Movie Revenue Prediction Considering Gender Diversity and Star Power

DS-GA 1001 Term Project Report

Team Unsupervised Learners
John Fitzpatrick, Mars Huang, Charlotte Ji, Hengjiali Xu

Business Understanding

Every business strives to predict consumer behavior and deliver what their consumers want. In 2019, the movie industry grossed over $100 billion[1], but the astronomical costs of producing a movie mean that even experienced and well-funded studios can struggle to produce multiple consistently profitable films each year. According to our data, the average major studio produced only 1.8 films per year at an average cost of $40.5 million per film.[2] Through machine learning and data analysis techniques, we hope to identify key factors that predict box office success, and how changes in them affect a movie's revenue.

Recent years have seen the rise of streaming platforms and the demise of traditional movie theaters. In 2019, digital content eclipsed theatrical content[3], and that gap grows wider during the current COVID-19 pandemic. Movie theaters across the world are closing, major studios are sitting on films that they cannot release, and streaming platforms like Hulu and Netflix expand faster and faster. Now, more than ever, being able to select and produce financially successful films could make the difference for a studio between being competitive and becoming a relic of a bygone era.

We want to focus on two initiatives in particular: gender diversity and star power. Since gender inequality in the entertainment industry has come under the spotlight with the advent of #MeToo movement, we are curious about how the level of gender diversity in a movie would affect its financial success. If we could identify financial incentives to increase female representation, it could accelerate the push for equality. Moreover, we research whether the film lead's star power influences audiences' viewing decision. We discuss how we define gender diversity and star power in Data Preparation below.

We believe there are some other important features that might help predict box office and we hope to explore them in three aspects: basic information (e.g. genre); production (e.g. budget); credits and

[1] Theme Report 2019. Motion Picture Association of America, https://www.motionpictures.org/wp-content/uploads/2020/03/MPA-THEME-2019.pdf.

[2] According to our data from 1998-2014. The average number of films per year was based on the leading studio listed in the credits. If we drop studios that only produce one film in the entire period to account for potential shell companies set up by the production houses, the average rises to 2.1 films per year.

[3] Theme Report 2019. Motion Picture Association of America, https://www.motionpictures.org/wp-content/uploads/2020/03/MPA-THEME-2019.pdf.

awards (e.g. director, oscar winner). Using the features identified in our models, a production company can better select which films to produce, and how to produce them. Many decisions made during the production process can yield very large differences in terms of financial success and budget requirements. For example, big names in Hollywood can often receive eight-digit paychecks for their performances in films, so understanding the value proposition they bring may greatly influence a movie's bottom line.

Data Understanding

We primarily leverage the Kaggle Movies Dataset containing information on over 45,000 movies collected from TMDB and Grouplens.[4] We also incorporate a second dataset containing the names of all Oscar winners to build one of our feature variables.[5] The Kaggle dataset was not part of a competition, so we cannot compare our model outputs to others who used it.

To make our movie comparisons more applicable, we limited our focus to English-language films produced in the US/Canada. That limits the differences in market sizes, access to funds, and our metrics for gauging actor star power. Many films were missing most of the attributes and revenue amount so we excluded them from the dataset. We also dropped movies released before 1990, mainly because information on those movies are limited. This enabled us to capture a consistent picture of prominent actors, popular genres, etc, which we expect to be changing over time. That dropped our dataset from 45,429 rows to 4,314 - a large drop, but we believe that we have enough data to reliably run our experiment.

When checking our target variable, we noticed that movie revenues are very skewed to the right. That is to be expected where in one year we can have a few very large blockbuster movies and many films with tepid or cold consumer responses. We transformed revenue into log revenue to achieve an approximately normal distribution, which is shown in Figure 4.

[4] Banik, Rounak. "The Movies Dataset." *Kaggle*, 10 Nov. 2017, www.kaggle.com/rounakbanik/the-movies-dataset.

[5] https://www.openintro.org/data/index.php?data=oscars

When looking at the cast members of the films, we were curious about whether some main actors of a movie contribute to its revenue. We came up with an idea of creating a feature reflecting star power. Based on the sampling of films we looked at, the cast list seems to order the cast members in terms of importance in the film. We restricted the list to the first five actors, and by looking at some specific actors' average log revenue of the movies they starred in the past five years' running window, we found some interesting discrepancies between popular and unpopular actors.

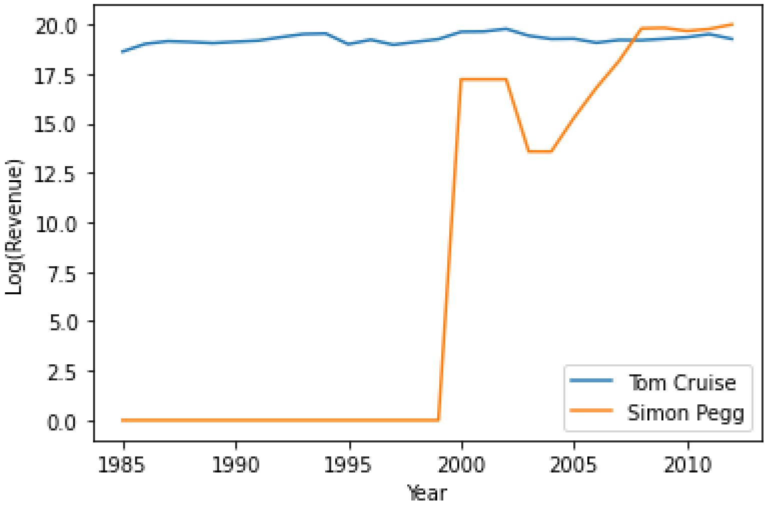


Figure 1: Average log revenue of starred movies in a 5-year running window

This figure shows that Tom Cruise had a quite stable running log average revenue, in which Simon Pegg had a major drop in around 2007. To capture the variations in average log revenue over the years, we built our feature for star power, which will be discussed in more detail in the following section.

We checked if our revenue data contains trends outside of the scope of our predictive analysis that we would need to correct for such as inflation. The plot below of log revenue against release year on our training dataset shows no significant upward or downward changes in average and median log revenue over the time window.

