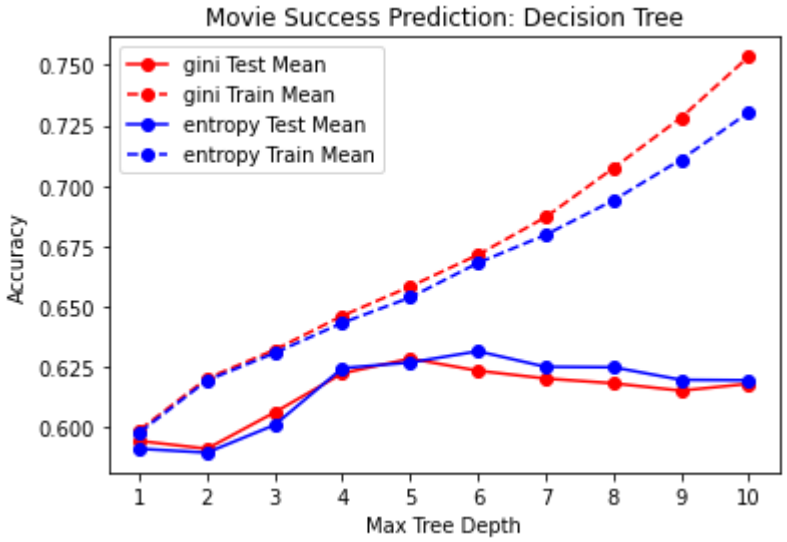
For imdb rating: 0.7369



For critic_score: 0.6793

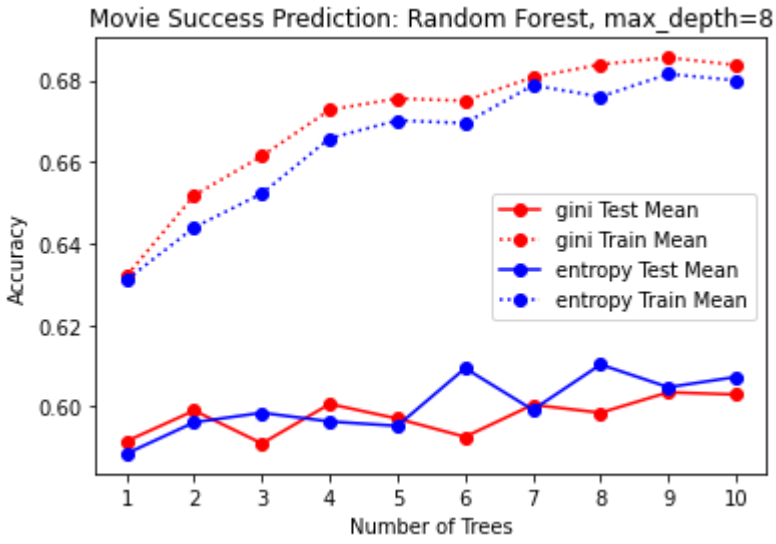


For audience_score: 0.6316

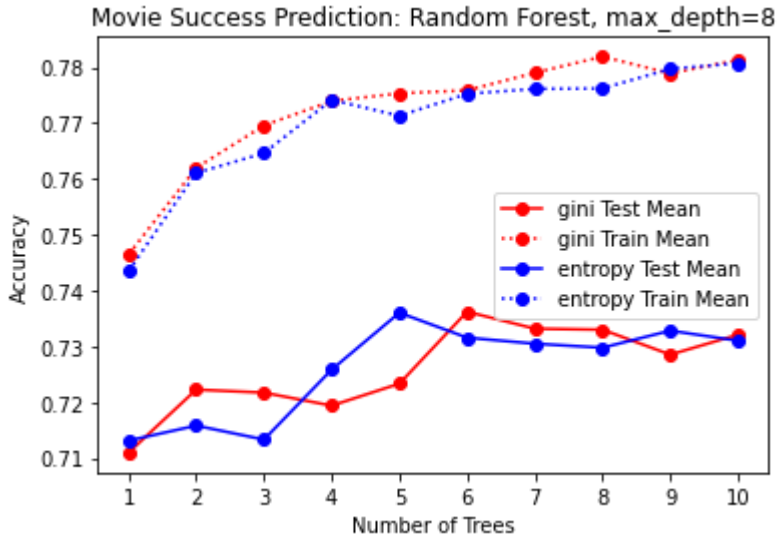


d. Random Forest

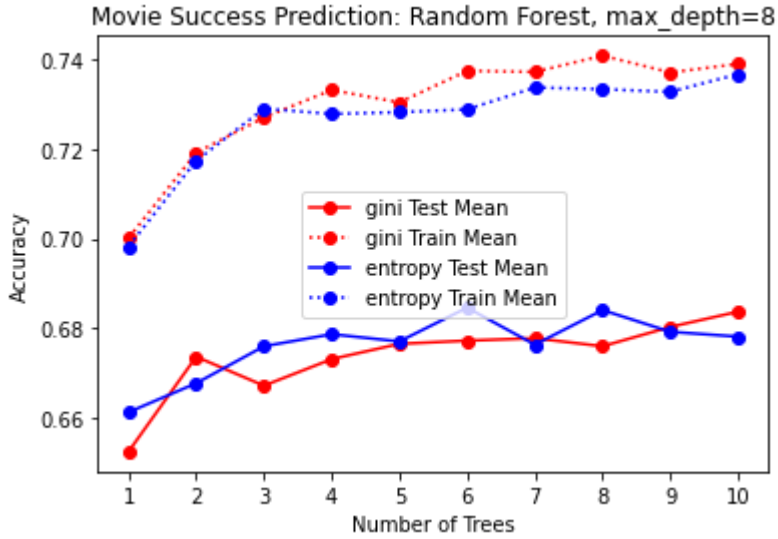
For gross: 0.6103 (Hyper-parameters = {'n_estimators': 8}



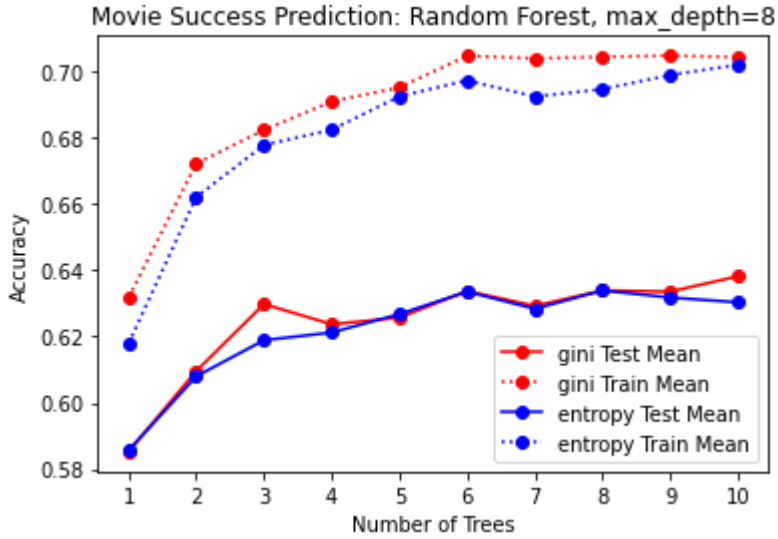
For imdb_rating:0.7362



For critic_score: 0.6846



For audience_score: 0.6382



4. Conclusion

After building the four models we found out that the random forest model represents the movie features more accurately. The success percentage for all models, while not good enough for industrial use, is in the close proximity of values obtained in previous studies. Some of the results obtained are better than that of some standard libraries and similar studies. Even though results are not good enough for industrial purposes the models built can be used in online applications. A larger training set is the key to improving the performance of the model. We need to consider additional features to achieve this. Director and cast analysis, plot summary analysis could be done and the information thus obtained could be added to the training set.

5. Summary

   Code repository link: https://github.com/NguyenTrang308/predict_movie_success

6. References

   Jeffrey Ericson & Jesse Grodman, A predictor for movie success, CS229, Stanford University

   Introduction to Business Analytics course notes

   https://github.com/timothyng-164/Movie-Success-Predictor

   www.imdb.com

   www.rottentomatoes.com

   https://www.calculator.net/inflation-calculator.html?