Capstone Project 1
Predicting Return on Investment at the Box Office

Benison Pang, Ph.D

1. Introduction

Problem Statement

The question of what makes a movie successful at the box office can seem somewhat intractable at times. There are movies which are critically acclaimed, but completely fail at the box office (The Shawshank Redemption, Blade Runner), and movies which are critically savaged but are massive successes (the Transformers franchise, for example). Even beloved classics such as the Wizard of Oz were box office failures in their day. As such, the biggest question that faces every studio in Hollywood is simple, but difficult: How do we maximize the profitability of a movie for a given budget?

Client Perspective

The question is simple, but the approach taken to that question can differ significantly between studios. If a studio is one of the 'Big Six' (Disney, Paramount, Columbia, Universal, 20th Century Fox, and Warner Brothers), then the performance of high-budget blockbusters may be imperative to the overall stature of the company, and less so in the less popular genres. Similarly, an up-and-coming studio may also be interested in high-budget movies in trying to gain a foothold in the industry, but may also need to invest in lower-budget films as a result of less capital to begin with. Lastly, small studios may be forced to take a completely different approach given limited capital to begin with.

In all these cases, understanding how features of a movie contribute to its box office performance is of paramount importance, as having a firm understanding of the importance of various movie features in generating revenue will allow companies to better tailor their filmmaking strategies. As such, the goal of this project is to build a regression model that will be able to predict the return on investment for a movie.

2. Data Acquisition, Cleaning and Wrangling

Acquisition

These data were acquired from two sources: Boxofficemojo.com, and the Open Movie Database (OMBD).

Boxofficemojo.com keeps a yearly list of movie grosses, which I scraped primarily through the use of the *BeautifulSoup* module in Python, searching each movie's specific page to retrieve their domestic gross, budget, genre, budget, and release date. Looking over the past 20 years

of data, I managed to scrape around 6800 movies' worth of box office information, which was then stored as a Pandas dataframe.

OMDB was an important source of information for movie ratings, runtime, and various metrics of critic and audience evaluations, all of which play into the profitability of a movie. These data were requested through the OMDB API; however, the data was relatively lacking compared to Boxofficemojo's database, and so the number of results returned for many of these variables fell in the 2000+ range, significantly smaller than the number of results from Boxofficemojo.

Joining

To combine the two dataframes, I performed an outer join. This was done to ensure that movies which had only Boxofficemojo data would still be kept in the final dataframe even if they were missing in the OMDB dataframe, and vice versa. This ultimately led to a database of 7130 movie titles and 13 total variables.

Defining the metric of interest

We are primarily interested in understanding what variables contribute to a movie's profitability. To that end, I will be using the Domestic ROI Multiplier (DRM) as the dependent variable, which we use to define what constitutes a profitable movie. The DRM is simply defined as the movie's absolute gross divided by its budget; as such, a DRM value of 1 is a break-even point; any movie with a DRM > 1 is profitable, and any movie with a DRM < 1 is unprofitable.

Cleaning

Improper Data Types

One problem that arose frequently in several of the variables was that they were recorded as string data types, even if floats or integers were preferable. A cell in the domestic gross column, for example, would be recorded as '$120 million', instead of '120000000'. Since this project was developed with regression in mind, it was necessary to clean any columns that recorded float values as strings.

There were two primary steps I took for this column and others like it. Firstly, I applied an element-wise lambda function in order to remove any non-numerical symbols in each of the cells using regular expressions; this made the process faster and the code cleaner, which was preferable to creating custom functions. To allow mathematical operations, I then converted each column into float data types by using the native Pandas conversion method.

Missing Values

Because of how the data was initially recorded from the scraped webpages, missing data was often recorded as a string: 'N/A'., or as None values. For several columns, I wanted to perform mathematical operations, and so I had to recode these as the standard numpy NaN values. This was done through the .replace() method native to Pandas dataframes.

As each data point was unrelated to each other (e.g. movie budgets) and because there were substantial amounts of missing data in some variables, it did not seem appropriate to use any approximation to fill in missing values.

