Analysis of Factors Influencing the Probability of a Nominated Film Winning an Academy Award for Best Picture

Abstract: This paper investigates factors that influence the probability of an Oscar-nominated film winning an Academy Award for Best Picture with a focus on race and gender variables. Using data from various movie databases, the project utilizes a multiple logistic regression model that correctly predicts the outcome of a nominated film 89% of the time. The model shows that gender diversity of cast, racial diversity of production, and the square root of fame of production are significant predictors in predicting if a film won. Of particular interest, there was a negative association between gender diversity of cast and probability of winning and a positive association between racial diversity of production and probability of winning. Further research could focus more on issues of racial diversity in the content of films as well as other types of diversity, such as LGBTQ+ representation.

Background and Introduction:

As America faces a reckoning with the racist and misogynist structures of power that have shaped its history, culture, and institutions since its inception, the question of positive representation in media has grown increasingly important. Nowhere has the want of diversity in media been as prevalent as in Hollywood. According to UCLA's Diversity Report 2020, despite improvements as compared to 2019, black and indigenous people of color (BIPOC) are still consistently underrepresented in all sectors of the industry, with only 27.6% cast as film leads and 15.1% as film directors (Hunt, Ramón). Award ceremonies like the Oscars and the Golden Globes have been plagued with controversy surrounding the prevalence of white individuals as actors, directors, and members of selection committees (Rottenberg, Perman). As recently as the 2021 Golden Globes, at the same time that Chloé Zhao made history as the first Asian woman to win the Golden Globe for Best Director (Stevens), *Minari*, a movie following the experiences of an immigrant Korean family in America, sparked controversy when it won the Golden Globe for Best Foreign Language Film because it did not meet the 50% English language threshold, barring it from competing in any other category, including Best Picture (Yuen).

In the wake of this controversy, the aim of this project was to build a model that would be able to reliably predict the chances that a certain film had of winning the Oscar for Best Picture once it had been nominated. The greatest focus was on determining how issues of racial and gender representation would feature in this model, but building a comprehensive model through considering potential predictors was also important.

Data and Exploratory Analysis:

<u>Data and Variables</u>

The dataset for this project consisted of the films nominated for the Oscar for Best Picture from 2010 to 2020. Within this, 21 variables were initially considered as predictors. These were: year nominated, gender diversity of the cast, general racial diversity of the cast, Black cast diversity, non-Black BIPOC cast diversity, Latinx cast diversity, racial diversity of production, gender diversity of production, Bechdel test (categorical indicator), rating (split into indicator variables), genre (split into indicator variables), runtime, budget, lifetime gross, audience review, critical review, Oscar nominations, Oscar wins, fame of cast, fame of production, and true story (categorical indicator). The response variable was the categorical variable of whether or not the film won the Oscar for Best Picture. The data was collected from the following websites: IMBD, TMBD, Bechdel Test Movie List, Box Office Mojo, Rotten Tomatoes, and the Official Oscars Database.

While certain variables were easy to define and find data on, such as rating and budget, others required significantly more subjectivity in defining and collecting data. Specifically, variables on diversity and fame were difficult to measure. For diversity, one initial issue was the lack of data on a racial test analogous to what the Bechdel test does for gender in determining whether the content of a film is inclusive. The races of cast and production members were difficult to verify and sometimes necessitated our own biased classifications, and because of the small proportional values, it became necessary to combine the categories of Asian/Indigenous and Multiracial. Additionally, due to time constraints and a lack of easily accessible information, it was necessary to only focus on the first 10 cast members listed in the IMDB credits as well as only the director and writers for the production. This was also the case when measuring fame, which was defined as the total number of Oscars that all cast or production members were nominated for or won *prior* to the year of the current film.

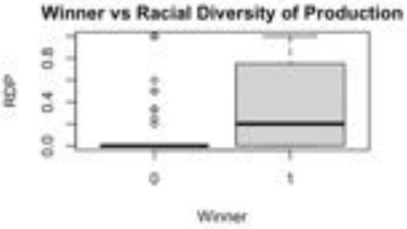
<u>Exploratory Data Analysis:</u>

To check for the linearity assumption in our empirical logit plots, we graphed every quantitative variable to check for violations. Due to some variables having a small sample size, we were unable to divide them into as many categories as we would have liked. While most of our variables seemed to meet the assumption, there was a curve in the variables of fame of production and budget. To account for this, we took the square root of each of these variables, which resulted in our assumptions being met.

We also initially attempted to classify race in smaller categories. However, these turned out to be so small that we instead chose to collapse race into three categories: black, non black BIPOC, and white.

In the interest of seeing whether representation in films has been improving recently, we also wanted to see how the race and gender variables changed over time, as well as how they compared to the winner for each year. To do this, we created graphs over time as seen in Appendix A. From these we can see that overall, the winner follows the general trends of the other movies, with a few marked exceptions.

