Maddi Ross

Final Report:
Box Office Movie Revenue

Problem Statement

The global box office film industry makes tens of billions of dollars every year, peaking at over $40 billion dollars before the coronavirus pandemic, and but only hitting around $20 billion in 2021. In a time where the box office is losing market share to other spheres of the entertainment industry, the question of what makes a blockbuster is more relevant than ever.

By using a variety of metadata obtained from The Movie Database (TMDB), we will produce a model for predicting box office revenue that can be used to justify, plan, and optimize movie features before (and during) production. We used a variety of exploratory and analytic techniques to analyze features such as crew and cast composition, languages, production companies, countries, plot key words, budget, and more.

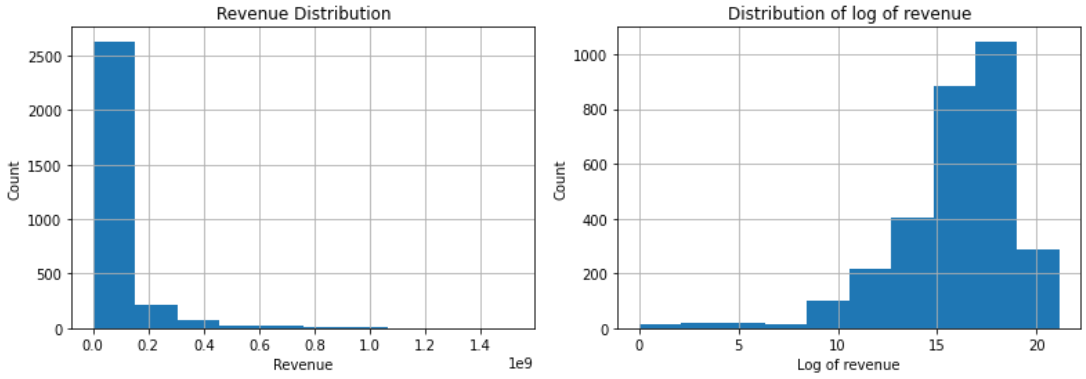
Dataset

The raw data came from TMDB's website (as downloaded from Kaggle). Films included in the dataset spanned release dates from the 1920s until 2018. The original 23 features included for the 3000 films in the dataset were as follows:

id:	the film's unique id on TMDB
belongs_to_collection:	if the film was part of a collection/series, the series title
budget:	film budget, USD
genres:	a list of genres
homepage:	if the film has a homepage, the link to the site
imdb_id:	the film's unique id on imdb
original_language:	the film's original language
original_title:	the title of the film in its original language
overview:	the description of the film provided on the TMDB site
popularity:	a metric tracking positive user interaction with the film on TMDB
poster_path:	a link to film poster(s)
production_companies:	list of production companies that produced the film
production_countries:	list of countries involved in the film's production process
release_date:	date of film release (day, month, and year)
runtime:	film runtime length in minutes
spoken_languages:	list of languages spoken in the film
status:	note of if the film is already released or still in pre-release stage
tagline:	the film's tagline (if applicable)
title:	the film's English language title
Keywords:	keywords associated with the film (if applicable)
cast:	information on cast including actor names, genders, and roles
crew:	information on crew including names, genders, and roles
revenue:	ultimate film box office revenue, USD – our target variable

Data Cleaning, Wrangling, Exploration

We began by inspecting our target variable: revenue. It was deeply right skewed, so we took its natural log for further modeling.



We then cleaned errors in release dates, budget, and revenue that had been scraped wrong (or perhaps were noted incorrectly on the TMBD at the time of scraping). Features that served purely for identification purposes (id and imdb_id, title, original_title, poster_path, overview) were all dropped from our dataset.

