

Movie Predictor: Write-up

SMU Data Analytics Bootcamp | Capstone Project 4

Group 1: Alyssa DiFurio, Alex Flores, Bekhem Horne, and Vanessa Martinez

I. Introduction

The purpose of this project was to create a movie predictor that will be able to predict movie success based on specified metrics. We used a dataset from Kaggle called “The Movies Dataset” to create our model. Our inspiration came from looking at movie recommender models, projects from previous bootcamps, and our own love for movies. We originally planned on solely making a movie recommender machine learning system, but found that the files were too large and the time was too short.

II. Data

Data Source & Data Cleaning

Our data source was The Movie Database from Kaggle that contained 5 csv files. The size of all the files was very large, and we decided to use the main one called movies_metadata. This file contained information for over 45,000 movies with features including posters, budget, revenue, release dates, languages, production countries and companies. The data required a lot of cleaning before we were able to test out different machine learning models. We first dropped unnecessary columns such as poster links, original titles, original language, collection, and IMDb Id. Columns such as budget, revenue, and release_date, were transformed into float and datetime data types. We imported ast library, using a Try and Except statement and ast.literal_eval to parse genre, production_company, production_country, and spoken_language columns to evaluate

strings containing multiple values. The release date was separated out using apply/lambda to create release month, year, day of week, and weekend including Friday through Sunday. We then dropped null values and had 32,020 rows of data to work with.

Further data transformations included simplifying categorical columns. Data for genre was narrowed down from 20 to 11 values (Drama, Comedy, Action, Horror, Crime, Documentary, Adventure, Animation, Fantasy, Romance, Other). Production company data was also combined to narrow down the amount of values. We did some research to find that some of the companies listed had been bought by larger ones and combined those shorten the list of values. Production country and spoken language were also narrowed down. After the data frame was cleaned, we saved it to a csv to be able to create visualizations and dashboards in Tableau.

III. Machine Learning

For our machine learning model, we wanted to be able to predict if a movie will be successful based on specified metrics. We used the cleaned CSV to begin. We selected our target for the model to be voter averages greater than a value of six before dropping remaining columns not needed for the machine learning and then performing get dummies on the remaining data. Next, we began testing models to see which would be the most accurate for our desired results. Viewing our machine learning notebook you will see the Logistic Regression model to have the worst accuracy, with an F1 score of 65%, however no signs of overfitting. We continued to try several other models, Decision Tree Classifier had severe overfitting, Random Forest Classifier also had overfitting, AdaBoost was the first to yield good accuracy at 71% with no signs of overfitting, XGB Classifier at 73% , no overfitting and a Brier score of 1414. Our last model however, LightGBM yielded the best results. An F1 score of 74% with no overfitting, AUC of

.8177 and a Brier score of 1392. We decided between XGB Classifier and LighGBM that the LightGBM model was the clear winner for our model.

```
#Lgb Model
lgb = LGBMClassifier(random_state=42)

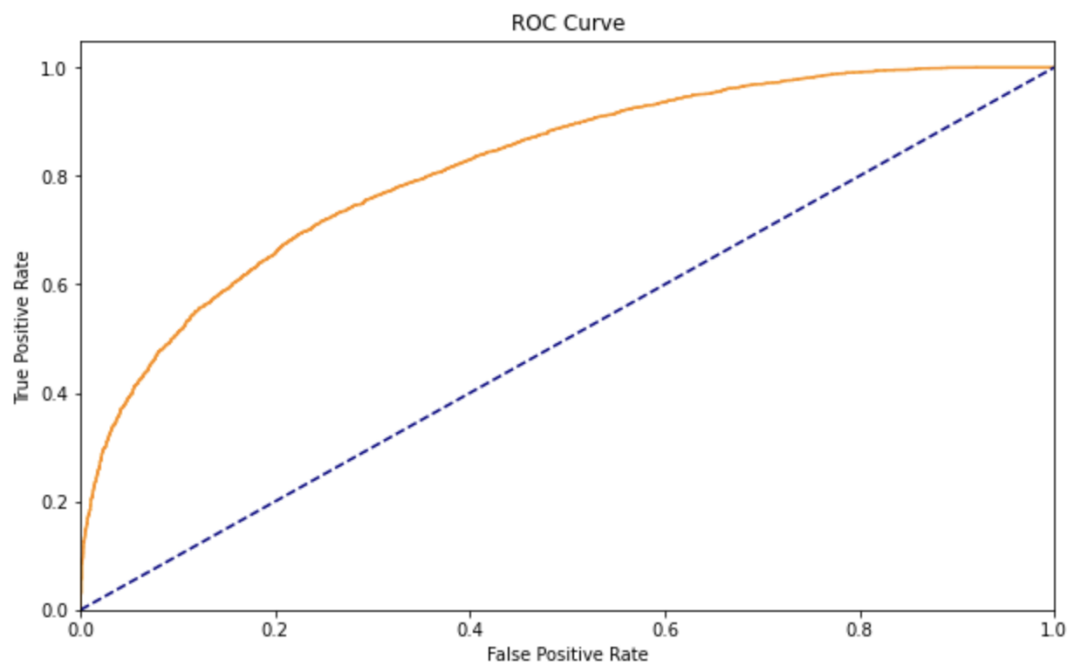
evaluateModel(lgb, X_train, X_test, y_train, y_test)
```