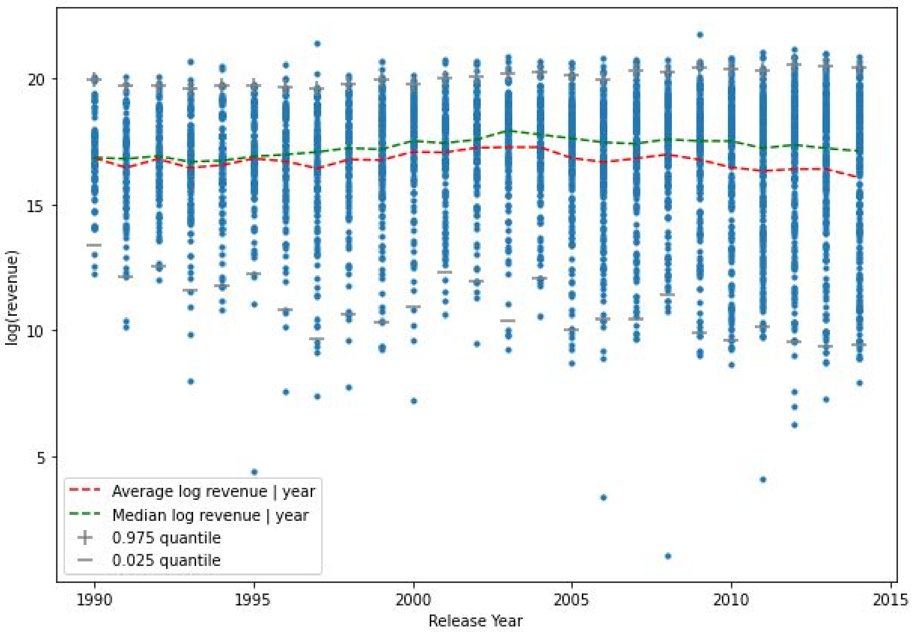


Figure 2: Scatter plot of log revenue against release year on the training dataset

We also looked at the distribution of movie genres. We believed that the genre of a film can have a significant impact on the financial performance and be correlated with other features in our data. We also had the viewer ratings of the films, and looked at which genres appeared to be fan favorites. The information we gleaned from that analysis became the basis for our popular genre variable which we discuss in the Feature Variables section below.

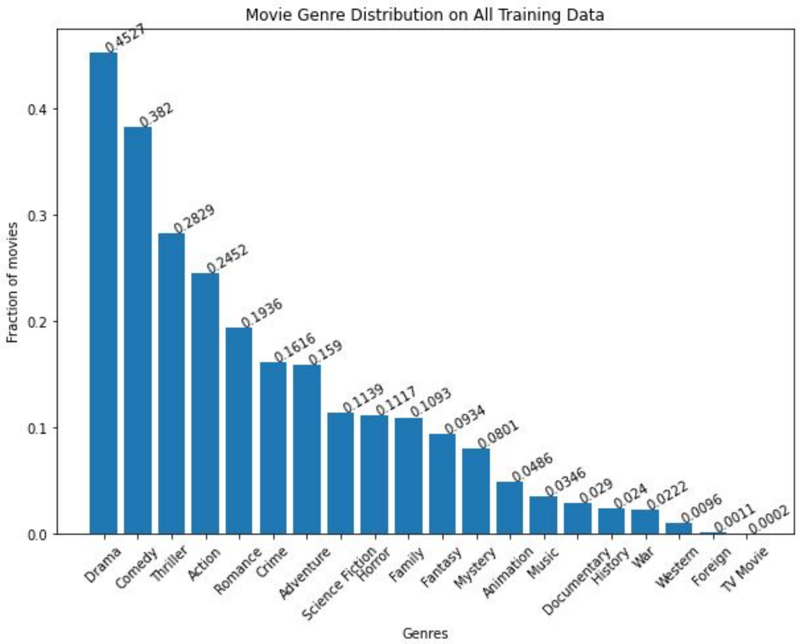


Figure 3: Movie genre distribution on the entire training dataset (train + val)

Data Preparation

The Kaggle Movie Dataset is broken up across multiple CSV files that we combined based on the movie ids. There were instances of the same movie being entered multiple times, so we dropped any duplicates in the process. The Oscar winner information was incorporated based on the names of the top five actors in each film.[6]

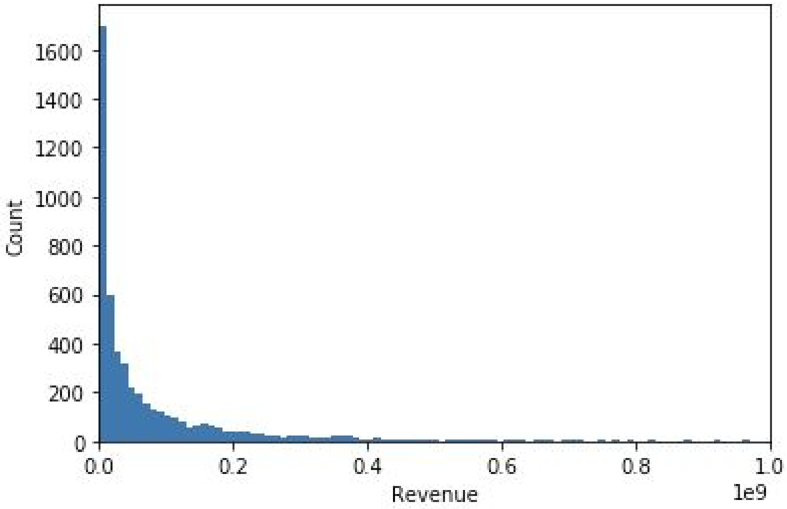
A lot of data preparation went into the project, beginning with filtering based on the film's language and country of origin. We then broke the data into three sets based on the year the film was released. That was done in order to capture features regarding star power and to more accurately reflect the real-life situation of determining what films will succeed financially next year based on prior years' information.

- **Data Sampling:** We split our dataset into training, validation and test based on the movies' release years. Since we framed our goal as a time series forecasting problem, we refrained from

[6] The overlap of professional actors' names is extremely low due to requirements to join the Screen Actors Guild (the union representing most actors in film) that tries to have unique stage names for each of its members. More information can be found at https://www.backstage.com/magazine/article/need-know-signing-sag-54530/.

doing random sampling on individual movies. Splitting data along the time dimension also allowed us to engineer features using historical revenue data without causing leakage. Our training dataset consists of movies released between 1990 and 2012 (inclusive), validation dataset between 2013 and 2014, and test dataset between 2015 and 2017. We used training and validation for hyperparameter tuning and model selection, and combined these two datasets as a larger training dataset for our selected models. We refer to the training dataset as "sub-training" and the combined dataset as "training" in the following sections.