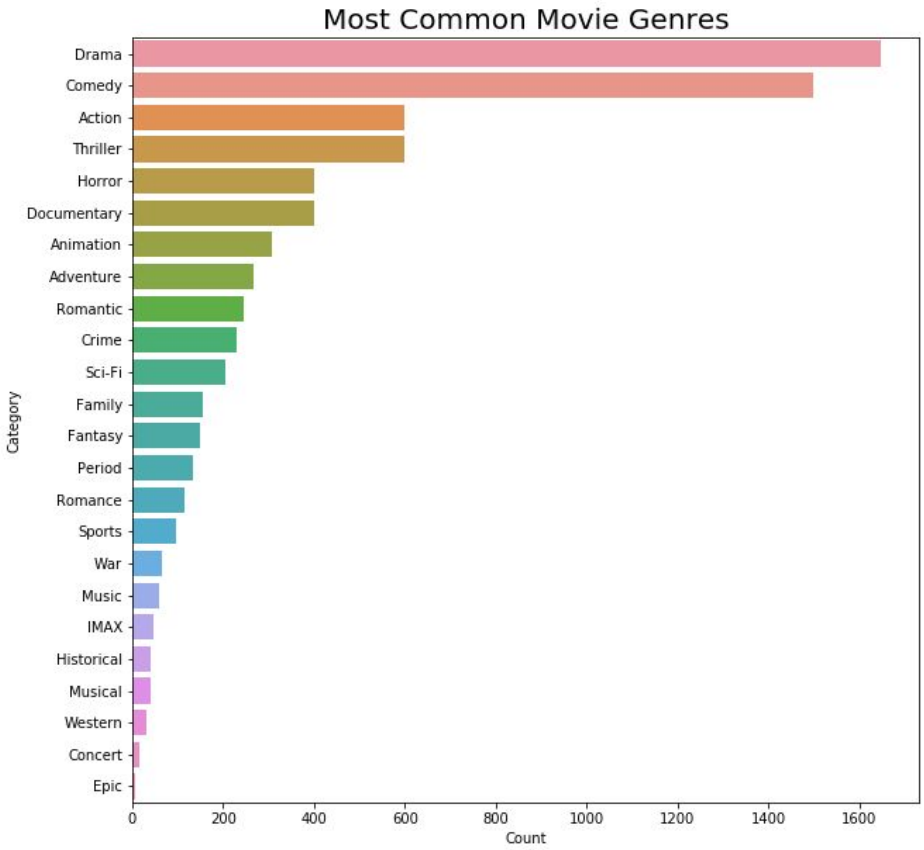
Outliers

There were definitely some outliers in the box office database. Star Wars: The Force Awakens, for example, is far and away the highest grossing movie in North American history. However, insofar as it's box office performance may have still have been related to its release date, production budget and high audience ratings, it was not removed. On that same basis, I did not think it appropriate to remove any other outliers as none of them could justifiably be considered to be the result of error, but rather a true reflection of the movie industry.

3. Exploratory Data Analysis/Statistical Analysis

Which are the most common movie genres?

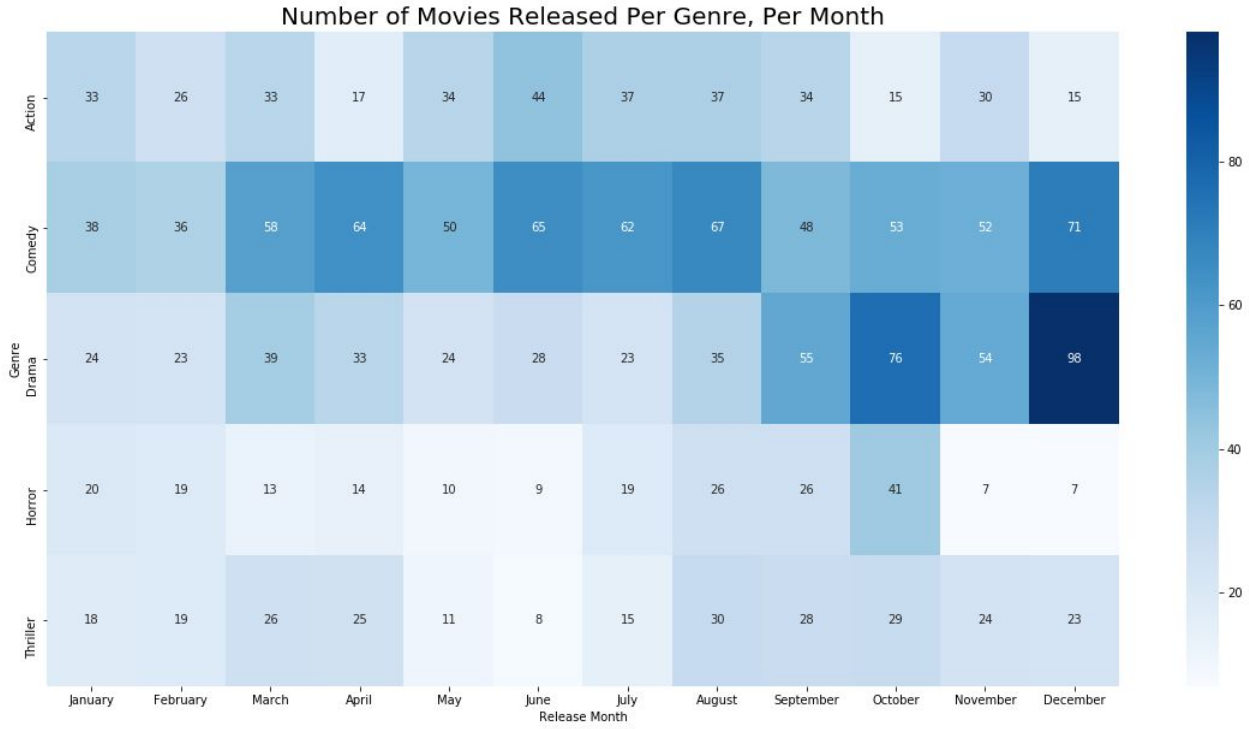
To get a sense of where movie studios are currently focusing their efforts, let's take a look at what movies are most often released in Hollywood.