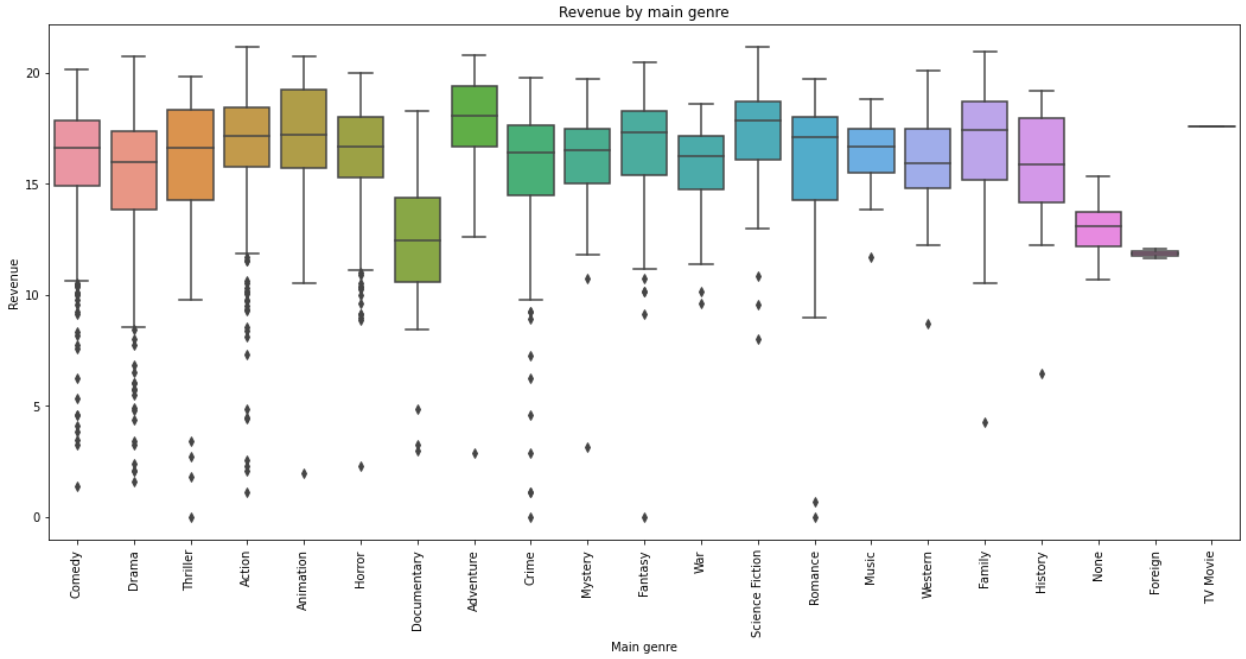
We then reorganized several variables as Booleans (homepage, status, tagline). Instead of tracking the links to the homepage, or the content of the tagline, we tracked the existence of a homepage, the existence of a tagline, and whether or not a film was released (status). Similarly, we replaced the information on collection name with a Boolean tracking simply whether or not a film was part of a collection.

Information on production company was used to create several new variables. We first created variables tracking how many production companies were involved in film production, and then new variables to track whether or not a film was produced at least in part by one of the top 50 production companies. The same was done for production countries – a new variable was created to store the number of countries involved in production, and, as the top 25 production countries accounted for more than 95% of all films, variables were created to track which of these top 25 countries in particular were involved in the production of each given film.

Similarly, a new variable was created to track the number of languages spoken in a film. As with production countries, the top 25 spoken languages accounted for languages spoken in more than 95% of films, so we created variables to track which of these top 25 languages was spoken in a given film in addition to the existing information on film's original language.

Information on cast and crew was treated similarly. Firstly, variables were created to store the size (number of people) involved in both cast and crew (cast_size and crew_size). Then the numbers of individuals of a particular gender (0 – meaning unlabeled, 1 – meaning female, or 2 – meaning male) involved in the cast and crew of the film were separately extracted. We also created variables to track the presence each of the 25 most prolific actors and crewmembers in a film.