- **Target variable**: log of revenue.



By taking the log of revenue, we can achieve a distribution much closer to Normal.

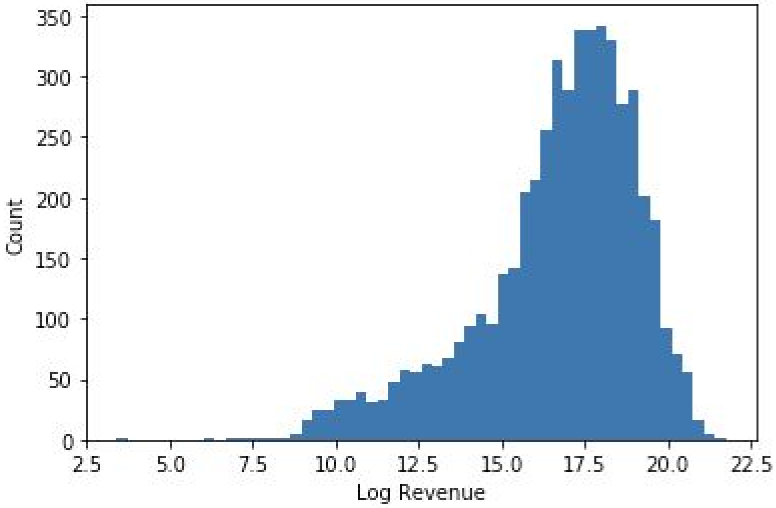


Figure 4 (top, bottom): Histogram of revenue before and after taking the logarithm

Predicting log revenue has benefits in that it can define different levels of misprediction losses for different magnitudes of true revenue. For example, for revenue of one-billion dollars, making a one-hundred-million-dollar wrong prediction is mild and acceptable. By contrast, when the true revenue is only one-hundred million, the same scale of misprediction is tremendous, which accounts for a one-hundred-percent error rate. With the log revenue prediction, MSE loss is more sensitive for smaller true revenues but less sensitive for large ones, so it can guide the model in the right direction.

- **Feature variables**:

  Key features:

  - Gender score: To see if the gender distribution of the cast affects the performance of the film, we created a gender score based on the proportion of the top five actors in each film that were female. For some actors, no gender was listed Based on our sampling of such instances, we decided to count the actor as female only if they were positively identified as female in the dataset.
  - Top 5 actor's average log revenues in the past 2, 5, 10 years: To trace and estimate the trends of the top 5 stars' earning power in each movie, we calculated the average/median log revenue of their starred movies in the past 2, 5, and 10 years from sub-training/training data. For movies before 2000, we used pre-1990 data to engineer these features[7]. For the movies containing less than 5 top actors, we considered they missed some actors. To handle these missing actors, we padded <unk> tokens as their names to ensure the length of the lists to be 5 and estimated their log revenues by <unk>'s. For validation/test data, we cannot compute these features directly from these datasets. If we calculate the top 5 actors' log revenues in the past few years, we would

[7] More specifically, we used 1980 to 1990 movie data to generate these features. Some of our other features discussed in the Appendix also utilize this set of data.

inadvertently access the true revenues in the validation/test set, which can introduce information leakage. Therefore, we had to instead estimate them from sub-training/training data. We inferred actors' log revenues by taking their latest log revenues of their annual log revenues in the sub-training/training data. For instance, an actor's log average revenue in the time range from 2011 to 2012 would be estimated as it in the time range from 2010 to 2011.

Production:

○ Log budget: To normalize the skewed distribution, we took the logarithm of budget as well.

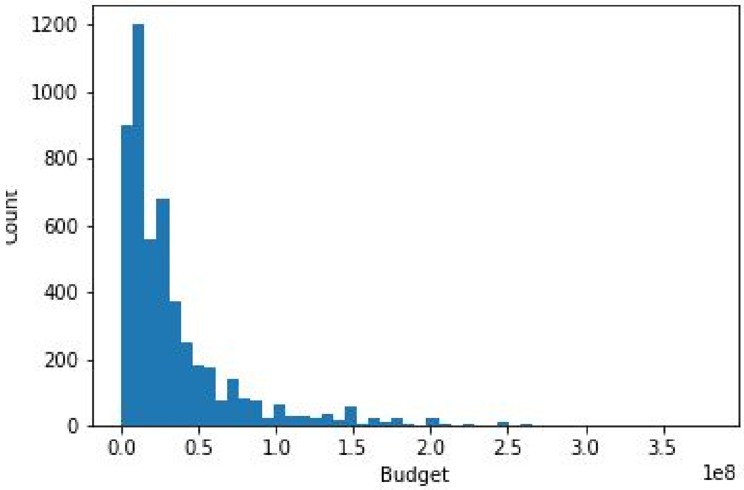


Figure 5: Histogram of budget

○ Production company class: To check the relationship between movie revenue and production company size, we created a categorical feature for production company. We only considered the first production company listed in the data as its main production company. We divided the production company into four groups: "major"- companies that produced more than 100 movies in dataset; "medium"- companies that produced 10-100 movies; "minor"- companies that produced < 10 movies; "other"- for those movies which do not specify any production companies. To avoid leakage, we performed this labeling process on the sub-training set and training set, and estimated the class of production company in the validation and test set according to that production company's class in the sub-training and training set respectively. I.e. If company A was classified as "major" in the sub-training set, it was also labeled as "major" in the validation set; if a company did not appear in the sub-training set but appeared in the validation set, it was labeled as "other".

We listed all other features we extracted in Appendix 1.

- **Feature Selection:**

  As a first step, we calculated mutual information between each feature and the target variable on the sub-training dataset.