Despite blockbuster movies looming larger in the public consciousness, it appears that drama is actually the most popular investment by movie studios, followed closely by comedy. These are the two most popular genres by far.

When are movies typically released for each genre?

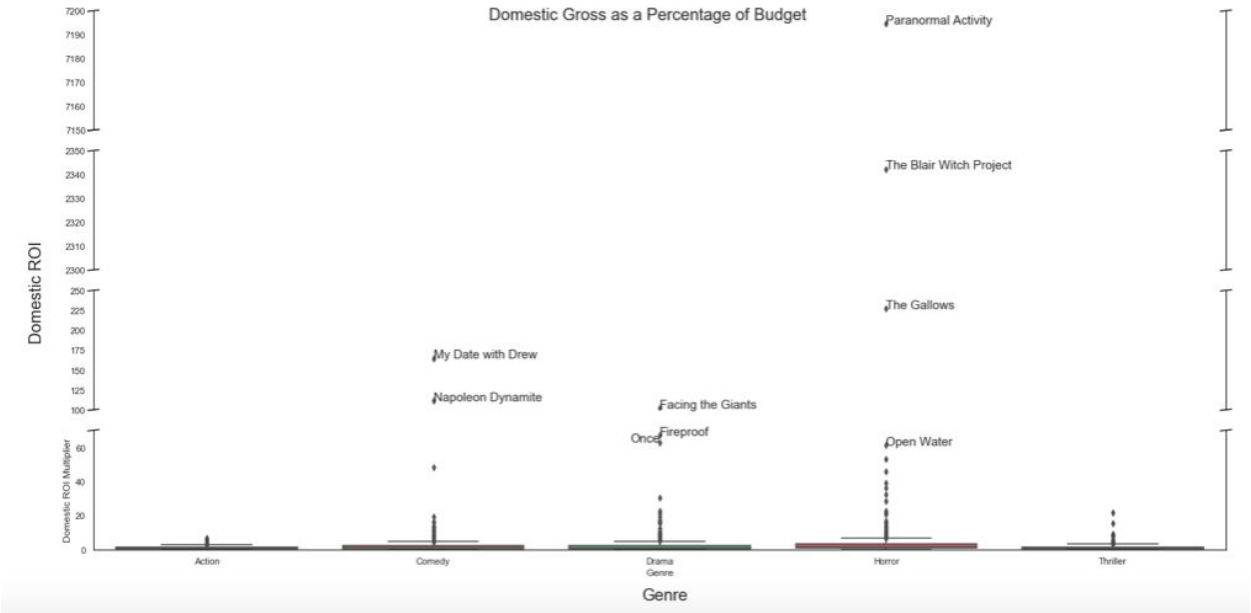
Knowing that release dates are one of the most important decisions made in the movie-making process, I wanted to look at historical data showing the typical behavior of movie releases based on the genre.