Figure 1



Model and Results:

Analytic Methods:

A multiple logistic regression model was used to model the log odds of winning an Oscar versus our categorical and quantitative variables. After building the initial model with all 21 variables, we removed the genre variables, as they were too correlated with winning and other variables. We fit our initial model using the other variables as well as the interaction terms between racial diversity of the cast and fame of cast, gender diversity of cast and racial diversity of cast, and gender diversity of cast and Bechdel test. Graphing this initial model, it appeared as though the linearity model assumption was met. While there were some issues with the random/independence assumption because the films inherently were not chosen randomly, there was no possibility of bettering this, so we continued. Our initial model predicted the outcome correctly 90% of the time, but it had too many variables to be easily interpretable.

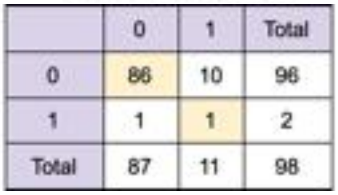
To remove some variables, we ran stepwise selection starting from the full model in both directions. Appendix D shows the stepwise selection done by R. The resulting model had three variables: the gender diversity of cast, racial diversity of production, and square root of fame of production. After checking model assumptions, we found they were relatively met and kept this as the final model.

Final Model and Results:

Our final model predicts the log odds of winning an Oscar using gender diversity of cast, racial diversity of cast, and square root of fame of production. At a 0.05 significance level, all of our predictors were significant. See Appendix C for a summary table.

When compared to an intercept only model, the likelihood-ratio test gave us a Chi-Squared value of 20.02, and a p-value of 0.001681. This indicates that our model fits more effectively as compared to an intercept only model. Our model correctly predicted the outcome of each movie 89% of the time, as is further detailed in Appendix C and pictured in Figure 2. The model diagnostics showed no significant violations, but we had some issues elaborated on in the conclusion.

Figure 2: Prediction Table



	0	1	Total
0	86	10	96
1	1	1	2
Total	87	11	98

Appendix C displays the 95% confidence intervals for the fold change in log odds associated with each variable. The gender diversity of cast and square root of fame of production having a fold change less than one suggests that an increase in gender diversity or fame of production decreases the odds of winning. While it was not too surprising that we see this for gender diversity, we were expecting the opposite relationship for fame of production. For racial diversity of production, we see that an increase in racial production is associated with an increase in the log odds of winning an Oscar. This was a bit surprising to us, but is a good thing.

Conclusions and Discussion:

On the morning of January 15, 2015, upon the release of a list of all-white nominees for Oscars acting awards, activist April Reign picked up her phone and sent out a Tweet that would fundamentally alter the narrative of diversity and representation in Hollywood: “#OscarsSoWhite they asked to touch my hair” (Reign). By lunch, #OscarsSoWhite was trending worldwide, and the industry’s history of excluding not just non-white but also female, queer, and other historically marginalized groups, stories, and identities was placed under intense scruity (Ugwu)—the same scrutiny we set out to apply to the Best Picture nominated and winning films from 2010-2020.

Our statistical model shows that while there is some evidence of exclusion of underrepresented groups in the history of the Best Picture winners, the results aren’t as drastic as expected. While there is a significant negative association between the gender diversity of a film’s cast and the film’s probability of winning Best Picture over this 10-year period, there also seems to be a positive association between racial diversity of the production and the probability of winning Best Picture. Other gender or racial variables were not considered significant enough to be included in the model. However, it is also important to address the limitations of our model, which stem in large part from the process of our data collection. There are three main concerns: 1) Size, 2) Race, and 3) Fame.

1) As in the Data section, it was necessary to limit our sizes and definitions of cast and production. This creates potential issues as different films often have different numbers of more visible and/or influential cast and production members, and the small sample sizes mean that any unusually small or large measurements of gender/racial diversity had a greater ability to skew our model.

2) It was difficult to measure racial diversity due to the absence of a formal database for such personal information. In order to calculate these proportions, we had to rely on what these individuals have said publicly about their racial identities. As noted in a previous section, small numbers of Asian/Indigenous and Multiracial cast members forced us to collapse our racial categories into White, Black, and Non-Black BIPOC, which perpetuates the erasure of certain groups of people. Furthermore, we must also consider that when it comes to race, perception of a person’s race may be as, if not more, influential than the person’s actual race on decisions such as deciding what film should win Best Picture, so our definition of racial diversity does not perfectly simulate the impact of race on the members of the selection committee.

3) Fame is similarly difficult to quantify. While our definition (described in Data section) doesn’t take into account other factors of fame (social media following, individuals are part of a family of famous figures, etc.), we chose it largely due to its simplicity, our time constraints, and the likelihood of the selection committee possibly being biased toward actors and producers they had previously awarded.

Beyond our data collection, there were some issues with the assumptions of our model. Linearity for all the variables was difficult to assume, once again due to the small sample sizes. And due to the nature of these films, which are nominated and winners for a reason, our data inherently is not random or independent.