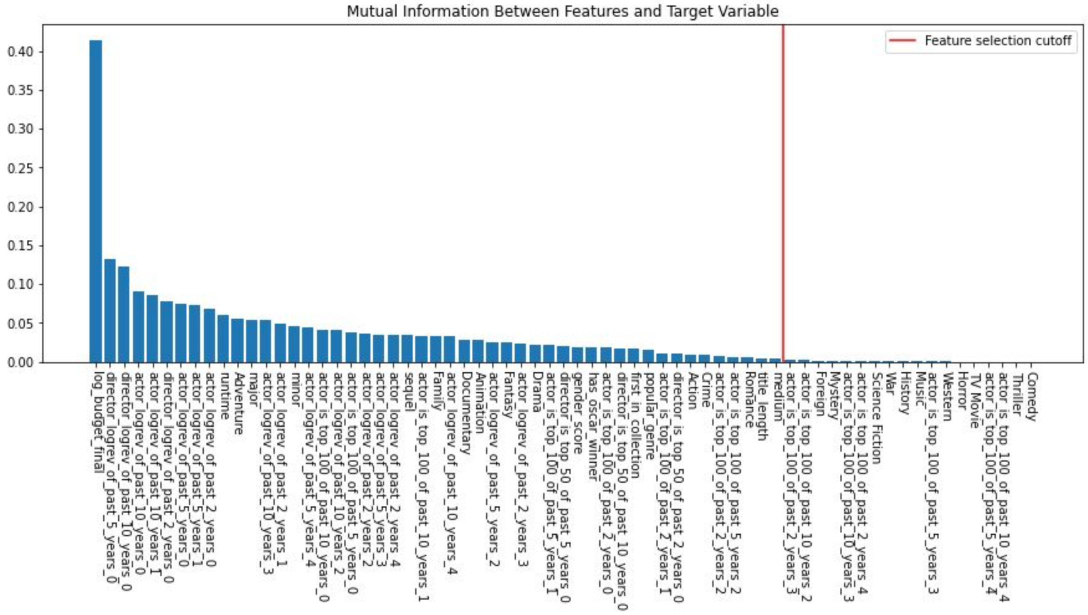


Figure 6: Mutual information between features and log revenue

We can see several genre and actor hotness features do not contribute much to the explaining variations in log revenue. We decided to drop all features after the feature selection cutoff shown in Figure 6 for all of our models to reduce noise and avoid overfitting.

On the other hand, to identify multicollinearity in our features, we plotted a correlation matrix on all variables that we believe would contribute to our models using the sub-training dataset.

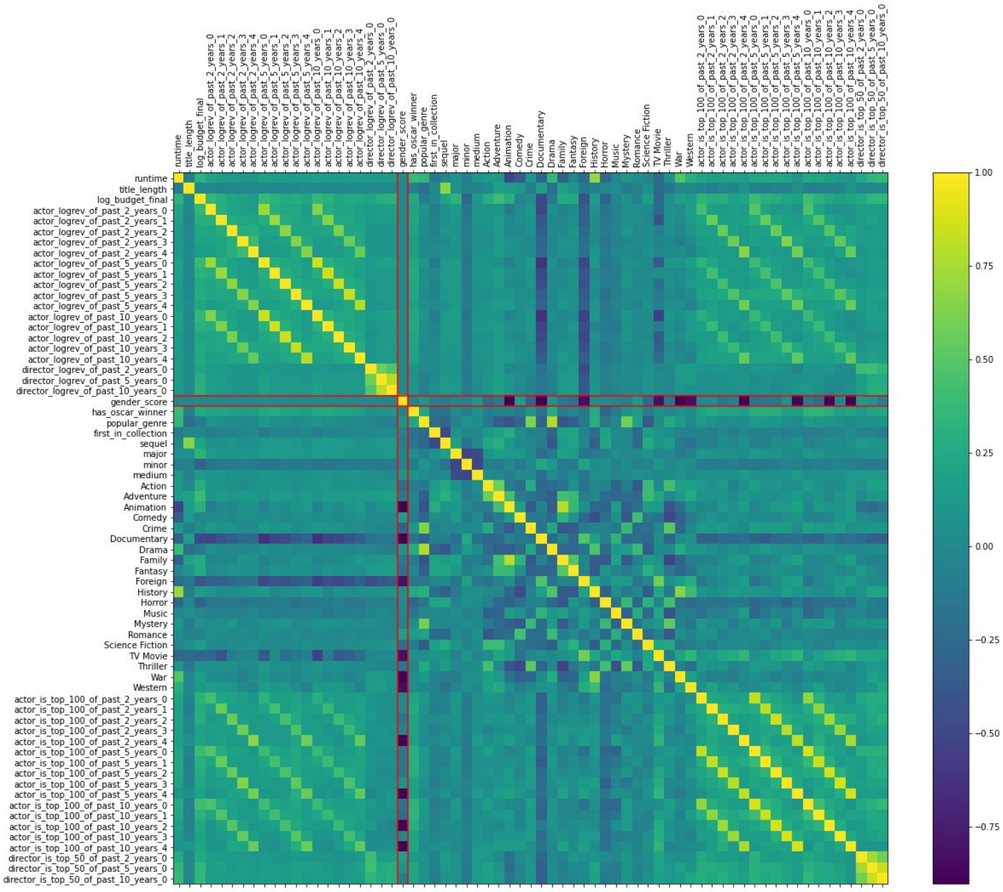


Figure 7: Mixed correlation matrix of all features

Since these variables have mixed data types, including continuous (e.g. log_budget), ordinal (gender_score) and dichotomous (e.g. genre variables) variables, using a single correlation measure would not be appropriate - for example, measures that are appropriate for estimating correlation between continuous variables, such as pearson or spearman rank-correlation, might fail to capture relationships between a continuous and a dichotomous variable. Point-biserial correlation would be more appropriate in the latter case[8]. Therefore, we

[8] Khamis, Harry. “Measures of Association: How to Choose?” Journal of Diagnostic Medical Sonography 24, no. 3 (May 2008): 155–62. https://doi.org/10.1177/8756479308317006.

generated a mixed correlation matrix according to data types in Figure 7. Variables before gender_score are all continuous, gender_score is the only ordinal variable with 6 levels, and the rest are dichotomous. Correspondingly, we plotted results pearson correlation for continuous-continuous, tetrachoric/polychoric correlation for dichotomous-dichotomous and ordinal-dichotomous, biserial correlation for continuous-dichotomous, and polyserial for continuous-ordinal.

This figure contributed to our understanding of the data by revealing connections between variables that we were unable to discover using only Pearson correlation. For example, we see various degrees of correlations between genres, as some genre categories often appear together in a movie. It's worth noting that some of these correlations need to be taken with a grain of salt, because as we have seen before, some of the genre categories are represented by incredibly small amounts of movies (e.g. TV Movies). We also see strong negative correlation between the gender score and certain genre categories, such as War and Western, which intuitively makes sense as female characters are indeed less common in these categories. From the corners of this map, where cast information is presented, we see positive correlations exist between actors' and directors' log revenues in the past 2, 5 and 10 years respectively, which indicate the existence of a trend in an actor's or director's average log revenue within the 10-year window.