This visualization reveals some interesting trends! We immediately see the following:

- Dramas are primarily released towards the end of the year, starting around September. Why might this be the case? Probably because the Oscars take place at the beginning of the upcoming year! This gives studios enough time to showcase and campaign for their Oscar-worthy movies, and most studios clearly bet on their dramas more than any other genre.
- Comedies tend to be released consistently throughout the year, but show a December spike; this probably correlates with Christmas, when studios tend to screen family-friendly movies to take advantage of the holiday season.
- Action movies, when they are released, appear to cluster in the summer months, which is typically seen as blockbuster movie season!
- Horror movies are released at very low frequency, but show a notable spike in October, which, of course, is when Halloween takes place. Notably, there are very few action movies in this time, which makes sense as the demographic for action movies and horror movies tend to overlap significantly (young people).

What is the typical pattern for return on investment in each genre?



Most movies do not actually generate a significant ROI multiplier, as seen by the flattened boxes near the bottom, where most movies fall somewhere in between 0 and 5. Movies with large budgets tend to have low multipliers, which explains why action movies do not have any notable outliers.

The more interesting story comes in looking at the genres in which outliers are practically common: as might be expected, the genres with lowest budgets tend to have the most outstanding outliers. Paranormal Activity was a movie that had a legendary box office performance, in large part because the movie was shot with practically no budget, much like the it's precursor, the Blair Witch Project.

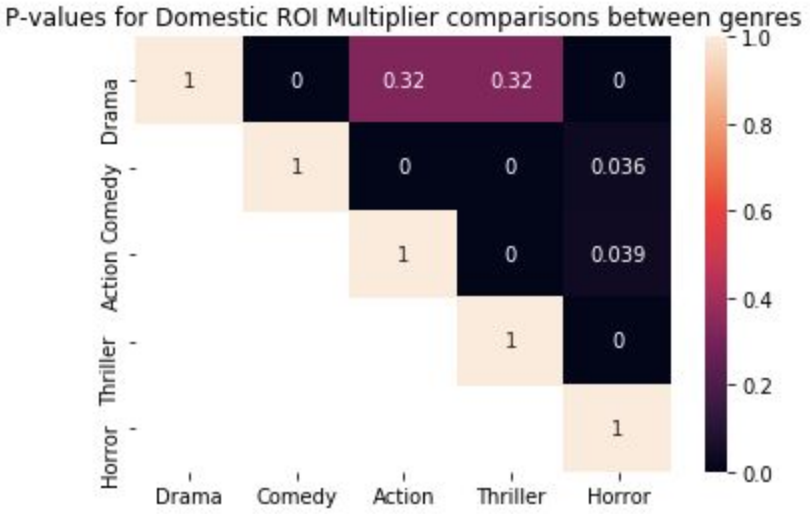
The sheer number of overperforming outliers in Horror suggests an alternative strategy in filmmaking, from a financial perspective. Horror movies can represent a low-risk strategy with a better-than-average chance of breaking out as a box office success. Of course, the absolute profit is never as high as that of a big-budget action movie,which is something else to consider.

Are some genres a safer bet than others?

As a studio, every movie represents a risk, particularly for high-budget features. For smaller studios with limited capital, even lower-budget movies can represent significant risk. From that perspective, it raises the question of what movies are better to invest in, from a studio's risk perspective.

Genre	Budget (millions)	Profit (millions)	Domestic ROI Multiplier
Action	70.00	-0.989502	0.970905
Comedy	30.00	6.868437	1.296492
Drama	22.25	-0.701991	0.943328
Horror	20.00	10.490808	1.498722
Thriller	35.00	-1.841102	0.916314

When looking at the median ROI (where ROI > 1 represents a profitable movie), we find that the two safest genres appear to be Comedy and Horror! If we then perform a bootstrapped comparison of medians to see if any of these differences are statistically significant, we get the following result:



As we can see, there actually is a significant difference in the majority of pairwise comparisons, suggesting that Horror and Comedy are, in fact, the safest genres to invest in.

Which variables significantly correlate with a film's success?

To answer this question, we will make use of the following metrics: a movie's budget, its absolute revenue, its runtime, and critic ratings such as Metascore and Rotten Tomatoes score, as well as audience scores. The correlation between these metrics and the Domestic ROI Multiplier can be assessed using a Pearson test to see if any of these show a significant correlation to the performance of a movie, with an confidence level of $\alpha = 0.05$. Note: As this involves performing multiple tests on the same dataset, a Bonferroni correction will need to be applied to the p-value threshold to account for the possibility of Type 1 error (false positives), altering our target p-value to a more stringent level.

