Gender disparity in Hollywood (1980-2020)

Nguyen Nguyen

November 8, 2021

**Abstract**

“You play ball like a girl!”, a famous line from The Sandlot movie in 1993, “I don’t play sports with women” from the movie I Love You Man (2009) with many others stereotyped and sexist screenplay does have an effect on audiences’ real-world gender-based attitudes, beliefs, and behaviors. The media has the power to educate and shape people’s perceptions of the world. Images and stories on the screen are crucial in telling society what and whose values are important. All of the images in films influence how people treat each other, how policy is formed, and how ideas are built. Much attention has been drawn to the lack of diversity and inclusion in Hollywood, especially towards gender disparity and women’s role in the U.S. film industry. This project examines gender bias practices in Hollywood over the time frame of 40 years (from 1980 to 2020) by analyzing historical data of over 2000 movies across different genres and sizes. The study inspects Hollywood’s gender gap between female and male filmmakers in aspects such as involvement distribution by utilizing data exploration and visualization. Importantly, in order to see the correlation of the film industry’s assumption about female-centric movies and lower returns (i.e., rating score and gross revenue), the study predicts a movie’s score and gross revenue based on the movie attributes and content using the Bechdel test, a critical test which tracks the completeness of women representations on screen. The project evidently refuses the Hollywood assumption of female filmmakers and low movie’s returns.

**1. Introduction**

The U.S. media and entertainment (M&E) industry is the largest entertainment industry in the world, it represents a third of the global M&E industry including motion picture, television, music, streaming content, and all other media products (Select USA, 2020). In this industry, filmed entertainment (motion pictures, television, and video) plays the leading role with more than $25.9 billion in revenue last year, ranking first in the world (Statista, 2020). Research has shown that film has a measurable impact on an individual’s perception and values (Cape, 2003), these beliefs then influence a person’s attitude and behaviors. It is fair to say Hollywood and the U.S. film industry are responsible for exporting American culture and values, lifestyles, ideologies and beliefs. Hollywood has massive power in telling the world whose values and stories are important. Therefore, it is crucial to see who is writing, directing, and telling stories with such huge impacts.

Numerous studies have linked the analogous influence of mass media and films on developing society’s expectations of gender roles and conduct, self-evaluative standards, and self-efficacy beliefs (Bandura’s social cognitive theory). Undeniably, Hollywood has been a male-dominated business for many decades. According to UCLA Hollywood Diversity Report in 2020, white men continued to dominate executive positions of top Hollywood studios, 84.9% of directors in theatrical films are male, 82.6% of writers are male, and 55.9% of lead actors are also male. Accounted for 50.8% of the U.S. population (U.S. Census 2020), females are woefully underrepresented and under-employed in the country’s film industry for over 50 years. As a result of remaining significantly discounted in behind-the-scenes critical employment arenas, women on screen are often stereotyped, sexually objectified, confined to domestic roles with little to no career-driven passion.

In order to unfold the gender gap in Hollywood hiring and casting practices, it is necessary to analyze the industry assumption of low returns (i.e., rating score and gross revenue) when movies are made by female filmmakers. Female directors and their works are deemed to be not as lucrative and hold smaller value compared to male directors’ stories. This leads to the fact that female filmmakers face a limited scope of career opportunities (Smith et. al, 2017), and are often seen as presenting more financial risk than male filmmakers. As a result, female filmmakers rarely received the same resources and budget as their counterparts. In addition, movies starring female actors and are about female-centric stories are often seen as not profitable.

To examine both assumptions of Hollywood about movies made and starred by women, a predictive analysis angle is used. The project is focused on examining the industry ‘misperception of audience’ that movies made by and about women would not generate great outcomes, by predicting a movie’s IMDb score and gross revenue, given the movie attributes and Bechdel test features. Audiences are requesting more diversity and inclusion in movies for the past decade (UCLA report, 2020). It is time for Hollywood to speed up its changing progress towards more women and minority inclusion.

The remaining structure of the paper is organized as follows: In Section 2, I provide an overview of related literature with a focus on (1) studies of Hollywood film’s impacts on society, especially on gender conceptions, and (2) data analytics research on gender disparities in films; in Section 3, I discuss the methods and analytic framework; Section 4 provides results of the analytics; and finally, in Section 5, conclusion and limitation are discussed.

**2. Literature Review**

This section is focused on (1) understanding the influence of Hollywood motion pictures on society and especially on the gender social conducts, beliefs, and behaviors; (2) reviewing studies that utilize data analytic tools to discover gender disparity in Hollywood.

*2.1. Influence of Hollywood films*

The immense influence of motion pictures on human conceptions of society and everyday life has been studied for decades by psychologists and sociologists.

Fearing (1947) studies the influence of movies on attitudes and behavior. Motion pictures are constructed by a coherent sequence of audio-visual images about human stories, actions, and interactions, which at the same time present attitudes, rewards, or ideological beliefs toward the characters’ actions. Therefore, the fact that any movie, regardless of its genre (romantic, musical, or documentary, etc.), has some measurable effects on specific delivered attitudes to its audiences is inevitable.

O’Toole (2001) analyzes gender and racial representation in films with a sociocultural approach. Humans recognize and empathize with films images because films resemble our real life and also our imagination. By creating the capacity for us to understand ourselves and others in a broader social context, motion pictures assist humans in structuring our society and our place in it. In films, characters’ life aspects such as time and space are controlled and manipulated by filmmakers. O’Toole defines the creation of stories, experiences, and knowledge from films as its ability to shape humans’ perception of reality and culture. In other words, motion pictures are a source of ‘constructed reality’ that can convey an ideological message to the audience. Because of the movie’s significance in shaping our culture and beliefs, the author emphasizes the link between gender/racial representation and people that produce, direct, and finance movies also control the message of how individuals should act as well as what society should believe and value.