Modeling & Evaluation

- **Choices of Data mining algorithm:**

  To preserve interpretability in our results considering our goals, we selected linear models as our first candidates. However, given the non-trivial amount of issues that can arise with linear models other than multicollinearity (e.g. the homoscedasticity assumption might be violated), we also tested on tree-based models to increase prediction accuracy, even though we lose interpretability as a result. Our final list of models are: Linear Regression (Ordinary Least Squares and Ridge Regression), Support Vector Regression, Decision Tree Regression, Random Forest Regression, and Gradient Boosted Decision Tree.

- **Evaluation metrics:**

  Since we have a regression problem, we used mean squared error (MSE) and the coefficient of determination (R^2) as our metrics.

- **Baseline model:**

  Since our goal is to both predict movie revenue as accurately as possible and study how gender diversity and star power affect revenue, we built two baseline models. For each of them, we built the model both on the sub-training set and training set, and evaluated it on validation set and test set respectively. In the first baseline model, we only used log budget and production company class to predict log revenue. Ideally speaking, a movie with a high budget should also receive high revenue. Large production companies tend to cost more for making a movie and their reputation might help a movie's financial success. Therefore, we built a linear regression model on log revenue with log budget and production company class. We got a MSE of 5.31 on the validation set and a MSE of 4.67 on the test set.

  In the second baseline model, we used gender score and star power to predict log revenue of movies. Again, we built a linear regression model on log revenue with gender score and the average revenue of movies that the top 5 main actors of the movie starred in the last 2, 5, and 10 years. As a result. we got a MSE of 8.57 on the validation set and a MSE of 6.99 on the test set. We found that both of the MSEs on the second baseline model are larger than the ones on the first baseline model.

  These mean squared errors showed that our baseline models, which used only two features, were quite naive and we needed to add more potential features to improve upon it.

- **Modeling:**
  - Ordinary Least Squares

Because OLS is susceptible to multicollinearity, we dropped correlated features combining evidence from the correlation matrix and variance inflation factor (VIF) of each feature. By convention, a VIF over 5 usually indicates multicollinearity. See Appendix 1 for more details. On top of dropped features using mutual information, we dropped popular genres and built three different OLS models using actor/director log revenue from on a single time frame (2,5, and 10 years) in each. These actor/director variables showed high collinearity across time frames, and we wanted to see which time frame produces the best model. The results of each model trained on the sub-training dataset and tested on the validation dataset are reported in Table 1.

	Time Frame of Top 5 Actor / Top 2 Director Average Log Revenue Features		
	Past 5 Years	**Past 10 Years**	**Past 2 Years**
MSE on val	4.3444	4.3507	4.3632
R^2 **on val**	0.5650	0.5644	0.5631

Table 1: Best parameters of OLS models with selected features including actor/director average log revenue features in different time frames

Including the 5 years features outperformed 2 years and 10 years, because a 5-year window contains enough data for the feature to capture the trend in actors' and directors' past box office performance. 2 years are too short for capturing such a trend, while we often see dramatic changes in the trend over 10 years, such as what we see in Simon Pegg's career in Figure 1, which might make the estimates of log revenue inaccurate.

  - Ridge Regression

We also performed ridge regression, which we expect to address the multicollinearity in our features directly through the L2 regularization term. To have a working regularization term with low bias, we performed grid search to choose 0.01 as our final alpha. We used all features selected using mutual information without further dropping correlated features. This model achieved 4.5354 in MSE and 0.5459 in R^2 , which is not ideal compared to the OLS results. We suspect that the collinear features introduced additional noise to our model, as we achieved better results using the same set of features used in OLS with 5 years actor/director average log revenue variables. The results are reported in Table 2. These results are similar to OLS results, which is expected given the similarity between the two models.

One possible limitation of both linear models is whether they capture variance in our data appropriately. OLS requires our target variable to be homoscedastic, which might not be the case. We explore more on this topic in Appendix 5.

○ Decision Tree Regressor

We used Decision Tree Regressor for one of our tree-based models since its algorithm is very easy to understand. However, it has the downside of instability and easily overfitting the data. We therefore performed grid search to tune three hyperparameters: min_samples_leaf, min_samples_split and max_depth.

○ Support Vector Regression

We used Support Vector Regression for improving our baseline models for it is very effective in higher dimensional spaces. We used the default “rbf” kernel and tuned C and gamma using grid search.

○ Random Forest Regression

For our random forest model, we only tuned n_estimators and min_samples_leaf while keeping other parameters default due to both computational limitations and that we don't want to limit the size of our model - random forest hardly overfits and usually performs better when it is large in size. Our best parameters and the best model's performance is reported in Table 2.

- Gradient Boosted Regression Tree

  For the gradient boosted regression tree model, it is easier to overfit due to its high variance. To control the model complexity, we set the minimum leaf size at least 16 and train it with a larger number of trees and smaller learning rates. The best selected hyper-parameters and the performance is shown in Table 2.

- **Results:**

Models	**Hyperparameters**	**MSE**	R^2
Ridge Regression	alpha: 0.01	4.3446	0.5660
Ordinary Least Squares	None	4.3444	0.5650
Decision Tree Regressor	min_samples_leaf: 16 min_samples_split: 128 max_depth: 6	4.005	0.599
Support Vector Regression	C: 600 gamma: 0.001	4.19	0.58
Random Forest Regression	n_estimators: 1000 min_samples_leaf: 4 max_depth: None	4.0865	0.5908
Gradient Boosted Regression Tree	learning_rate: 0.01 n_estimators: 1000 min_samples_leaf: 16 min_samples_split: 32 max_depth: None	**3.7476**	**0.6248**

Table 2: Best parameters of each model and the corresponding MSE and R^2

Our result suggests that the gradient boosted regression tree is the optimal model for log revenue prediction. Using the chosen hyperparameters for that optimal model, we trained the model again on the training dataset (sub-training + validation) and achieved a MSE of 2.9946 and a R^2 of 0.6117 on the test dataset. This means that our model could explain about 61.17% of the variance of log revenue in the test set, which gives us confidence that our model could make fairly good predictions on movie revenue.