	Correlation	Observed_pval	Permuted_pval	Significant
Budget (millions)	-0.0386211	0.0558048	1	no
Domestic Gross (millions)	0.0109657	0.587234	0.2021	no
Runtime	-0.0439689	0.0408363	0.9966	no
Metascore	0.0430278	0.0464118	0.0209	no
IMDB score	0.00482542	0.822489	0.4348	no
RT score	0.0443767	0.0390011	0.0076	yes

As we can see, most of the metrics do not appear to correlate significantly with the movie's performance, except for its Rotten Tomatoes score! Is this truly the case? In and of itself, it appears as if a studio should only be concerned with the Rotten Tomatoes score, but we also know that these relationships do not exist in a vacuum. As such, the next stage of this project is to develop a model that takes all these features into account. This will help us with the grand purpose of the project: to better inform business decisions when it comes to deciding what areas of a film project needs to be focused on in order to generate maximum revenue.

4. Building a Predictive Model

Data preprocessing

The first step in building a model for predicting ROI is to preprocess the data we have. Here, we converted all categorical variables such as genre or release month to dummy variables via one-hot encoding. This was done using scikit-learn's OneHotEncoder.

The issue facing numeric variables is what to do in the event of missing data; in other words, the strategy for data imputation. In this case, I used SimpleImputer from scikit-learn to impute the median value for numeric variables in this dataset, knowing that there were some outliers which makes using the mean value a less viable strategy.

Data standardization

The need for data standardization arises from the fact that not all numeric variables have the same range; for example, our budget variable ranges from 0.001 to 317, in millions of dollars, while our critic score variables (e.g. Rotten Tomatoes), ranges fro 0-100. This is not a particularly extreme situation, but normalization is a good practice, and so we apply said normalization to allow any model to treat these variables equally. In this case, a StandardScaler was used, which centers the variables on 0 and scales the variance to 1 for each numeric feature.

Model Scoring

One of the more important aspects of choosing a model is to determine a metric by which they can be compared in terms of performance. In regression analyses such as the one we are doing, common options are Root Mean Squared Error or Mean Absolute Error. In many models, RMSE is preferred for mathematical reasons such as smooth differentiation, or when it is desirable to harshly penalize differences between predicted and actual values (the 'squared' component of the metric).

However, MAE is more interpretable because it is simply the absolute average of offsets, which is a much more intuitive concept. Further, it tends to be robust to outliers relative to RMSE. For these two reasons, I went with MAE as the basis of comparison for our models.

Baseline model

For the purposes of having a naive model to compare to, I created a baseline model which, for every movie in our dataset, predicted the ROI to be exactly the median log-transformed ROI of all movies. In this way, any model adopted for the client would have to perform better than the baseline at minimum, which had an MAE of 1.063.

Linear Regression models

Linear models are nice and easy to understand, and as such are always a worthy first step when developing a predictive model. The choice here was between Ridge Regression, Lasso Regression, and Elastic Net models; because Elastic Net generally avoids the pitfalls of Ridge Regression and Lasso Regression while including both of those approaches as special cases, it seemed like a natural choice to use here.

To perform the Elastic Net regression, I applied a grid search cross-validation approach, which sets up a hyperparameter grid along which cross-validation is performed and a performance score calculated. The grid search approach is such that the model with the best score is selected as the best estimator, from which we can then extract the best hyperparameters. In this case we were tuning the L1 ratio as well as the alpha hyperparameters, both of which are regularization parameters. The optimal model, once the grid search was performed, was then fit to the training data and scored against the testing data in order to derive an MAE score.

One important note is that even if the performance of the linear model was not as ideal as the tree-based approaches, linear coefficients provide an indicator of directionality for the features, which is useful when determining actionable insights from the models.

Tree Based models

In this section, I tested two popular ensemble tree-based methods in Random Forests and XGBoost, which is a gradient-boosting approach but with similar underlying principles. In the

case of Random Forests, decision trees are grown from the data with replacement, where a subset of features (here, the square root of the total number of features) is used to generate the tree. This allows the variance to decrease and is an attempt to mitigate the overfitting problem common to tree-based approaches. The result from this approach is the 'forest' of decision trees which make a prediction by aggregation, which is more accurate than if we were to look at any given tree.

XGBoost and Random Forests are similar in being tree-based, and do have similar hyperparameters, such as the depth of the trees, although the XGBoost model is a stepwise learning one, and as such has further hyperparameters dealing with the learning rate. Essentially, in a gradient-boosting approach, the algorithm focuses on data points which were difficult to fit by the previous model, and in the next iteration, tries to get them right. Over time, this leads to a better model which makes better predictions. Although XGBoost is a new and highly-regarded approach, it is not without its issues. It is prone to overfitting even more so than the random forest approach, which already has overfitting issues. As such, there is a need for both to be considered.