Behm-Morawitz (2008) utilizes **‘**social cognitive theory’framework to study gender portrayals in teen movies and the effect it has on gender-based beliefs about friendship, social behaviors, and women’s role in society. The study first analyzes gender’s portrayals depicted in teen movies, then measures the extent to which exposure to these images influences young adults’ beliefs about females in society. Social cognitive theory, in simple terms, suggests that portions of a person’s knowledge acquisition and social behaviors can be directed related to observing others’ social interactions and experiences. Individuals adopt gender characteristics by observing and monitoring the consequences and rewards associated with others’ behaviors. The theory holds media messages as a meaningful source for an individual’s acquisition of ‘gender-linked knowledge and competencies’. Women’s representations on-screen then directly contribute to building viewers’ perceptions of gender identity (gender expectations in society).

*2.2. Gender disparirties in Hollywood*

Smith et. al (2017) study the inclusion of Hollywood directors’ diversity in terms of gender, race, and age in 1,000 popular films using a data analytic framework. Key findings of the study include a prevalence of female directors across the top films from 2007 to 2016 to be woefully suppressed: out of 1,114 directors in the dataset, only 45 of them are female (4%); females rarely direct in popular genres such as action or thriller; female director are usually “one and done”, which means that it is extremely difficult for female directors to be financed and hired to direct a second movie. The study discovers the reasons behind the upsetting number of female directors. One of the reasons is that the film marketplace is gendered: the industry leaders believe that women make small and independent films, thus the work and movies assigned to women have little value compared to men. Female filmmakers face a limited scope of career opportunities. The study proposes evidence-based solutions in solving the problem of gender bias. The authors strongly encourage all related parties in the film industry to first acknowledge the data that confirms impressive gender disparity, then progressively change the hiring practice as well as the problematic beliefs towards female directors.

Kagan et. al (2020) investigate gender bias in on-screen female characters over the past century. The study focuses on the portrayal of women in movies by fusing data from the IMDb movie database with a dataset of movie dialogue subtitles. The largest corpus of movie social networks is created from this study. From this dataset, the authors evaluate female roles in movies as an alternative to the well-known Bechdel test by answering these questions: “Are there movie genres that do not exhibit a gender gap?”; “What do characters’ relationships reveal about gender, and how has this changed over time?”; “Are women receiving more central movie roles today than in the past?”; “How has the fairness of female representation in movies changed over the years.” The findings of the study include a trend of improvement in all aspects of women’s roles in movies.

**3. Methodology**

In this project, two targeted variables (i.e., gross revenue and IMDb score) are predicted to inspect the mentioned Hollywood’s assumption about females in the film industry: (1) Movies made by women are not profitable, and (2) Female-centric movies don’t have great returns.

In step 1, the IMDb dataset obtained from Kaggle was explained and preprocessed; data exploration and visualization are utilized to examine gender distribution, differences in movie contents when the gender of the filmmakers changes. In step 2, multiple predictive models including (a) multiple linear regression and its regularization methods, (b) regression tree and its ensemble methods, (c) k-nearest neighbors were employed.

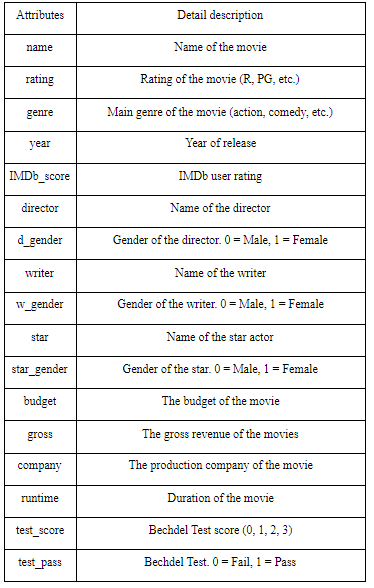
*3.1.1. Data preparation*

The dataset is obtained from the Kaggle website, the author scraped this data from the IMDb movies database. Originally, the dataset has more than 7000 records and 15 variables. The dataset was modified to only include movies made in the U.S., gender attributes for the director, writer, and star were added manually to ensure analysis’ accuracy since gender is considered to be one of the most important predictors.

After some modification is done for gender features, the movies’ content features are obtained from using Python’s Beautiful Soup library to scrape the Bechdel test website. Scraping the Bechdel test website *‘*[*https://bechdeltest.com/*](https://bechdeltest.com/)*’* was done by utilizing the open-source code for scarping the website on GitHub (MagicMilly GitHub). The Bechdel is significant to the analysis because the test inspects how humanized and complete are the representations of women in movies. The test asks three questions: (1) Are there at least two women in the movie who have names? (2) Do these women talk to each other? (3) Do they talk to each other about something other than a man? The test seems to ask simple questions to show the agency of women portrayed on screen. Passing or failing the Bechdel test gives insight into the movie’s story how is it related to the movie’s outcomes. Scraping the Bechdel test website provides the Bechdel test features of [test\_score] and [test\_pass] that give information about female representation in each movie.

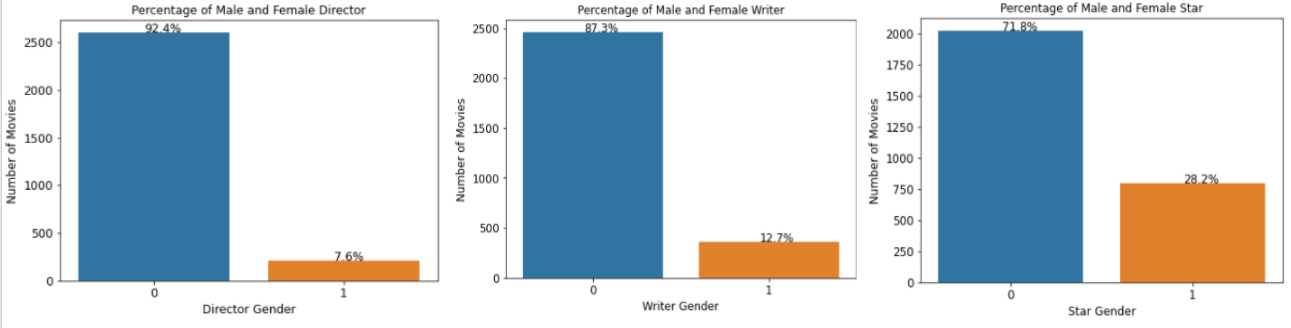
Table 1 shows a detailed data dictionary of the dataset. The dataset includes a variety of 2815 movies from 1980 to 2020. Around 800 movies do not have corresponding budgets. The data cleaning process includes using Python to identity missing movie budget that is directed by females, then doing manual input on those missing values. The same task was performed on missing a movie budget that is written and starred by females. Handling budget missing value for female filmmakers required a more detailed and accurate approach since the female class is significantly smaller in all three attributes (director, writer, and star gender) compared to male filmmakers. Budget missing values for male filmmakers were replaced by the mean of the corresponding movies’ score bin. The score was sectioned into three categories (less than 6, from 6 to 6.9, equal and more than 7).