Due to lack of interpretability of gradient boosted regression trees, we referred to the OLS model to examine the effect of gender and star power on revenue prediction. We trained the OLS model on the training dataset using selected features. We achieved 3.9503 in MSE and 0.48783 in R^2**.** Our regression results are reported in the Appendix. The coefficient for gender score is 0.0476, which suggests a slightly positive relationship between casting more female actors and log revenue of a movie - adding one more female actor to the main cast is correlated with an increase in movie log revenue by 4.7%. However, the result needs to be taken with a grain of salt as the coefficient's p-value is 0.128, indicating the high possibility that we got this positive coefficient by chance. Still, the positive relationship between gender diversity and log revenue in our dataset is a good starting point, and we could redo the modeling to see how this coefficient changes in value and significance if we gather more historical data in the future.

On the other hand, we did get significant results for some of the actor features that we created. The coefficient for the main actor's average log revenue of the past 5 years is 0.2473, which suggests that after we estimate the trend in average log revenue of an actor using historical data, hiring those with 1% higher estimated average log revenue is associated with a lift in log revenue by 0.24%. We see similar levels of positive coefficients for the average log revenue variable of each of the 5 leading actors, despite one of them being insignificant.

Deployment

Our model is not designed to be integrated into a pipeline and run automatically - the decision of whether to produce a movie in a given year and how is one that costs tens of millions or even hundreds of millions of dollars, and requires careful planning and consideration. Our model would complement that decision making process by helping executives understand what factors affect a film's financial success. Because so few films are produced each year by each production company, they would have a lot of flexibility about when and how often the model should be run. Ideally, it should run to provide updated information before they greenlight a project and when making high-level decisions about budget allocation. Given the considerations below, the data and outputs should be considered and interpreted in the larger context of changing demographics and trends.

The model should be monitored and evaluated often. Consumer interests change quickly and movie stars continuously move in and out of vogue. Just having a top 100 actor might not be enough in our current culture where one tweet or one hack could lead to a very popular actor becoming radioactive overnight.

It is also important to keep in mind which consumer and what market they are targeting. Our data is a little dated by today's standards and doesn't incorporate much of the recent explosion in digital content. What makes a film successful on Netflix could be very different than what makes it successful in a theater. But to apply our methodology with a restricted time window might be a mistake; limiting the time window also limits the number of datapoints they have to draw from. And even with a shortened window with enough data points, cultural shifts that would pop up next year might not be reflected in prior data. Instead, alternative data sources like social media could provide better insight into what consumers would like to see.

In terms of ethical considerations, this model could directly affect hiring decisions, so care should be taken to ensure that is not abused. For example, even though our gender score variable is not significantly correlated with revenue, if in the future that changes and the model indicates that gender diversity negatively impacts financial performance, that could disincentivize production studios from hiring women in their films or choosing scripts that would feature female protagonists. This could lead to more niche or independent films being the only ones that prominently feature female leads, lowering the revenue feature even further since they would not be targeted at the larger population.

Another ethical consideration arises when we try to quantify star power. Because if a lead actor's star power is a strong indicator of a film's success, that might incentivize production companies to rely on existing movie stars to take lead roles in their films. But, because of historical and existing biases, the actors who are most often included in films and who might garner the most star power are overwhelmingly white and male. That could prevent new actors from joining the industry and inhibit future diversity.

Appendix 1: Contributions

- **John Fitzpatrick**
    - Assisted with planning and scheduling
    - Created gender_score and has_oscar_winner feature variables
    - Created some of the graphs and EDA
    - Consulted on model selection process
    - Wrote much of the first draft of the report and helped with editing

- **Mars Wei-Lun Huang**
    - Created is_topk_director/actor, directors' and actors' log(revenue) feature variables.
    - Created graph of the trend of actors' log(revenue) and EDA
    - Trained and Tuned Hyper-parameters for Gradient Boosting Regression Tree
    - Added explanation for target variable
    - Assisted in model explanations

- **Hengjiali Xu**
    - Created log budget, log revenue, production company class, title length variables.
    - Created graph of log budget and EDA
    - Created baseline models
    - Trained and tuned decision tree regressor and SVR
    - Wrote some parts of the write-up and assisted in write-up editing

- **Charlotte Ji**
    - Contributed to data cleaning and understanding, plotted revenue over years
    - Created genre and collection variables
    - Generated correlation matrix, mutual information and VIFs and their plots and tables

- ○ Built and tuned OLS, ridge regression and random forest models
- ○ Interpreted model results and wrote respective parts of the write-up

Appendix 2: Feature engineering on other variables

Basic information:

- Sequel: The dataset included information about whether the film was part of a movie series. Because sequels are only created if the first film was a financial success and the subsequent films' successes are in large part due to the success of the entire franchise, we wanted to capture that information. We created separate binary variables for films that are first in a series and films that are sequels.
- Title length: To see whether the length of movie title affects the movie's revenue, we counted the number of words in the title and created the feature 'title length'. Since larger length may convey more information about the movie itself, we want to explore how it may affect people's decision on whether to watch the movie in theatre. We also excluded stop words such as 'the', 'of' etc, because they do not contain much information.
- Runtime: This feature is in our original dataset. It is in the unit of minutes. Longer run times would be correlated with higher budgets and could represent a niche market if the runtime conveys information about the movie's content. A possible scenario is films with longer runtime are sometimes based on written work, which cannot be shortened without omitting key aspects of that story, and these films often carry with it a specific fan base.
- Genre: In our original dataset, each movie is categorized into one or multiple genres. There are 20 distinct movie genres in total (see previous figure on movie genre distribution). For each genre, we generated a binary variable indicating whether a movie belongs to that genre.
- Popular genre: On top of our dummy genre variables, we created a binary variable 'popular genre' to indicate whether a movie belongs to any of the popular genres. For the time periods of sub-training and training datasets, we determined popular genres respectively by the following procedure: first, we calculated the average IMDb user ratings for each genre, and the percentage of movies in each genre category. Then, we sorted genres first by average user rating and second

by movie percentage. In the end, we chose the median of the average ratings and percentage = 0.05 as our cutoffs for being popular, that is, only when a genre has a high enough average rating and at least 5% of the movies belonging to its category, we consider it a popular genre. We ended up with Drama, Crime and Mystery as our popular genres for both time periods (sub-training and training).