Tree-based models allow us to extract feature importances from the models i.e. the most relevant features to our actionable insights. Unlike linear methods, however, there is no way to infer directionality, which is why a combination of both approaches is preferred here.

Model comparisons

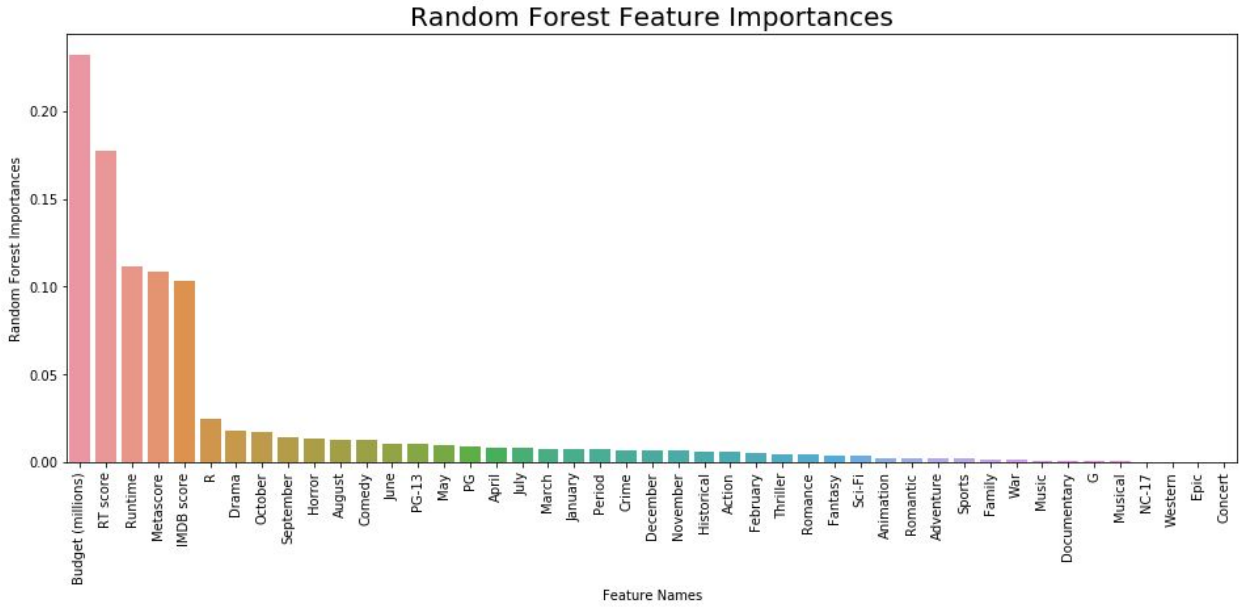
The MAE scores for each of the three major models are shown below, as well as the naive model.

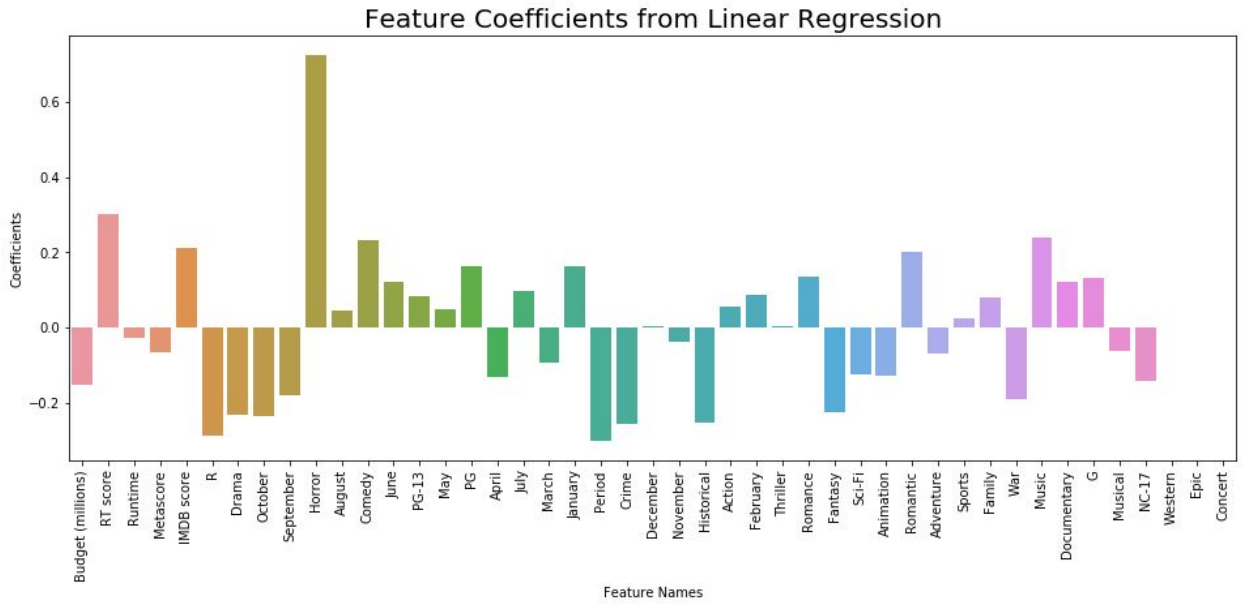
Models	MAE
Naive	0.866205
Elastic Net	0.777221
Random Forest	0.636799
XGBoost	0.657015