METRICS FOR THE TESTING SET:

[[2883 1011]
[1126 2985]]

	precision	recall	f1-score	support
False	0.72	0.74	0.73	3894
True	0.75	0.73	0.74	4111
accuracy			0.73	8005
macro avg	0.73	0.73	0.73	8005
weighted avg	0.73	0.73	0.73	8005

AUC for the Model Test Set: 0.8177237726534982



With our model chosen we moved forward to look at the feature importances. This was somewhat surprising as our top importance was not what we expected to see. Voter count, release year and runtime to name a few were the top three.

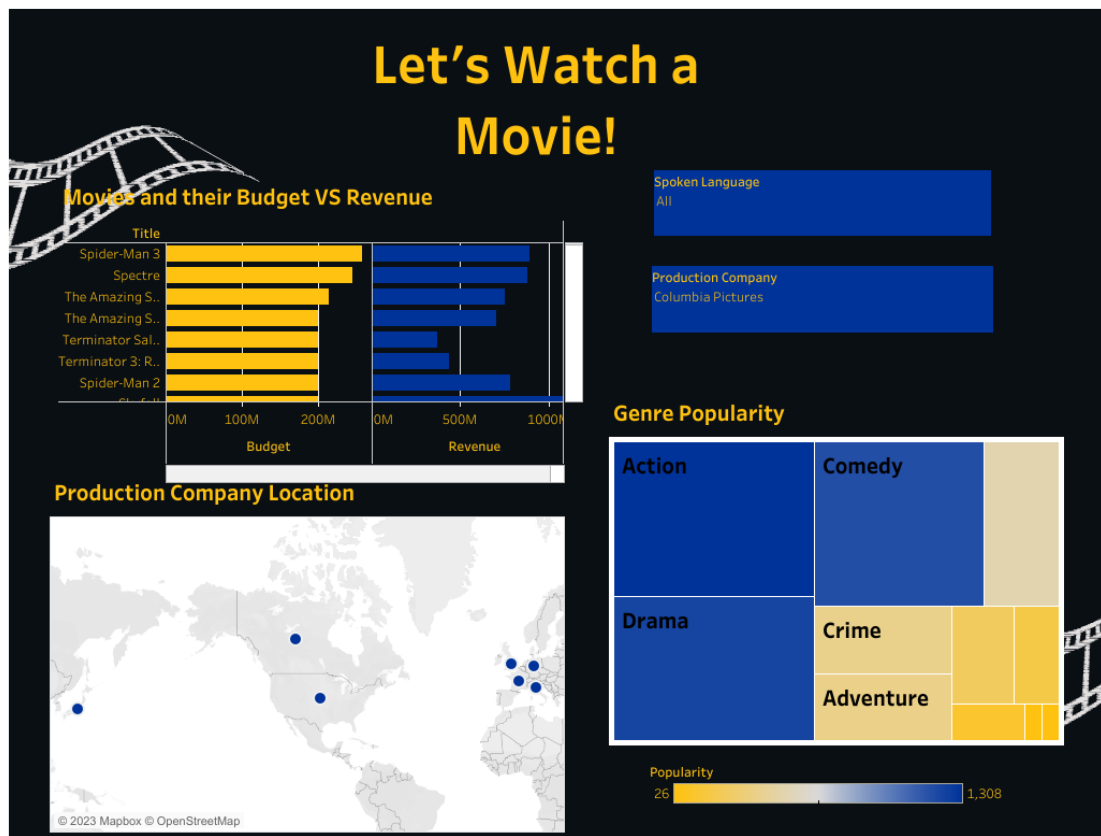
	Feature	Importance
4	vote_count	589
5	release_year	482
3	runtime	445
1	popularity	293
0	budget	196
2	revenue	140
6	release_month	132
13	genre_Documentary	52
40	production_country_United States of America	52
16	genre_Horror	50
14	genre_Drama	48
8	genre_Action	41
43	spoken_language_English	41
7	release_is_weekend	37
11	genre_Comedy	31
48	spoken_language_Other	29