Table 1. Data description



*3.1.2. Data visualization*

Women are significantly underrepresented in the dataset. Fig. 1 displays the percentage of female to male directors, percentage of female writers and female lead actors. Out of 2185 movies frm 1980 to 2020, only 7.6% of movies are made by female directors, 17.5% of those movies are written by female, and 28.2% are starred by female actors.

  
Fig. 1. Percentage of female and male filmmakers

The data confirms the gender gap in movies budget received by female filmmakers compared to males (see fig. 2). However, the is no significant difference in the mean rating scores. Fig. 3 shows the boxplot of IMDb score. Movies directed by females have the same maximum, while there are many movies directed by males that fall under the score minimum.

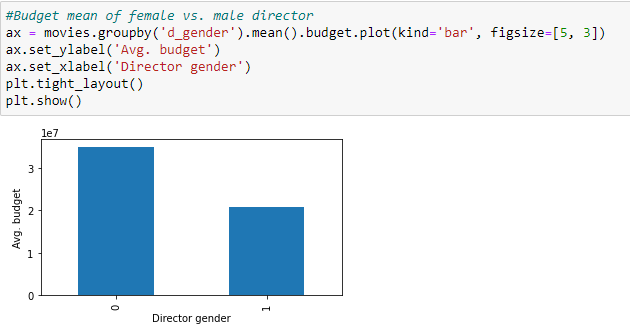


Fig. 2. Movie budget by director gender.

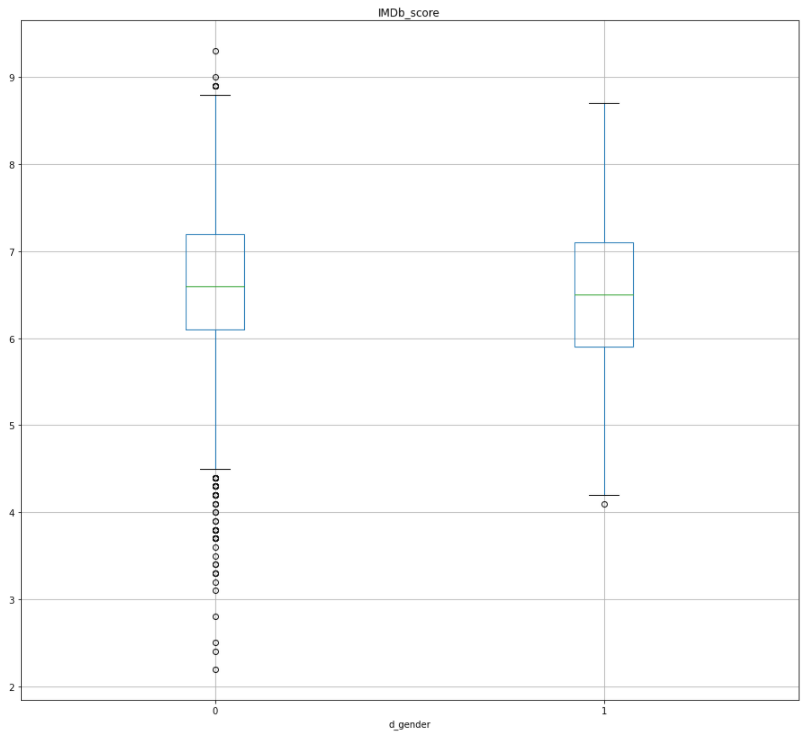


Fig. 3. IMDb score boxplot by director gender.

Female characters’ representation fluctuates when the gender of the director changes. Fig. 4 shows the violin plots of director genders to the passing/failing the Bechdel test as well as the test score to director gender. Female filmmakers are more likely to make films about women with more ‘humanized’ conversations. There are 2601 male directors in the dataset, only 54% of the movies directed by men pass the Bechdel test. The average score of the Bechdel test for movies directed by men is 2.12. This suggests that male directors don’t give as much attention to how they portray female characters on-screen as female directors do. The dataset has 92.4% of male directors, which conveys that most movies are directed through the male gaze and on-screen female portrayals are woefully biased.

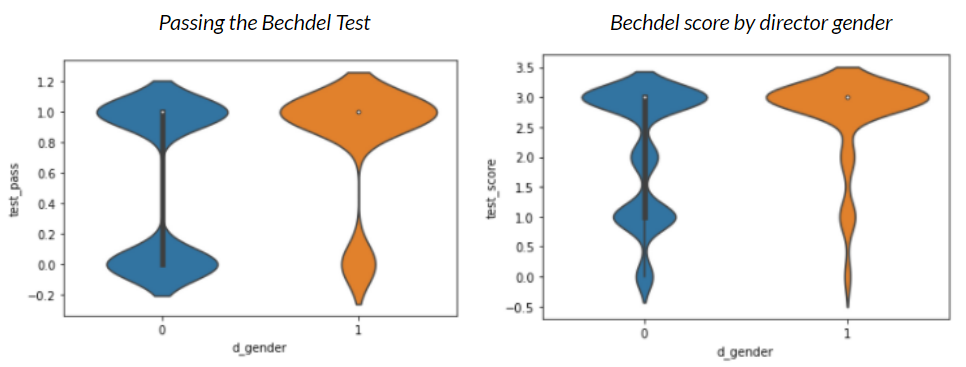


Fig. 4. The Bechdel test and director gender.