Credits and Awards:

- Oscar winner: We created a binary variable if the film has an Oscar winner in the top five members of the cast. This is where we combine the Kaggle Movie Dataset with the Oscar winner dataset. To prevent leakage, we restricted the time periods to actors that would have already won the award before the year the movie was released to prevent knowledge of their future win leaking backwards.
- Hotness of the first 5 actors in the past 2, 5, 10 years (is_top100_actor): We calculated and ranked the number of filmed movies for the first 5 actors of all movies in the past 2, 5, and 10 years. Then we defined they are hot if their number of productions are in the top 100 actors of the past years.
- Hotness of the first director in the past 2, 5, 10 years (is_top50_director): We computed and ranked the number of directed movies for the first director of all movies in the past 2, 5, and 10 years. Then we defined they are hot if their number of productions are in the top 50 director lists of the past years.
- Director log average/median revenue in the past 2, 5, 10 years: Similar to top 5 actors, we would like to understand how the earning power of the top 1 director in a movie affects its log revenue. We computed the average/median revenue of their directed movies in the past 2, 5, and 10 years from sub-training/training data. For the movies without director lists, we replaced their missing log revenue with all directors' median. Likewise, for validation/test data, to estimate directors'

log revenue, we took their log revenue in the last year, and the median/linear extrapolation of their log revenue in the sub-training/training dataset.

Appendix 3: Variance inflation factor greater than 5 calculated using all features

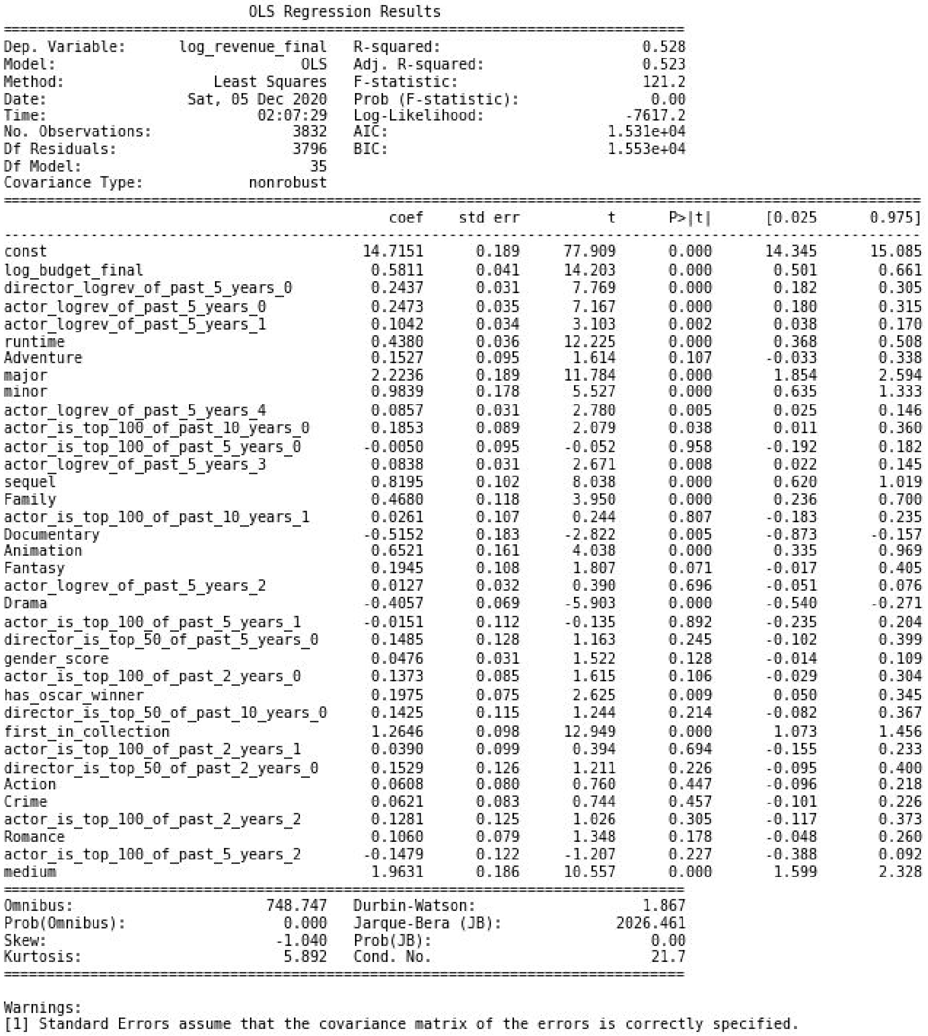
feature	VIF
major	8.603999011
minor	8.529933524
medium	8.100122785
actor_logrev_of_past_5_years_0	7.583381983
actor_logrev_of_past_10_years_0	6.582281107
actor_logrev_of_past_5_years_1	6.415715639
actor_logrev_of_past_5_years_2	5.606879152
actor_logrev_of_past_5_years_3	5.433655426
actor_logrev_of_past_10_years_1	5.321512306
popular_genre	5.270249561
actor_logrev_of_past_5_years_4	5.044906499

VIF for a feature X_i quantifies multicollinearity:

$$VIF_i = \frac{1}{1-R_i^2}$$

where R_i^2 is the coefficient of determination in the regression of X_i on all other features. The results on major, minor, medium are less interpretable since they are dummy variables created for a categorical variable. Considering the evidence from the mutual information ranking which indicated their importance, we decided to keep them in our final OLS models.