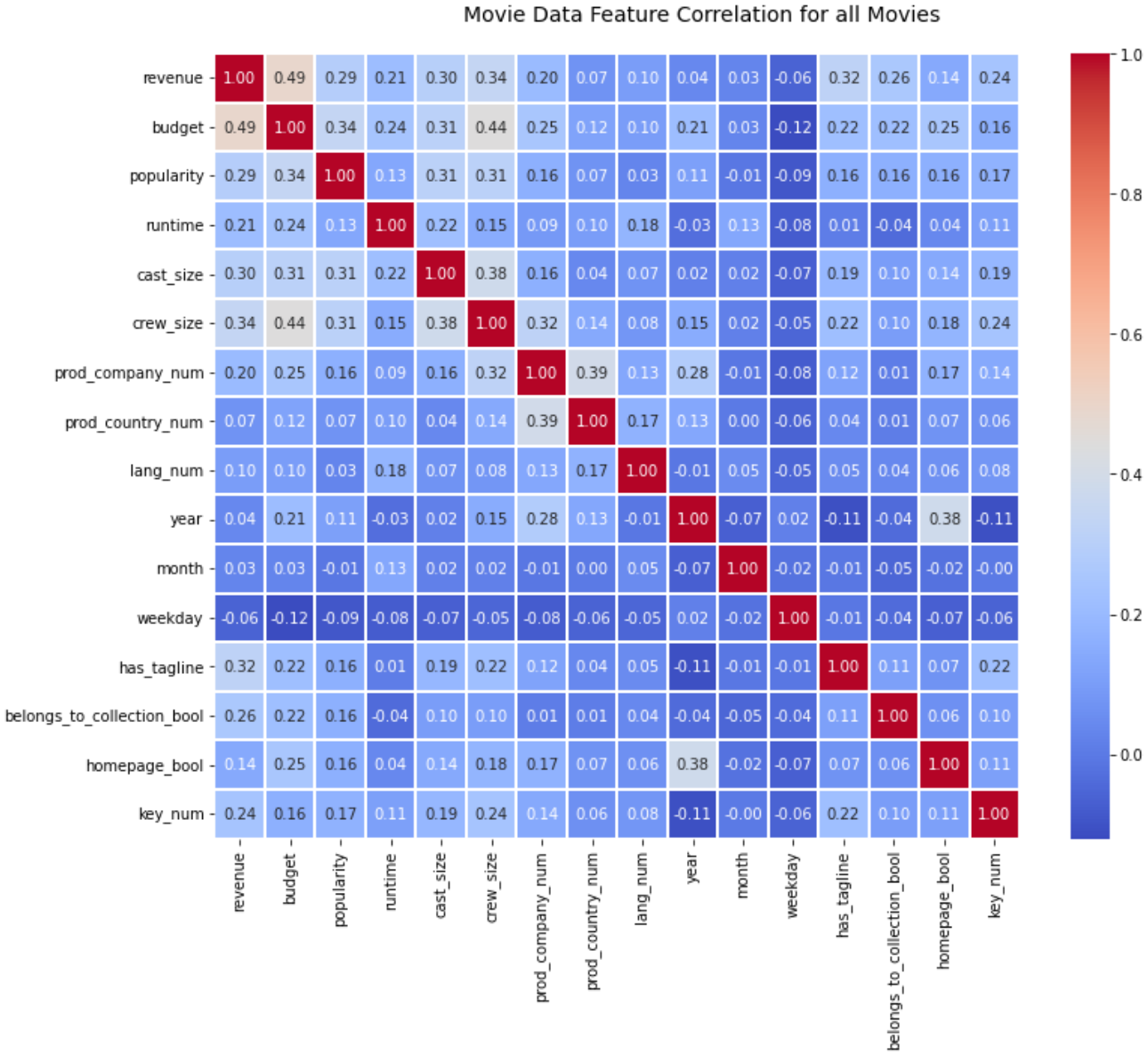
We created variables to track whether a film was classified as belonging to each of the 21 listed genres. Films of different genres had noticeably different revenue distributions (arranged below by the first, or "primary," genre associated with a film):



We also created a new variable to track the number of keywords (a modern marketing technique) associated with a film.

Finally, we used the information on release date to store information on day of week, month, and year of release.

Our numeric data showed certain promising correlations with revenue, particularly with budget, but also with cast size, crew size, popularity, runtime, collection status and more. However, it should be noted that even a simplified feature correlation matrix missing many of our newly added features show some collinearity. Doubtlessly, in our final dataset there are many more. We move onto modeling with 230 features for each film.