Based on these results we mapped out the selection features we wanted to provide the users with on our search engine to ensure the results from the users inputs could potentially yield good results. Finishing the machine learning side of our project we created the form needed in our HTML code to build our search engine.

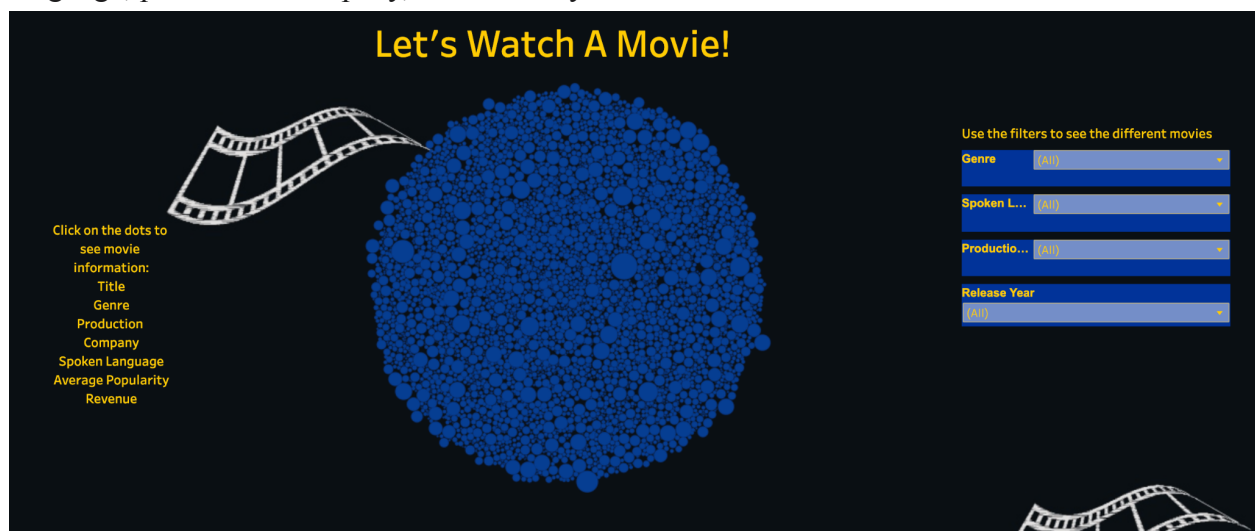
IV. Data Visualization in Tableau

We created 3 dashboards in Tableau to show different aspects of our data. Our color choices were representative of the Blockbuster logo to add some movie nostalgia into the project. The background included a plain black background with white film strips, to not distract from the main focus of the visualizations. The first dashboard shows all the movie titles comparing their budget versus their revenue, a map showing the production company locations, and a

treemap depicting the popularity of each genre. This whole dashboard can also be filtered by production company and spoken language.



The second dashboard is a movie title focused dashboard that shows a packed bubbles chart, with an overview of the movie data including title, genre, production company, spoken language, average popularity, and revenue. It can be filtered to show different movies by genre, spoken language, production company, and release year.