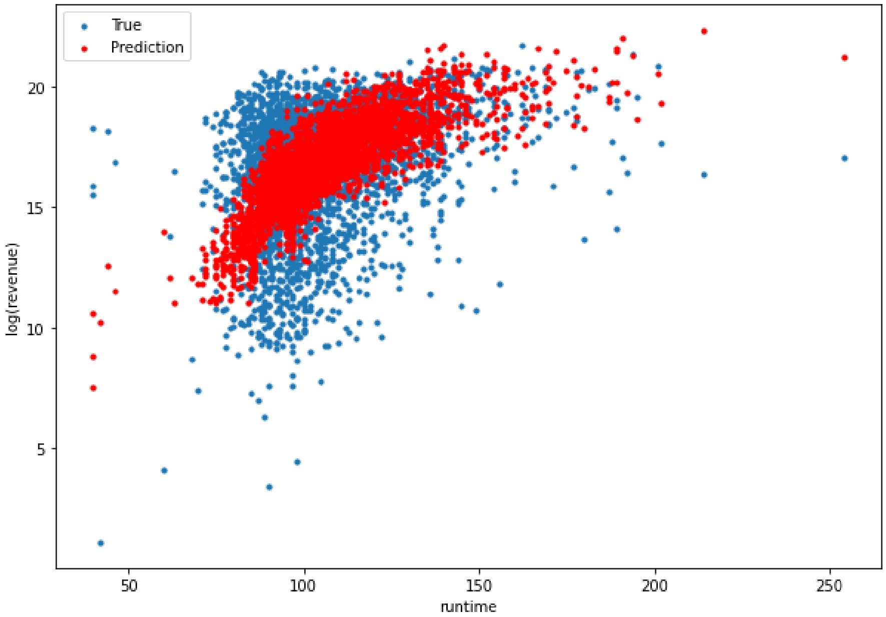
Appendix 4: OLS Regression Results using statsmodels



```
                            OLS Regression Results
==============================================================================
Dep. Variable:      log_revenue_final   R-squared:                       0.528
Model:                            OLS   Adj. R-squared:                  0.523
Method:                 Least Squares   F-statistic:                     121.2
Date:                Sat, 05 Dec 2020   Prob (F-statistic):               0.00
Time:                        02:07:29   Log-Likelihood:                -7617.2
No. Observations:                3832   AIC:                         1.531e+04
Df Residuals:                    3796   BIC:                         1.553e+04
Df Model:                          35
Covariance Type:            nonrobust
==========================================================================================================
                                              coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------
const                                      14.7151      0.189     77.909      0.000      14.345      15.085
log_budget_final                            0.5811      0.041     14.203      0.000       0.501       0.661
director_logrev_of_past_5_years_0           0.2437      0.031      7.769      0.000       0.182       0.305
actor_logrev_of_past_5_years_0              0.2473      0.035      7.167      0.000       0.180       0.315
actor_logrev_of_past_5_years_1              0.1042      0.034      3.103      0.002       0.038       0.170
runtime                                     0.4380      0.036     12.225      0.000       0.368       0.508
Adventure                                   0.1527      0.095      1.614      0.107      -0.033       0.338
major                                       2.2236      0.189     11.784      0.000       1.854       2.594
minor                                       0.9839      0.178      5.527      0.000       0.635       1.333
actor_logrev_of_past_5_years_4              0.0857      0.031      2.780      0.005       0.025       0.146
actor_is_top_100_of_past_10_years_0         0.1853      0.089      2.079      0.038       0.011       0.360
actor_is_top_100_of_past_5_years_0         -0.0050      0.095     -0.052      0.958      -0.192       0.182
actor_logrev_of_past_5_years_3              0.0838      0.031      2.671      0.008       0.022       0.145
sequel                                      0.8195      0.102      8.038      0.000       0.620       1.019
Family                                      0.4680      0.118      3.950      0.000       0.236       0.700
actor_is_top_100_of_past_10_years_1         0.0261      0.107      0.244      0.807      -0.183       0.235
Documentary                                -0.5152      0.183     -2.822      0.005      -0.873      -0.157
Animation                                   0.6521      0.161      4.038      0.000       0.335       0.969
Fantasy                                     0.1945      0.108      1.807      0.071      -0.017       0.405
actor_logrev_of_past_5_years_2              0.0127      0.032      0.390      0.696      -0.051       0.076
Drama                                      -0.4057      0.069     -5.903      0.000      -0.540      -0.271
actor_is_top_100_of_past_5_years_1         -0.0151      0.112     -0.135      0.892      -0.235       0.204
director_is_top_50_of_past_5_years_0        0.1485      0.128      1.163      0.245      -0.102       0.399
gender_score                                0.0476      0.031      1.522      0.128      -0.014       0.109
actor_is_top_100_of_past_2_years_0          0.1373      0.085      1.615      0.106      -0.029       0.304
has_oscar_winner                            0.1975      0.075      2.625      0.009       0.050       0.345
director_is_top_50_of_past_10_years_0       0.1425      0.115      1.244      0.214      -0.082       0.367
first_in_collection                         1.2646      0.098     12.949      0.000       1.073       1.456
actor_is_top_100_of_past_2_years_1          0.0390      0.099      0.394      0.694      -0.155       0.233
director_is_top_50_of_past_2_years_0        0.1529      0.126      1.211      0.226      -0.095       0.400
Action                                      0.0608      0.080      0.760      0.447      -0.096       0.218
Crime                                       0.0621      0.083      0.744      0.457      -0.101       0.226
actor_is_top_100_of_past_2_years_2          0.1281      0.125      1.026      0.305      -0.117       0.373
Romance                                     0.1060      0.079      1.348      0.178      -0.048       0.260
actor_is_top_100_of_past_5_years_2         -0.1479      0.122     -1.207      0.227      -0.388       0.092
medium                                      1.9631      0.186     10.557      0.000       1.599       2.328
==============================================================================
Omnibus:                      748.747   Durbin-Watson:                   1.867
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2026.461
Skew:                          -1.040   Prob(JB):                         0.00
Kurtosis:                       5.892   Cond. No.                         21.7
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

<u>Appendix 5: Heteroscedasticity in target variable</u>



Log revenue conditioned on certain features such as runtime does not have the same variations over different values. We plotted our prediction using ridge regression over the original sub-training data and observed that the model doesn't capture variations in log revenue over the lower range of runtime.

Appendix 6: all of our codes and data are in this Google Drive folder:
https://drive.google.com/drive/folders/1tcZKaokUFASKSBmcWpVUh8OX_dEt9lzG?usp=sharing

Bibliography:

Theme Report 2019. Motion Picture Association of America, https://www.motionpictures.org/wp-content/uploads/2020/03/MPA-THEME-2019.pdf

Banik, Rounak. "The Movies Dataset." Kaggle, 10 Nov. 2017, www.kaggle.com/rounakbanik/the-movies-dataset.

"Oscar Winners, 1929 to 2018." Data Sets, www.openintro.org/data/index.php?data=oscars.

Khamis, Harry. "Measures of Association: How to Choose?" Journal of Diagnostic Medical Sonography 24, no. 3 (May 2008): 155–62. https://doi.org/10.1177/8756479308317006.