*3.2. Predictive models*

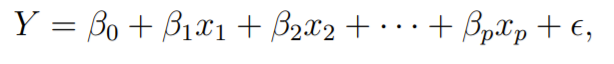
A variety of predicting models are used to predict the gross revenue and IMDb score, including Multiple Linear Regression (MLR) and its regularization regression models of Lasso and Ridge regression, regression tree and its ensemble methods of random forest and gradient boost forest, K-Nearest Neighbor (KNN) is used to predicting gross revenue of a movie that is not in the dataset. To assure fairness and comparison among different models, the dataset is split into three sets of training data (60%), validation data (20%), and testing data (20%). The split is used on all MLR models, regression trees and ensemble methods, as well as in KNN.

Features selection is based on the assumption of female filmmakers in Hollywood. Therefore, gender features such as director gender, writer gender, and star gender are important features; the Bechdel test features of test score and test pass features are also significant; movie features of the release year, movie’s budget and votes are chosen to as predators as well. Hence, the predictors for predicting gross revenue and IMDb score are the same [d\_gender, w\_gender, s\_gender, test\_score, test\_pass, votes, budget]

*3.2.1 Multiple Linear Regression*

Multiple Linear Regression is utilized in the predictive model as well as the explanatory model.

MLR is used to fit a relationship between a numerical outcome/dependent variable and a set of predictor/independent variables. The regression assumption of the linear relationship is expressed in the following function:



*β0, . . . , β*p are coefficients and *ϵ* is the noise.

Multiple linear regression, at the predictive analysis level, focuses on finding a regression model that best predicts continuous values for new individual records. MLR predictive model performance is measured on the holdout or validation set after the training set is used to estimate the model. The most common metrics for evaluating the MLR predictive performance are mean error (ME), root mean square error (RMSE), and R-squared. The errors are also known as the residuals or deviations of the observed and predicted values. The first metric, ME is simply the average of the model’s residuals, which does not account for the possibility that positive and negative errors could balance out. In most cases, ME does not explain the magnitude of the model errors. RMSE takes the square root of the average squared error, which eliminates the positive and negative signs of errors. The R-squared measures the fit of the MLR model by taking the variance (errors) explained by the model over the total variance (Shmueli et. al, 2018).

An explanatory model uses the data to estimate the coefficients and to quantify the noise (Shmueli et. al, 2018). The Ordinary Least Square (OLS) method and its statistical summary are helpful in assessing the significance of independent variables in explaining the outcome variable. OLS regression focuses on estimating the best-fit model by using the method that finds predictors’ coefficients to minimize the sum of squared deviations between the actual outcome values and the predicted values from the model. The explained information of each independent variable to the dependent variable in MLR could be estimated using two metrics: The regression coefficients and the p-value. The regression coefficients explain the average change in dependent variable for one unit of change in the predictor variable while holding other independent variables constant in the model. The p-value tests the significance of one predictor variable. Commonly, a low p-value (<0.05) indicates that the predictor variable is statistically meaningful to the linear regression model (Minitab, 2013).

Ridge and Lasso regressions are extended regularization methods of MLR. Regularization methods reduce the magnitudes of the coefficients and also the variance of the outcome variable in order to reduce model complexity and prevent overfitting. Ridge regularization puts a penalty threshold that is equivalent to the sum of squares of the magnitude of the coefficients. Lasso regularization puts a penalty on the sum of absolute values of coefficients. Lasso regression can lead to zero coefficients if some of the predictor features are completely neglected for evaluating the outcome. Normalizing the data to put them into the same scale is necessary for regularization methods (Bhattacharyya, 2018).

*3.2.2. Regression tree and ensemble methods*

Regression tree is a data-driven method from the decision tree algorithm family. The decision tree algorithm’s logic lies in the two key ideas of recursive partitioning in constructing the tree and pruning or cutting the growth of the tree (Shmueli et. al, 2018). The regression tree algorithm is extended from the decision tree method, in which the prediction of the outcome variable is numerical values, but the same procedure occurs to determine the best prediction. The fitting process for the training data proceeds as follows:

* Many splits are attempted by dividing the predictor variables into distinct regions, a split is selected to minimize the impurity - which is the sum of squared deviations from the region mean
* For each observation that falls into each region, prediction is computed as the average of response in the training data in that particular region (Vala, 2019).

Regression tree predicting performance is measured by RMSE, in the same way as linear regression. Letting the tree grow fully will cause the model to be overfitted to the training set and reduce prediction power on the holdout set. Therefore, pruning the tree by exhaustively building trees using different parameter values, coupling with cross-validation method to find the best tuning parameters might increase the prediction accuracy of the regression tree model. After fitting the training data into the best-pruned tree, the holdout data is used to assess the model performance.

Regression tree prediction can be improved by taking advantage of combining results from multiple trees. This multi-tree approach is known as the ensemble method, in which random forest (bagging) and gradient boosted forests are employed.

The random forest model process is explained as follows:

* Multiple random observations (samples) are drawn with replacement from the training set
* A random subset of predictors is used at each stage, and a regression tree is fitted to each sample
* Predictions are determined by averaging the results from the individual trees (Shmueli et. al, 2018).

Gradient boosted forest is more useful in classification tasks than prediction tasks since the method of boosted forest concentrates on the misclassification of minority labeled class. Thus, in regression tasks, boosted forest performance in predicting continuous numerical values is not significantly useful.

Regression tree models do not provide coefficients for predictor variables but they can produce variable importance scores (Shmueli et. al, 2018). Hence, regression tree models are utilized in order to measure the relative contribution of different predictor variables in comparison to multiple linear regression.

*3.2.3. k-Nearest Neighbors*

K-Near Neighbor (KNN) regression is an extended algorithm of the KNN highly automated data-driven method. To predict a continuous outcome, KNN regression averages the observations in the pre-determined k-nearest neighbors. Neighbors are determined by computing distances. Finding the best number k is determined by using an error metric - in this analysis is the root mean square error (RMSE).