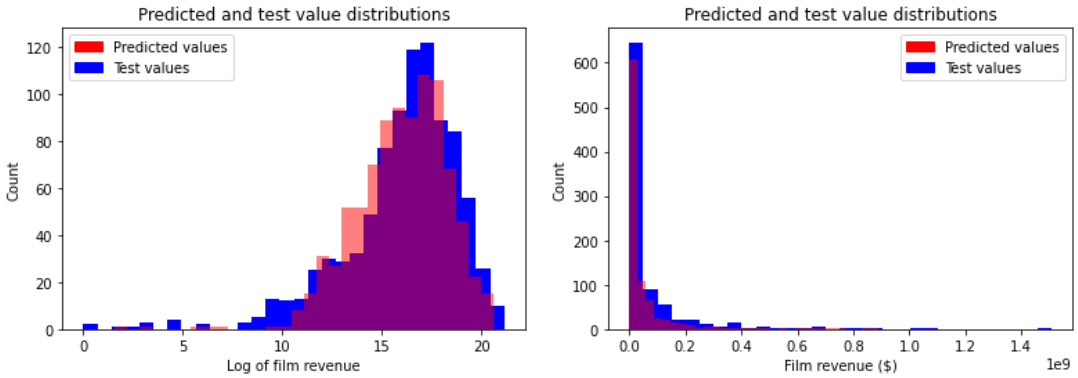


Modeling

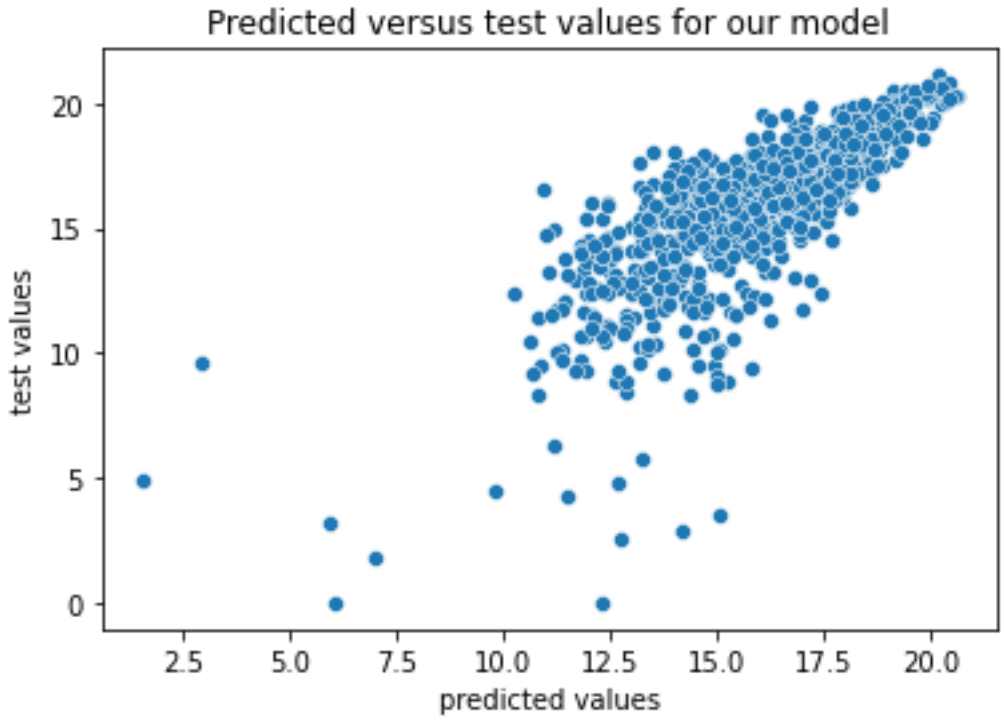
We experimented with four different model types: Linear Regression, Random Forest Regression, KNeighbors Regression, and XGBoost Regression.

The XGBRegressor performed best on our data, with an r2 of around 0.6 on both the test data and as a mean of cross validation on the training data. It had a root mean squared error of 1.847, the lowest for all of our models.

The distribution of the values predicted by our model and the actual test values are displayed below both for the log of revenue (as predicted by our model) and for revenue in USD post transformation:

