Capstone

Script Success Predictor

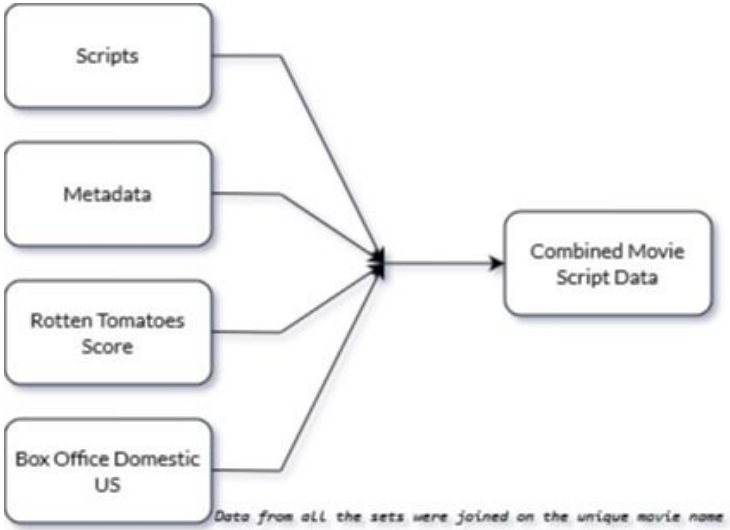
Viraj Kunthe• 07.05.2022

Overview

- Introduction
- Data Collection
- Data Description
- Clean-up and Modelling
- EDA
- Results
- Conclusion

Data Collection

- Boxoffice Mojo Alltime Revenue Data - Kaggle
- The Movies Dataset -Movie metadata - Kaggle
- Movie scripts- Scraped from the internet. (978 films)
- Rotten Tomatoes score- Audience and Critic - Scraped

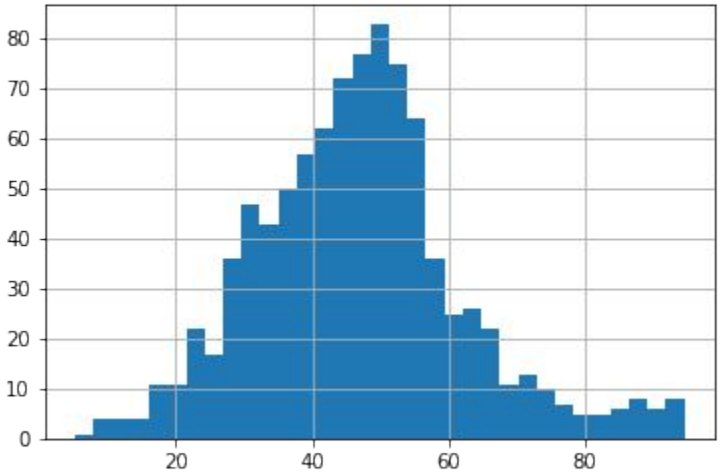


Data description

- Data points represent each individual movie being considered for this analysis.
- The columns represent the genre, imdb score, year, tags, script text, US box office performance and a calculated combined success score.
- The target variable is the success status based on the combined success score.
- Features such as parts of speech, genres, languages spoken, number of words, etc. were extracted from the script of the film.

Column	Description
movie_name	Name of the movie.
script	The entire script of the movie.
spoken languages	Languages spoken in the movie.
genres	All the genres that the movie falls into.
is_success	Whether the movie is / isn't a success (success_score >55).
success score	Calculated success score.
belongs to collection	Whether the movie belongs to a franchise of movies.
year	Year of release.
prop_count	Count of proper nouns.
verb_unique_percent	Percentage of unique verbs.
noun_unique_percent	Percentage of unique nouns.
verb_percent	Percentage of verbs.
noun_percent	Percentage of nouns.
adj_percent	Percentage of adjectives.
prop_nunique	Count of unique proper nouns.
adj_unique_percent	Percentage of unique adjectives.
adv_unique_percent	Percentage of unique adverbs.

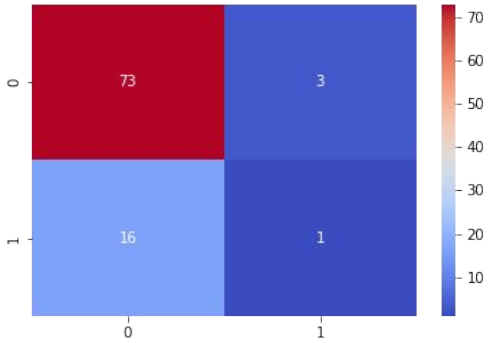
EDA - Wordclouds



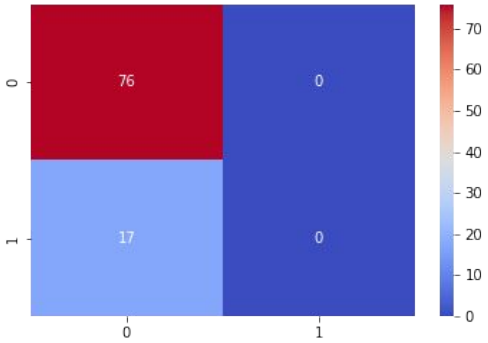
Clean-Up and Modelling.

- Have combined the data from all the sources.
- Downsized the metadata and box-office data to only the films for which the scripts were available.
- Applied NLTK tokenizer to get POS features.
- Implemented Wordclouds to get most common words by genre.
- Tried Logistic Regression, Decision trees and K Nearest Neighbours to predict the success status of the movie.

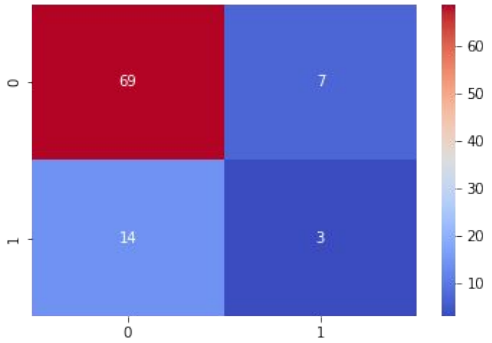
Results



Decision tree - 79.56% accuracy



KNN - 81.7% accuracy



Logistic Regression - 77.42% accuracy

Conclusions

- There is a positive indication towards our ability to predict whether a script would make a successful movie or not.
- The current dataset is too small to catch a pattern in the successful scripts.
- I envision continuing to work on this project and create a cleaner dataset of movie scripts that are uniform and ready to process and make it publicly available.

Questions ?