KNN algorithm makes no assumptions about the data since the prediction is based on similar features of records. Therefore, KNN is employed to estimate a new record using the same predictors as other previous models.

**4. Results of predictive models**

Multiple linear regression is considered to be the best model for predicting both gross revenue and IMDb score since MLR can explain the statistical relationship between predictors and outcome variables, at the same time predicting the values at an acceptable level of accuracy.

Predictor features include gender attributes [d\_gender, w\_gender, s\_gender], the Bechdel test attributes [test\_score, test\_pass], and movie attributes [votes, budget].

*4.1. Predicting gross revenue*

Gross revenue is used as a target variable in prediction to inspect the Hollywood assumption that movies made by women don’t generate great revenue. Gros revenue has a significantly large variance range: from $3,121 to $2,847,246,203 and the standard deviation is 215,367,822. The large variance range is important in interpreting the performance of predictive models.

*4.1.1. Multiple linear regression results*

First, an MLR model with no additional methods is utilized to fit the training data, fig. 5 shows the coefficients of predictors and the performance of the model on the training set. All gender attributes and Bechdel test attributes have generally large positive coefficients. The RSME is more than 120,000,000, this large value is expected since the standard deviation of gross revenue is more than 200,000,000.

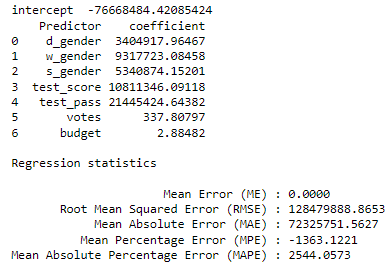


Fig. 5. MLR performance the training set.

After fitting the model with training data, prediction on the validation set has proceeded with additional information about residuals. Fig. 6 presents the performance of the model on the validation data. RMSE increases from the training set by around 6% in the validation set. This is not a significant decrease in accuracy performance. Fig. 7 visualizes the distribution of residuals on validation data, most of the errors occur in the ± 250,000,000.

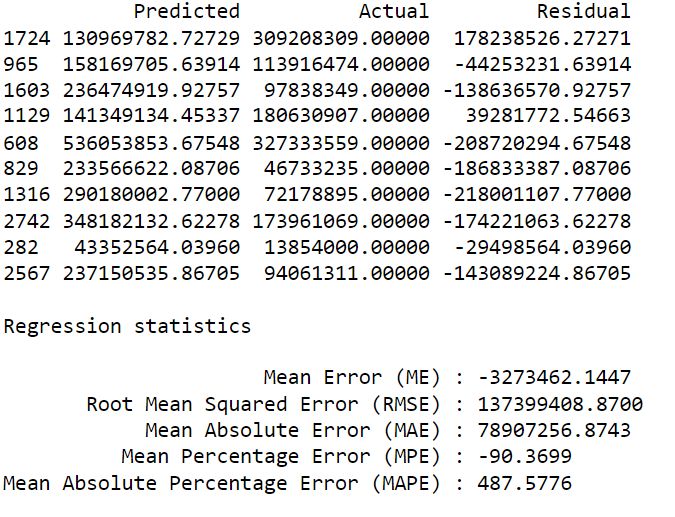
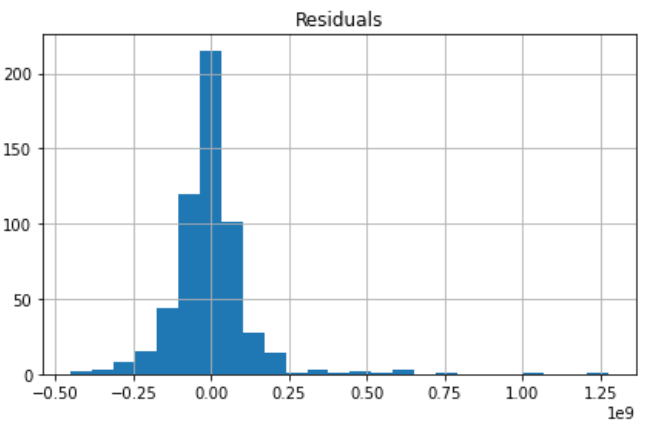


Fig. 6. MLR performance on valid. Fig. 7. Validation set residuals.

Second, the training set is used to fit the OLS regression model. The OLS explanatory regression model attempts to capture the average relationship between the predictors and the target. The accuracy of the best-fit line through the training data has R-squared equals 0.63 and adjusted R-squared equals 0.629. The best-fit line presents a respectively high performance in fitting the training data into a linear regression model. Fig. 8 shows the coefficients, standard deviation errors, t-statistics, and p-values of seven predictors. All of the gender features in the explanatory model have large positive values and high p-values. This suggests that although a movie’s gross revenue might increase if females direct, write, and star in that movie (since gender features have 1 as female) due to the large positive coefficients, high p-values imply that there could be no statistically significant relationship between gender features and gross revenue. The Bechdel test predictors have lower p-values: test\_score: 0.143, test\_pass: 0.174. Although these p-values might be larger than the common threshold of <0.05, high positive coefficients of these predictors may convince that passing the Bechdel test with high scores (i.e., the score of 3 is maximum) would generate greater gross revenue for a movie.