As we can see, Random Forest performs best for the data that we have, and as such we will consider the insights from the random forest model moving forward.

Feature Importances and Directionality - All movies

Before presenting the findings, we should note that the ensemble approaches are more accurate than the linear model, which suggests that there may be some non-linear or interaction effects at play. As such, any actionable insights should take into consideration this fact when looking at which features appear to be important.

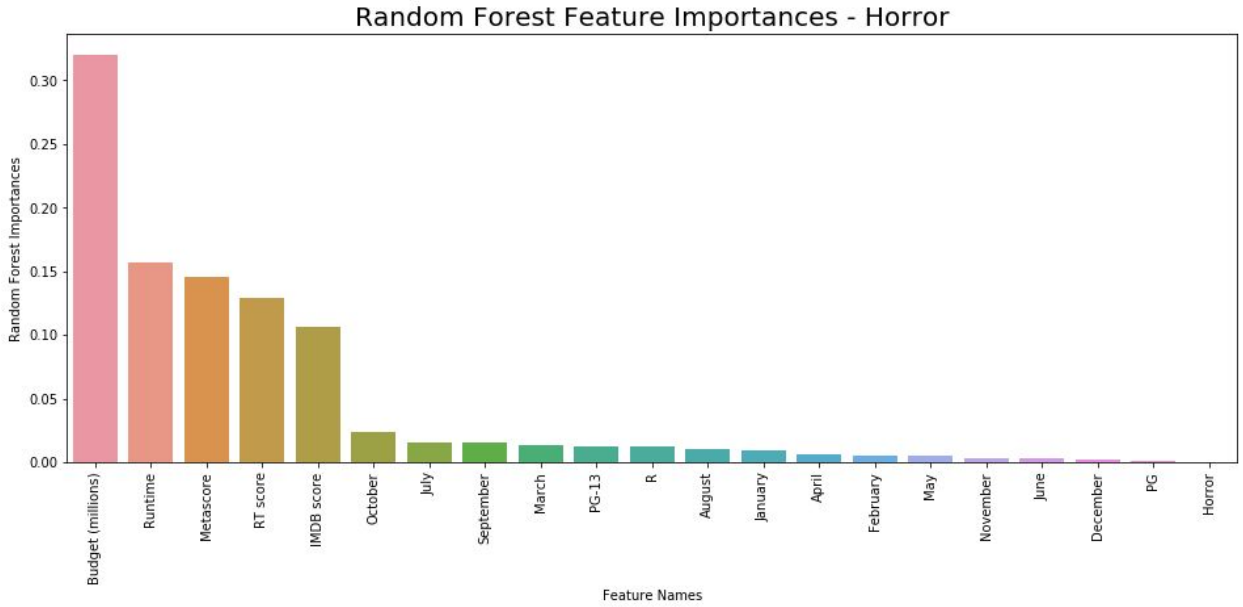


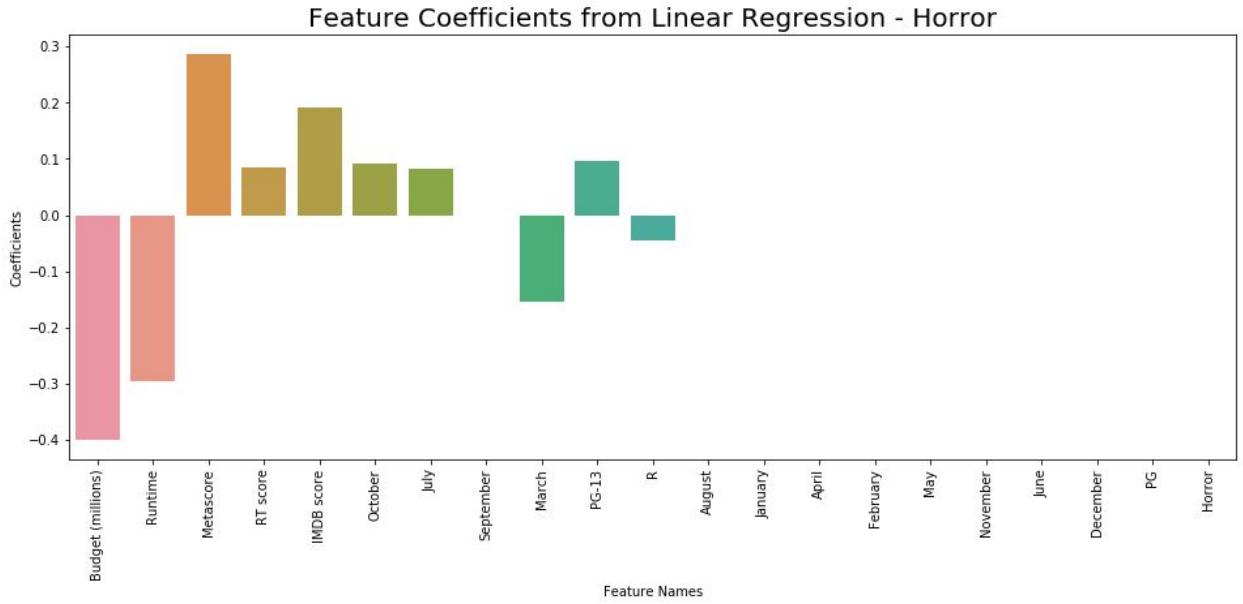


As we can see, the most important features in predicting ROI appear to be the budget, Rotten Tomatoes score, Runtime, Metacritic score, and IMDB score.

Although it is not one of the most important features, it should be noted that horror is the first genre feature which appears, and has a positive coefficient. In the XGBoost model, horror is one of the most important features, alongside others which are similar to the ones in the Random Forests model. Since genre is one of the most important considerations in movie production, it is worth repeating this analysis but with a trimmed down dataset to assess if there are any features specific to horror movies that tell us a different story, particularly which release month is most closely associated with good ROI for horror movies.

For horror movies alone, the results are as follows:





5. Actionable Insights

From the model results, some notable insights are as follows:

1. A low-budget movie is a better predictor of high ROI.

This makes intuitive sense, as a movie with an extremely low budget does not need to make much money in order to generate high ROI in terms of percentage.

2. Shorter movies are associated with high ROI.

This can be explained by the fact that shorter runtimes allow movie theaters to have more screenings of the movie in a given day, allowing for increased box office earnings.