The third dashboard contains a bar chart with the movie title, release year and average vote count. It also shows a line graph of vote count and runtime minutes by release year. This was meant to be used to see if the amount of runtime minutes, and the year it was released, affects the amount of votes a movie has. Movies from the 1800s and early 1900s have lower vote counts compared to movies released later in the 20th and 21st century. Both of these visuals can be filtered by genre, production company, spoken language, release year, and release month.

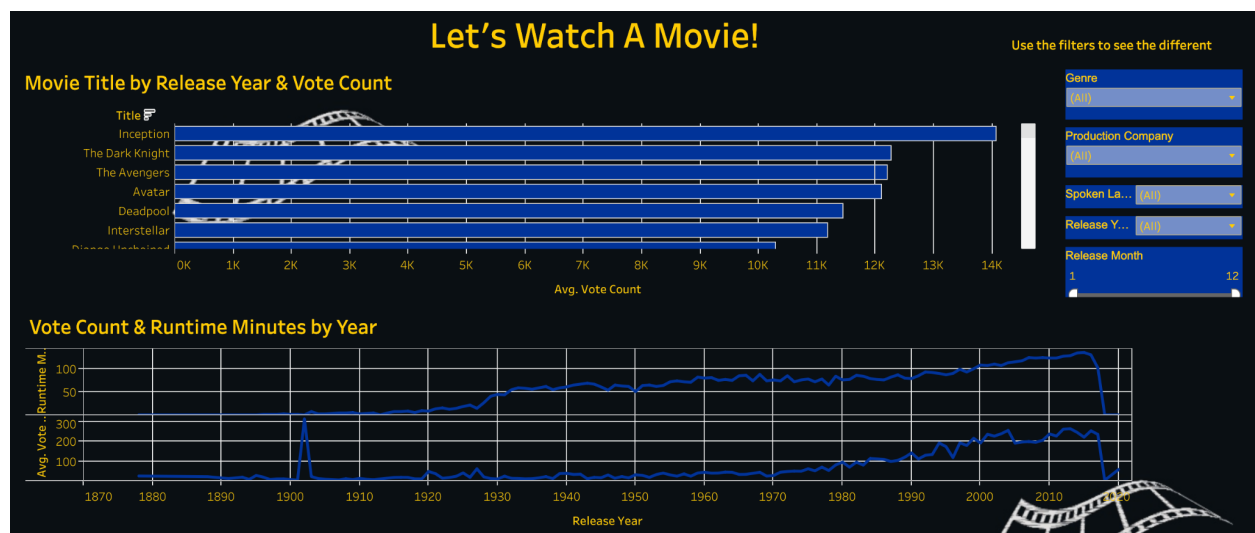


Tableau Inspiration

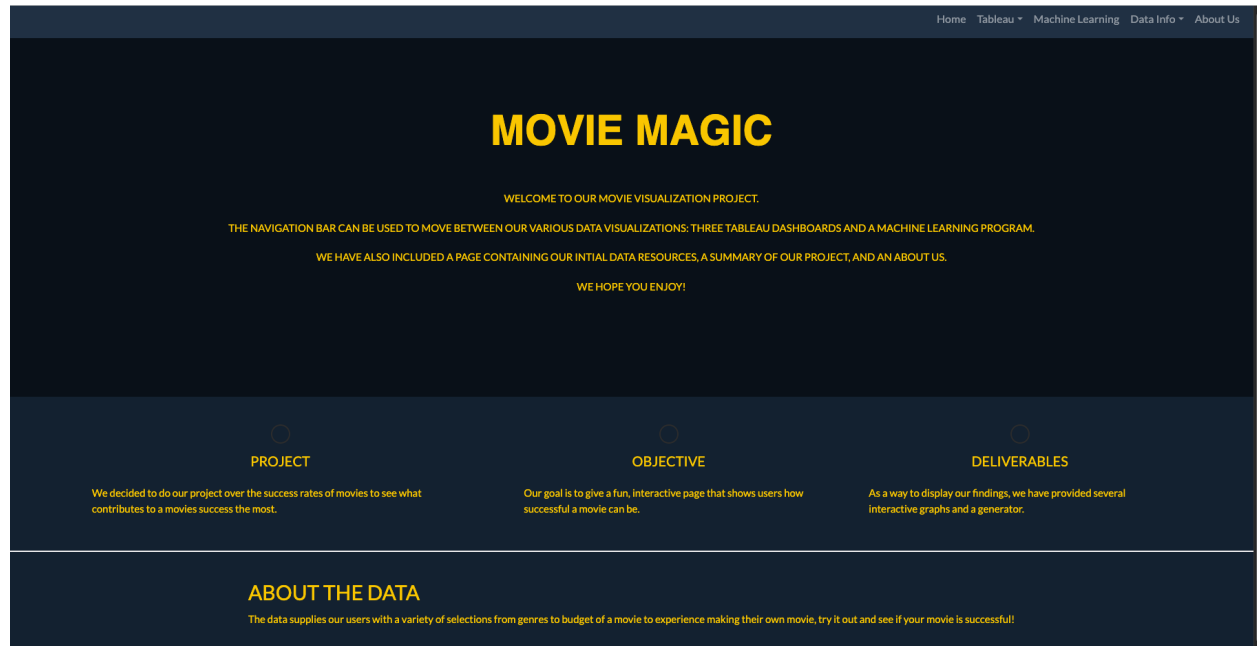
We looked up movie recommender dashboards on Tableau Public to use as inspiration for our dashboard design.

<https://public.tableau.com/app/profile/kalpitha4932/viz/MovieSystem/MovieAnalyser>

https://public.tableau.com/app/profile/kevin.flerlage/viz/MovieMoney_15813400492150/MovieMoney

V. Website App

For our website app we decided for the colors blue and yellow in the theme of the Blockbuster movie store colors. For the landing page we gave a brief description of what our webpage does, our objective, deliverables and information about our data.



ML Form - Serves live predictions to the user

Our machine learning page contains movie characteristics that can be altered and paired to see if the user's input will create a good movie or a bad one. These characteristics are budget, revenue, popularity, runtime in minutes, IMDB votes, year released, month released, genre, production company, and language spoken for the movie. Once a user inputs all the information they would like, the return will appear at the bottom of the search engine.

MAKE A MOVIE

SELECT FROM THE FOLLOWING MENU TO GENERATE YOUR OWN MOVIE!

Filters


Budget	Popularity	Revenue	Runtime (minutes)