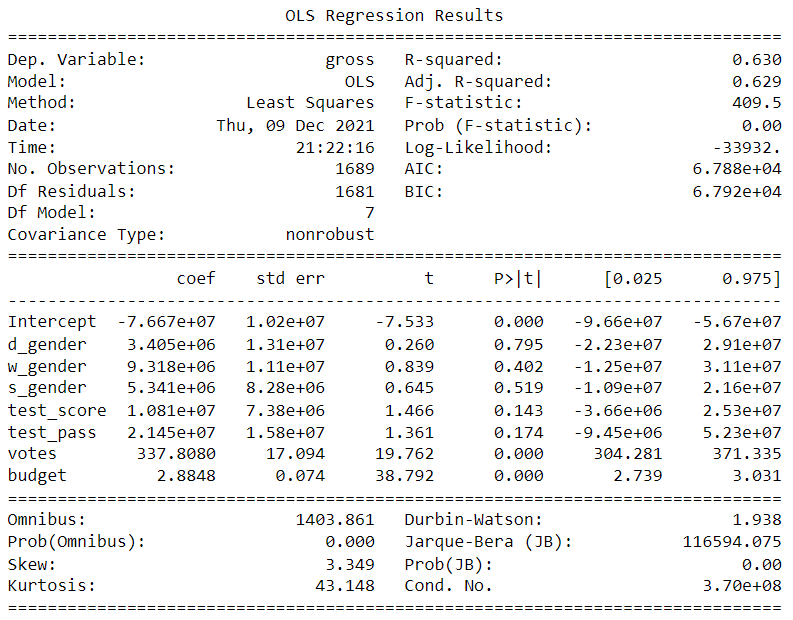


Fig. 8. OLS explanatory regression model on the training set.

The OLS regression model that fitted with training data is then used to predict the validation and testing set. The performance of OLS regression is shown in fig. 9, the adjusted R-squared is ~0.629, which is almost the same as the training set adjusted R-squared. The predictive performance of the OLS regression model on validation is considered to be good since there is no decrease in adjusted R-squared.

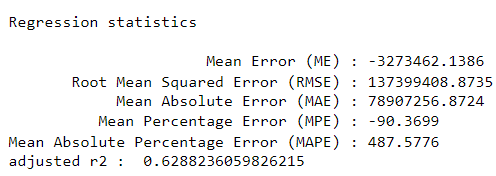


Fig. 9. OLS regression performance on the validation set.

On the holdout testing data, OLS regression performs better with a smaller RMSE compared to the validation set, as well as a higher adjusted R-squared of 0.644 (see fig. 10).

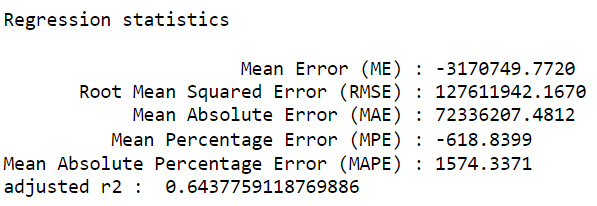


Fig. 10. OLS regression performance on the testing set.

OLS regression explanatory and predictive model provides such insights: (1) there is no evidence from concrete IMDb data to make the assumption that movies made by females don’t generate great gross revenue; (2) passing the Bechdel test may increase the gross revenue for a movie.

Lasso and Ridge regularization are employed to test for the improvement of the MLR model in predicting gross revenue. Lasso regression with its penalty approach of L1 - the sum of absolute values for coefficients, presents an almost identical performance on all of the accuracy metrics to OLS regression on the testing data (see fig. 11). Ridge regression, however, indicates that there is a negative adjusted R-squared of -0.268 (see fig. 12). Ridge regression performance on the testing set suggests that there is a very low explanation of predictors towards gross revenue. The performance of Lasso and Ridge regularization aligns with the insights gained from OLS regression that the assumption about female filmmakers and gross revenue is wrong.

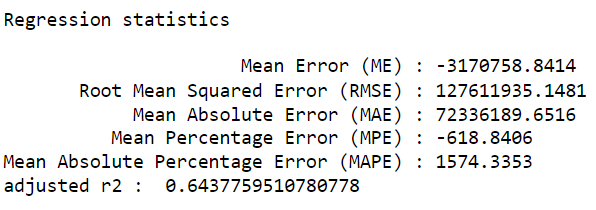


Fig. 11. Lasso regression on the testing set.

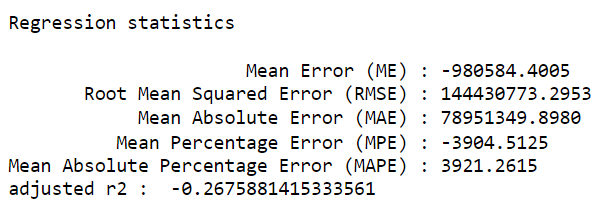


Fig. 12. Ridge regression on the testing set.

All of the employed MLR models agreed to refuse the assumption that movies made by female filmmakers don’t generate great gross revenues.

*4.1.2 Regression tree and ensemble methods*

The regression tree and its ensemble methods predictive models are utilized to predict gross revenue, and most importantly to inspect the importance of predictor variables in the models.

Fig. 13 shows the prediction performance of the regression tree model on the holdout testing set. The regression tree model is developed on the best-pruned tree, which is obtained by using a Grid search with different parameter values for the tree’s maximum depth, minimum impurity, and minimum samples split. The RMSE of the testing set increased by 67.88% from the validation set. This is a large increase in error, which indicates that the regression tree model is not ad suitable in predicting gross value as the MLR models.

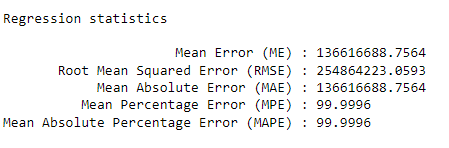


Fig. 13. Regression tree performance on the testing set.

Despite its limitations in predicting gross revenue, the regression tree method provides de insights into variable importance. Fig. 14 presents the predictors’ importance to the model. The gender predictors are considered to have little to no influence on the target variable; the Bechdel test score does have some significance in explaining the outcome. However, the regression tree algorithms do not explain the negative or positive relationship of the predictor features towards the outcome.

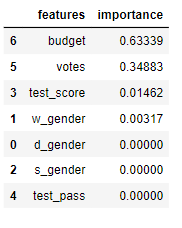


Fig. 14. Regression tree features importance.

Random forest and boosted forest regression predictive power is similar to regression tree model, in terms of lower prediction accuracy but provide features importance measurement. The importance of predictor features is similar to the regression tree model. Gender attributes are considered to be not influential in predicting the gross revenue of a movie (see fig. 15).