3. High Rotten Tomatoes and IMDB scores are associated with high ROI, but not Metacritic scores.

To understand why this is the case, we need to consider the role that these three metrics serve. IMDB scores are typically given by IMDB users i.e. the general public, and as such are a direct measure of audience favor. Rotten Tomatoes scores are derived from critics reviews; however, the score is based on the percentage of critics who simply would recommend a movie or not i.e. a binary system. Contrast that with Metascore, which assigns scores to a movie based on critics evaluations, in a manner that is more akin to a school grade i.e. it is a more objective measure of film quality rather than the Rotten Tomatoes approach. The difference can be summarized as such: A technically mediocre but enjoyable movie would likely receive high Rotten Tomatoes score, but a low Metascore.

As such, what we can gather from these metrics is that ultimately, audience enjoyment, and not critics enjoyment, is most closely tied to a good ROI.

4. Avoid drama movies, and choose horror movies. When releasing horror movies, release them in the months of October and July.

Despite Hollywood preferentially producing drama movies (see: Data Exploration), our model suggests that this may be not be a wise financial choice. Broadly speaking, drama movies have a tendency to play well to critics, but not necessarily appeal to general audiences for enjoyment. Horror movies are better in that regard, and October is an unsurprising month to choose given the presence of Halloween. However, why the model recommends July is uncertain, which is interesting! Although July is a month typically thought of as being the province of action blockbusters, it is possible that July horror movies appeal to the same younger demographics who constitute significant numbers of moviegoers in that time period, and as such may also be a worthy investment.

All code can be found at: https://github.com/benisonp/Capstone-Project-1---Box-Office