<input type="text" value="300000"/>	<input type="text" value="20"/>	<input type="text" value="500000"/>	<input type="text" value="125"/>
IMDB Votes	Year Released	Month Released <input type="text" value="July"/>	Released on the Weekend <input type="text" value="True"/>
<input type="text" value="700"/>	<input type="text" value="2020"/>		
Genres <input type="text" value="Horror"/>	Production Company	Production Country	Spoken Language <input type="text" value="English"/>
	<input type="text" value="Columbia Pictures"/>	<input type="text" value="United States of America"/>	
<input type="button" value="Make Prediction!"/>			

About Us

The ‘About Us’ page is the last page on our navigation bar. We have each created a short paragraph to describe our educational history as well as our work history. We have also attached a picture of each of us for the website and added some brief fun facts.


[Home](#)
[Tableau](#)
[Machine Learning](#)
[Data Info](#)
[About Us](#)

Alyssa DiFurio




Alyssa holds a degree in International Relations and a masters certificate in Cybersecurity Policy and Management. She hopes to move into the intelligence or defense field and is taking this bootcamp to broaden her range of skills. Alyssa loves taking her cat on walks and spending time with her niece.

Alexandra Flores




Alexandra is from Dallas, Texas and she holds a degree in healthcare administration. After some time working as a CNA and bartending, she wanted to do something completely different so she decided to enroll herself into the Data Science bootcamp. Alexandra loves spending time with her dog Apollo and finding cool spots to eat.

Bekhem Horne



Bekhem is a full-time CPI/Lean lead in the Semiconductor industry. Taking pride in the learning and studying with team members is intriguing; he always looks forward to intaking substance from an ever-changing field. Bekhem also loves boxing and painting.