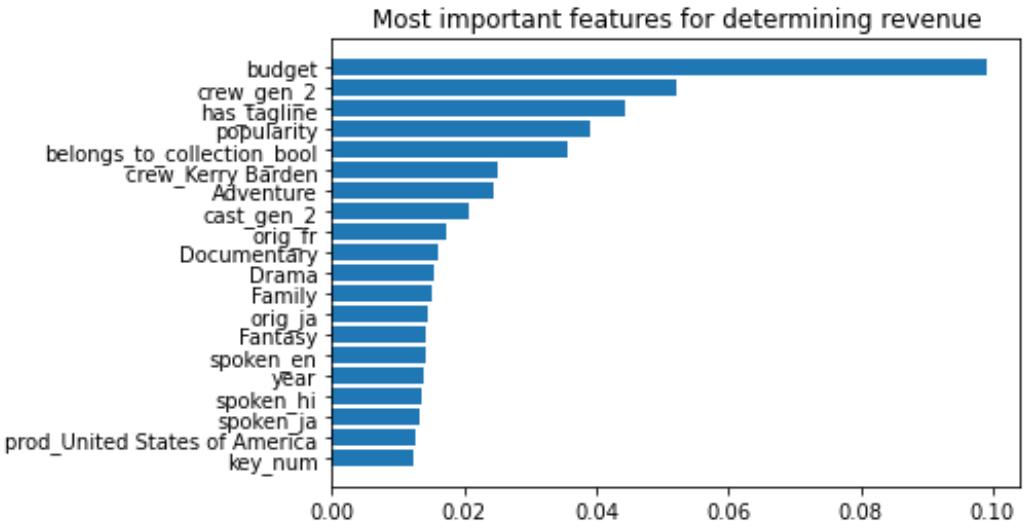


We can see that our model performs worse at lower values, but performance improves as revenue values increase.



This model is by no means perfect. It does not have a particularly high r2 score, but its comparatively low root mean squared error makes it the most useful of our models given the information available to us. As our target variable values range from 0 to around 21.5, the root mean squared error is not an overlarge number.

The model can also provide utility with regards to its most important features:



The most important feature for determining revenue according to our model is by far budget, which should come as no surprise.

Other important features relate to the gender of the crew and cast (male heavy films are associated with higher revenues), and existence of various marketing features (tagline, keywords). Films that are part of collections also are associated with higher returns, which might explain the number of sequels, prequels and remakes that continue to flood the movie market. Various genres and languages are also associated with higher revenue, as is casting director Kerry Barden.

Further analysis of feature importance also makes it clear that many of our added features are unimportant and could be removed, based on the extremely low importance value for many of our hundreds of added features (103 of which have absolutely no influence on the model).

Popularity is the fourth most important feature in determining revenue according to our model, which is something of a flaw of any model built using this data, as popularity is our most uncertain feature – it is determined by user 'activity' related to a film on the movie database site, which does exist pre-release but is much more of a post-release factor, and such user activity often only begins when the film is already advertising part way into the production process.

We therefore developed a variant of this model removing popularity, so that the model could be referred to in the pre-production phase as well as during production.

This model, also an XGBRegressor with all the same parameters and hyperparameters (as listed in model_metrics.txt), results in only slightly worse model performance, with an r2 of 0.592 on the test data and a mean cross validation r2 score of 0.589 on the training data.

Similarly, the root mean squared error increased to 1.878 (+0.031), which still leaves this model performing better than any of our other models with Linear Regression, KNeighbors Regression, or Random Forest Regression (all with r2 scores over 2).

	train score	test score	RMSE
Linear Regression	0.493	0.402	2.272
KNeighbors Regression	0.524	0.432	2.214
Random Forest Regression	0.813	0.498	2.082
XGBoost Regression	0.803	0.605	1.847
XGBoost no popularity	0.783	0.592	1.878

Ultimately, the XGBoost model is our model of choice, but the choice of using a the model with or without popularity may depend on the circumstances. In general, there is little reason not to use the model without popularity, as it does not significantly worsen model performance and avoids possible issues caused by film production and release status, but if trying to use the model to predict film box office revenue for a film already in production that is already being marketed, including online, and therefore has a presence on TMDB, the usage of the model including popularity is admissible.

Takeaways and Future Research:

This model has value for production companies, studios, and other stakeholders involved in the production and pre-production film creation process. It could be useful for optimizing features of proposed movies or movies already undergoing production or pre-production in order to increase revenue (such as establishing the revenue increases associated with adding languages in order to establish their net value, the predicted returns of a budget increase, the value of investing in various advertisement features or producing on-location in a given country, etc.). It could further be used to evaluate film proposals, or as a tool to suggest the commission of films with specific features.

The box office industry has suffered deeply during the covid era, and is still certainly affected by the mass closures and maximum capacity rules that are still in force in certain parts of the world. Therefore, there might be some issues when evaluating films from 2020 on. However, adding data from that time period is unlikely to help our model predict box office revenue going forward, as box office revenue was deeply affected by exogenous factors as listed above that our model could not account for.

It is likely better to continue to use our model built on data from “better times” for predictions moving forward as the box office industry continues to try to restore itself to its former highs. However, we should keep in mind the consumer switch to home theaters and streaming during the covid era, as it’s possible that changed habits over the last few years could have a negative effect on our model’s predictive capability. It is difficult to predict whether these habits will have sticking power as a post-covid world continues to develop and normalcy is slowly restored across much, but not all, of the world.