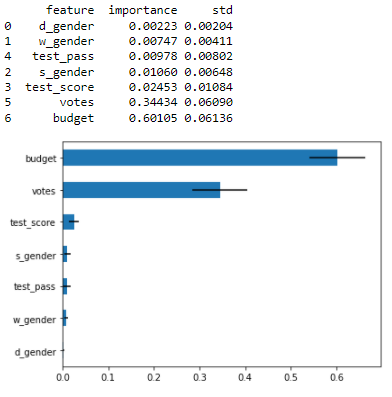


Fig. 15. The random forest features’ importance.

By employing two different regression approaches MLR and regression tree to predict a movie’s gross revenue, the statistical importance of predictors can be discovered from the predictive models. MLR and regression tree predictive models are similar in terms of what independent variables that these models considered to be important. The gender of filmmakers and the stars of a movie might not relate to the movie’s gross revenue. Therefore, the myth of movies made by female filmmakers doesn’t generate great gross revenue is proven to be untrue.

*4.1.3. K-Nearest Neighbors*

KNN algorithm is used to predict a new movie’s gross revenue, ‘Eternals’. Fig. 16 shows the new movie features. The actual gross revenue is around $300,000,000. The KNN model predicts the movie’s gross to be more than 1 billion dollars. This prediction is not that accurate, but it may suggest that the developed model recognizes the importance of predictor features to the outcome.



Fig. 16. The new movie features.

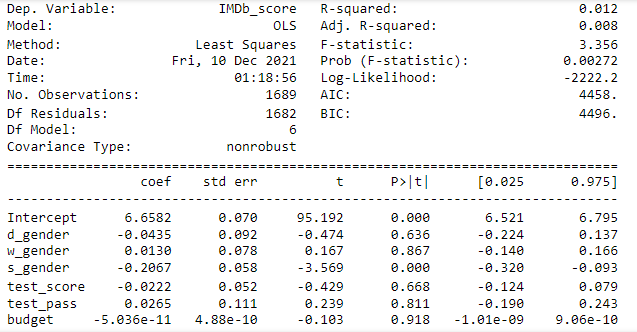
*4.2. Predicting IMDb score*

IMDb score is used as a target variable in prediction to inspect the Hollywood assumption that female-centric movies are not as good. IMDb score has a significantly small variance, from 2.2 to 9.3 with a standard deviation of 0.899. Due to the small variance, prediction accuracy on all models doesn’t have really good performances. The independent variables used to predict a movie’s IMDb score are gender attributes [d\_gender, w\_gender, s\_gender], the Bechdel test attributes [test\_score, test\_pass], and the movie’s budget. [votes] is excluded since the IMDb score is derived from the number of votes.

*4.2.1. Multiple linear regression*

The OLS regression model is utilized to first explain the relationship between predictors and IMDb score using the training data, and then predict the score on validation and testing set.

Fig. 17 discusses the explanatory OLS regression model. The best-fit linear line has an R-squared of 0.012 and an adjusted R-squared of 0.008. These really R line-fitted values convey a weak explanatory power of predictors towards the IMDb score. The Bechdel test attributes have high p-values, this suggests that there might be no relationship between these variables and a movie’s score. The OLS regression model also does not perform well on the testing set (fig. 18) since the adjusted R-squared is negative 0.029, although the RMSE of 0.89, which is equal to the targeted variable’s standard deviation.



Fig, 17. IMDb score OLS regression on the training set.

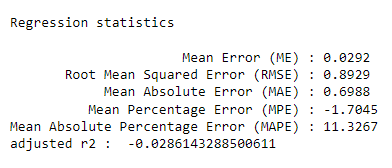


Fig. 18. IMDb score OLS regression performance on the testing set.

Ridge regression model prediction on the testing set is much different than the OLS model prediction. By penalizing the coefficients in the model using L2 - the sum of squared coefficients, Ridge regularization produces a more meaningful model in explaining the IMDb score using the chosen independent variables. Fig. 19 shows the performance of Ridge regression on the testing set. Adjusted R-squared is positive 0.0035, which is slightly bigger than the negative values from the OLS regression and also the Lasso regression model.

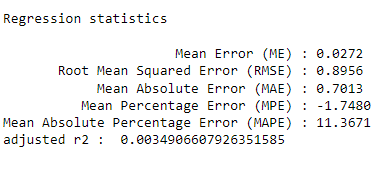


Fig. 19. IMDb score Ridge regression performance on the testing set.

All MLR models have low accuracy performance and significantly small coefficients for the Bechdel test variables. This suggests that a movie’s IMDb score does not depend on passing or failing the Bechdel test. Therefore, the assumption that female-centric movies are not good is untrue. The IMDb score is derived from the number of votes submitted by IMDb users and not determined by movie critics (Reynolds, 2017).

*4.2.2. Regression tree and ensemble methods*

The regression tree predictive model is developed to predict IMDb score by utilizing a grid search to find the best-pruned tree. Fig. 20 is the regression tree result on the training data. The variable [s\_gender] which is the gender of the movie’s star is considered to be the only important predictor in the best-pruned regression tree. according to this best-pruned tree, if [s\_gender] is lower than 0.5 (star gender is 0 for male, and 1 for female), the IMDb score will be slightly lower. Due to the fact that the best-pruned tree is too shallow and the averaging method behind the regression tree algorithm, the splitting criteria and the results do not convey much useful information about the relationship between predictors and the outcome.

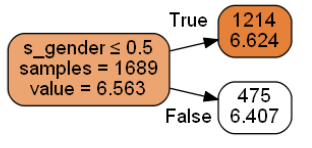


Fig. 20. IMDb score regression tree plot.

The predictive performance of this small tree also decreases in the metric RMSE from validation to testing data. Fig. 21 shows the regression tree model performance on the testing set, RMSE is 0.8929, which increases from the validation set (RMSE = 0.8682).

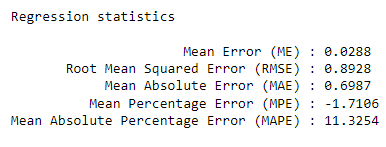


Fig. 21. IMDb score regression tree performance on the testing set.