Vanessa Martinez



Vanessa is from Dallas, Texas and has a Bachelors of Science degree in Business Administration from UCF. She has an extensive background in floral design and is pursuing data analytics and visualization through SMU's Data Science Boot Camp. Vanessa's hobbies include shopping, cooking, working out, ordering Crumbl cookies, and binge-watching movies and shows.

Data Table - can be a sample, example of your raw or clean data

In our website we also included our raw data so the user could see where we received the data from as well as how we used it. This data is also from the CSV we got from our Kaggle dataset.

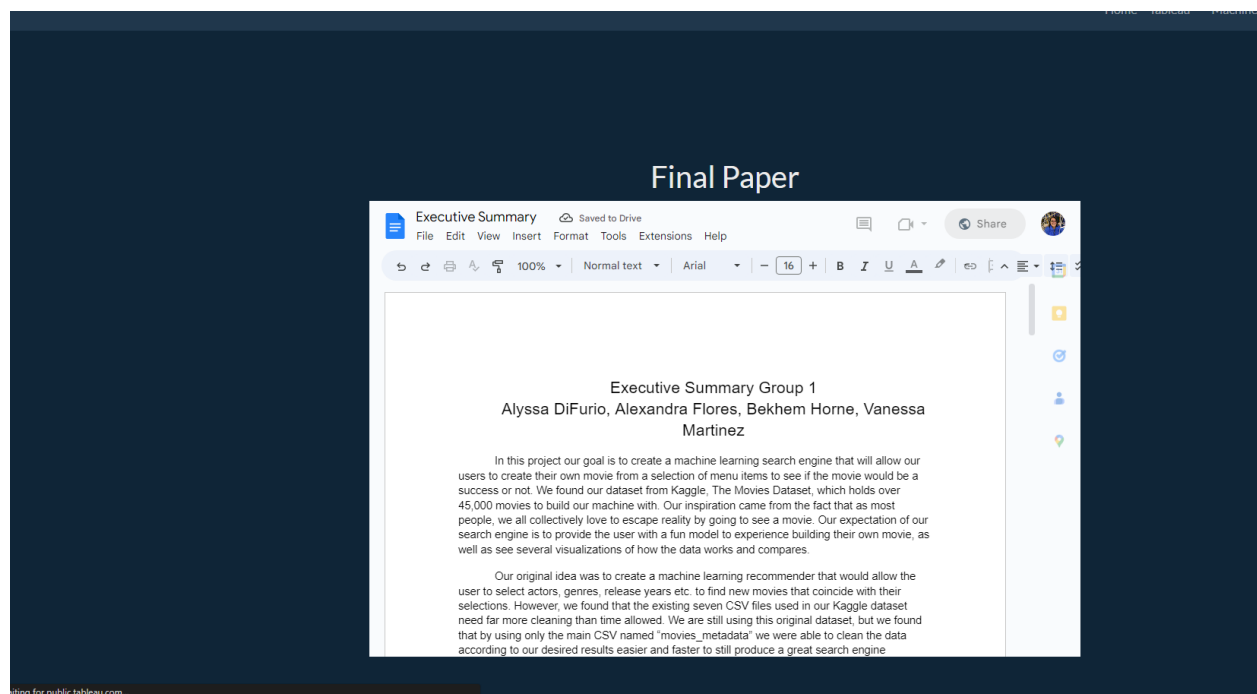
[Home](#)
[Tableau](#)
[Machine Learning](#)
[Data Info](#)
[About Us](#)

This data is all the data from the project.

	budget	popularity	release_date	revenue	runtime	title	vote_average	vote_count	genre	production_company	production_country	spoken_language	release_year	release_month	release_day	week	release_is_weekend
20073.0.0	2.382734	2014-08-09	0.0	116.0	The Fool	8.1	36.0	Drama	Other	Other	Russian	2014.0	8.0	5.0			True
22957.0.0	6.592284	2015-01-24	0.0	105.0	With This Ring	5.5	23.0	Comedy	Other	United States of America	English	2015.0	1.0	5.0			True
29245.2494400.0	1.779393	2017-01-27	0.0	89.0	The White King	6.3	10.0	Drama	Other	Other	English	2017.0	1.0	4.0			True
16296.0.0	6.651607	2012-02-25	0.0	101.0	Keep the Lights On	5.4	25.0	Drama	Other	United States of America	Other	2012.0	2.0	5.0			True
18641.0.0	0.142022	1942-08-18	0.0	58.0	Busses Roar	8.0	1.0	Action	Warner Bros.	United States of America	English	1942.0	8.0	1.0			False
13209.0.0	1.168074	1976-09-01	0.0	87.0	Massacre at Central High	6.3	7.0	Drama	Other	United States of America	English	1976.0	9.0	2.0			False
27166.0.0	0.000019	1970-08-02	0.0	101.0	May Morning	0.0	0.0	Horror	Other	Italy	English	1970.0	8.0	6.0			True
30319.0.0	1.069205	2008-10-11	0.0	83.0	Fist of the North Star: The Legend of Kenshiro	5.6	5.0	Animation	Other	Japan	Japanese	2008.0	10.0	5.0			True
11242.0.0	5.256960	2008-07-18	188126.0	98.0	Repo! The Genetic Opera	6.7	103.0	Horror	Lions Gate Films	United States of America	English	2008.0	7.0	4.0			True
3396.0.0	1.375675	1999-01-01	0.0	112.0	Nowhere to Hide	6.8	15.0	Action	Other	Other	Other	1999.0	1.0	4.0			True
16795.0.0	4.011610	2012-09-03	0.0	89.0	Beverly Hills Chihuahua 3 - Viva La Fiesta!	5.8	33.0	Comedy	Walt Disney Pictures	United States of America	Other	2012.0	9.0	0.0			False
3776.20000000.0	4.619593	1988-03-25	51684798.0	106.0	Biloxi Blues	6.4	72.0	Comedy	Universal Pictures	United States of America	English	1988.0	3.0	4.0			True
15429.20000000.0	3.174512	2011-04-14	17479.0	113.0	5 Days of War	5.8	63.0	Action	Other	United States of America	English	2011.0	4.0	3.0			False
8561.0.0	7.132709	2002-09-23	0.0	120.0	Solino	7.0	12.0	Comedy	Other	Germany	Deutsch	2002.0	9.0	0.0			False
8374.0.0	0.944547	2004-07-22	0.0	95.0	Vares: Private Eye	6.2	16.0	Action	Other	Other	Other	2004.0	7.0	3.0			False
26707.0.0	1.395260	2009-05-18	0.0	90.0	Daniel & Ana	5.9	11.0	Horror	Other	Other	English	2009.0	5.0	0.0			False
27669.0.0	5.453322	2016-04-22	0.0	100.0	Special Correspondents	5.6	159.0	Comedy	Other	United Kingdom	English	2016.0	4.0	4.0			True

Executive Summary

Lastly we included a summary of the information on our website. We included our struggles with the data cleaning and our thought process throughout this project. In this summary we also included information about our machine learning model as well as snippets and descriptions of each of the three Tableau dashboards. Lastly, we spoke about our objective of this project and its purpose for its users, to make a good movie prediction. This page can be found under the ‘Data Info’ drop down menu.



VI. Limitations & Future Work

A significant limitation came in our initial data cleaning. The Movies dataset has seven CSVs that we found needing immense cleaning, sadly we did not have time to clean all seven and correct some of the issues causing file corruption, so we changed our project from a movie recommender to a movie search engine.

In the future we would like to revisit this once we have more time to see if we can potentially get the CSVs in working order to build a movie recommender. We would also like to improve upon our Tableau dashboards to make them more dynamic and uniform.

VII. Conclusion

This project challenged us to apply all of the skills we've learned throughout this bootcamp and we thoroughly enjoyed building our machine learning model, dashboards, and webpage. We look forward to seeing the results our page generates for the user.

VIII. Works Cited

https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset?select=movies_metadata.csv

<https://public.tableau.com/app/profile/kalpitha4932/viz/MovieSystem/MovieAnalyser>

https://public.tableau.com/app/profile/kevin.flerlage/viz/MovieMoney_15813400492150/MovieMoney

<https://coolors.co/>