The ensemble methods, random forest and boosted forest, are done by setting the number of trees built to be 500 trees. The predictive performances of these two methods are shown in fig.22. The R-squared values are negative, which suggests that the regression tree approach is not as appropriate as the MLR approach in predicting the IMDb score.

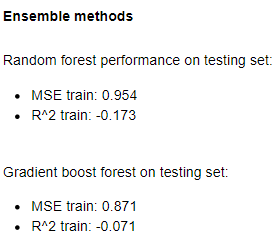


Fig. 22. IMDb score ensemble method performances on the testing set.

However, the random forest algorithm’s ability to display the importance of features is helpful in this case. Fig. 23 shows that the most important feature in predicting IMDb score is budget, the Bechdel test score is also considered to be somewhat important as well.

Although the regression tree and ensemble predictive models do not perform well, insights about the importance of features in predicting the IMDb score are provided. These findings align with the results from MLR models, in which the IMDb score might not depend on the Bechdel test criteria and the filmmakers’ gender.

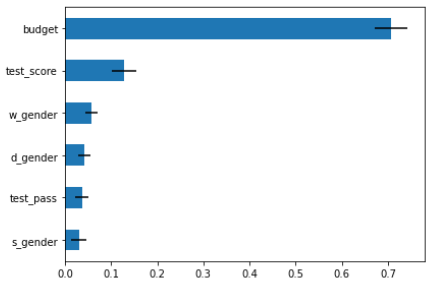
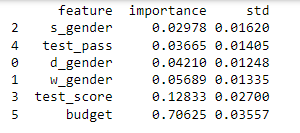


Fig. 23. IMDB score random forest importance of features.

4.2.3. K-Nearest Neighbors

As mentioned above in Section 3, the KNN algorithm is utilized to predict a new movie’s IMDb score, rather than predicting the whole targeted feature. The predicting movie is ‘Eternals’, which is recently released in November of this year, the movie is directed and starred by women (fig. 24). Chosen k in the analysis is k = 8, determining the number of k is based on running a loop through the training data and using the RMSE metric. The actual IMDb score for this movie is 6.8, the model predicted score is around 7.1, which is not so different from the actual score.



Fig, 24. The new movie features.

**5. Conclusion**

What we see on-screen directly influences how we navigate the world as well as understand our place in it. Men have been dominantly shaping Hollywood’s narrative of themselves and for women for decades. The project has evidently confirmed gender disparities in Hollywood: women are woefully underrepresented and under-employed in the country’s film industry. Through utilizing predictive analytics to predict movies’ gross revenue and IMDb score, the two famous assumptions of Hollywood about female filmmakers, female-centric movies to low returns are demonstrated to be untrue. The results from all the employed predictive models suggest that there is no strong relationship between female filmmakers or female-centric movies to the low gross revenue or IMDb score of the movie.

As our society slowly progressive changing toward equality for all (genders, sexual orientation, race, etc) Hollywood’s narrative desperately need to change in the same direction. On-screen stories and representations will not change until females and people of color receive the same employment opportunities as men in the film industry.

***References***

Kagan, D., Chesney, T., Fire, M. (2020). Using data science to understand the film industry’s gender gap. *Palgrave Communications*. Retrieved from <https://doi.org/10.1057/s41599-020-0436-1>

Media And Entertainment Spotlight. *Select USA*, <https://www.selectusa.gov/media-entertainment-industry-united-states>

Filmed entertainment revenue in selected countries worldwide in 2019. *Statista*. Retrieved from https://www.statista.com/statistics/296431/filmed-entertainment-revenue-worldwide-by-country/#statisticContainer

Hollywood Diversity Report 2020: A Tale of Two Hollywoods. (2020). *The University of California at Los Angeles*. Retrieved from <https://socialsciences.ucla.edu/wp-content/uploads/2020/02/UCLA-Hollywood-Diversity-Report-2020-Film-2-6-2020.pdf>

Fearing, F. (1947) Influence of Movies on Attitudes and Behavior. *ANNALS,* 254(1). Retrieved from <https://journals.sagepub.com/doi/10.1177/000271624725400112>

Bussey, K., Bandura, A. (1999). Social cognitive theory of gender development and differentiation. *Psychological Review*, 106(4):676-713. doi:10.1037/0033-295x.106.4.676

Gavin S Cape. 2003. Addiction, stigma and movies. *Acta Psychiatrica Scandinavica,* 107(3):163–169.

Morawitz, B., E. (2008). Mean Girls? The Influence of Gender Portrayals in Teen Movies on Emerging Adults' Gender-Based Attitudes and Beliefs. *Journalism & Mass Communication Quarterly*, 85(1):131-146.

Smith, L., S., Pieper, K., Choueiti, M. (2017). Gender, Race, & Age of Film Directors Across 1,000 Films from 2007-2016. *USC Annenberg*. Retrieved from <https://annenberg.usc.edu/sites/default/files/2017/04/06/MDSCI_Inclusion%20_in_the_Directors_Chair.pdf>

Grijalva, D. Movie Industry. Retrieved from <https://www.kaggle.com/danielgrijalvas/movies>

O’Toole, H.K. (2001). A Sociocultural Analysis of the Gender and Racial Representations in Films Offering a Vision of the Future of Humanity. *American University.*

Shumueli, G., Bruce, C. P., Yahav, I, et. al. (2018). Data Mining for Business Analytics. *Wiley.*

Reynolds, M. (2017). You should ignore film ratings on IMDb and Rotten Tomatoes. *Wired*.

<https://www.wired.co.uk/article/which-film-ranking-site-should-i-trust-rotten-tomatoes-imdb-metacritic>

Minitab Blog Editor (2013). How to Interpret Regression Analysis Results: P-values and Coefficients. *Minitab.* <https://blog.minitab.com/en/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients>

Bhattacharyya, S. (2018). Ridge and Lasso Regression: L1 and L2 Regularization. *Towards Data Science.* <https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b>

Vala, K. (2019). Tree-Based Methods: Regression Trees. *Towards Data Science.* <https://towardsdatascience.com/tree-based-methods-regression-trees-4ee5d8db9fe9>