Due to these limitations, we hope that future research expands upon the foundation we have laid through broadening the scope of our data collection: increasing sample sizes, looking at perceived racial identity as well as self-reported racial identity, and considering various measurements of fame. Including films outside of the 2010-2020 period would also provide more data with which to examine other predictors of probability of winning Best Picture such as interactions between variables of interest like racial diversity and gender diversity. Perhaps most important of all is expanding the scope of our current definition of diversity to include other types of diversity such as LGBTQ+ diversity and neurodiversity.

References

Background:

1. Reign, April. “#OscarsSoWhite They Asked to Touch My Hair. 😠.” *Twitter*, Twitter, 15 Jan. 2015, twitter.com/ReignOfApril/status/555725291512168448.
2. Hunt, Darnell, and Ana-Christina Ramón. “Hollywood Diversity Report 2020.” *UCLA Social Sciences*, UCLA College Social Sciences, 2020, socialsciences.ucla.edu/wp-content/uploads/2020/02/UCLA-Hollywood-Diversity-Report-2020-Film-2-6-2020.pdf.
3. Rottenberg, Josh, and Stacy Perman. “Who Really Gives out the Golden Globes? A Tiny Group Full of Quirky Characters - and No Black Members.” *Los Angeles Times*, Los Angeles Times, 21 Feb. 2021, www.latimes.com/entertainment-arts/business/story/2021-02-21/hfpa-golden-globes-2021-who-are-the-members.
4. Stevens, Matt. “Chloé Zhao Becomes the First Asian Woman to Win the Golden Globe for Best Director.” *The New York Times*, The New York Times, 1 Mar. 2021, www.nytimes.com/2021/02/28/movies/chloe-zhao-asian-director.html.
5. Ugwu, Reggie. “The Hashtag That Changed the Oscars: An Oral History.” *The New York Times*, The New York Times, 6 Feb. 2020, www.nytimes.com/2020/02/06/movies/oscarssowhite-history.html.
6. Yuen, Nancy Wang. “What Hollywood's Treatment of 'Minari' Says about the Asian American Dream.” *NBCNews.com*, NBCUniversal News Group, 4 Mar. 2021, www.nbcnews.com/think/opinion/what-hollywood-s-treatment-minari-says-about-asian-american-dream-ncna1259482.

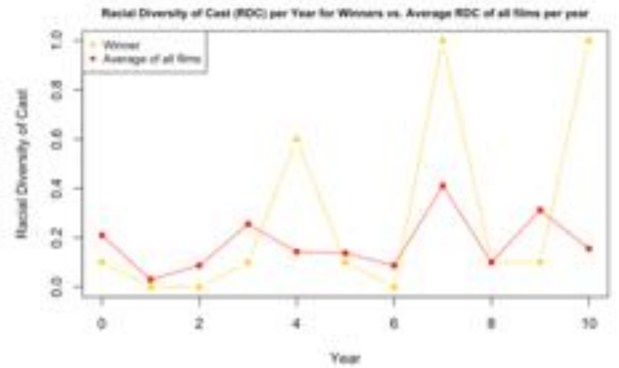
Data:

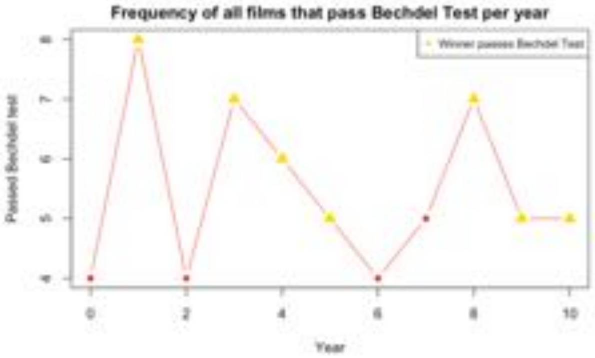
1. “Academy Awards Search.” *Academy Awards Database*, awardsdatabase.oscars.org/?search=Basic.
2. *Box Office Mojo*, www.boxofficemojo.com/.
3. *IMDb*, IMDb.com, www.imdb.com/.
4. *The Movie Database (TMDb)*, www.themoviedb.org/?language=en-US.
5. “Movie List.” *Bechdel Test Movie List*, bechdeltest.com/?list=all.
6. “Movies: TV Shows: Movie Trailers: Reviews.” *Rotten Tomatoes*, www.rottentomatoes.com/.

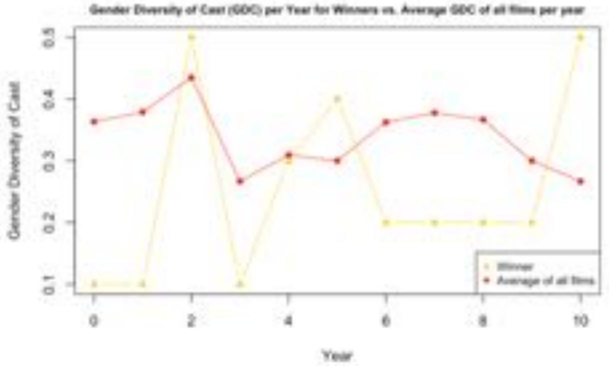
Appendices

Appendix A

Variables over Time



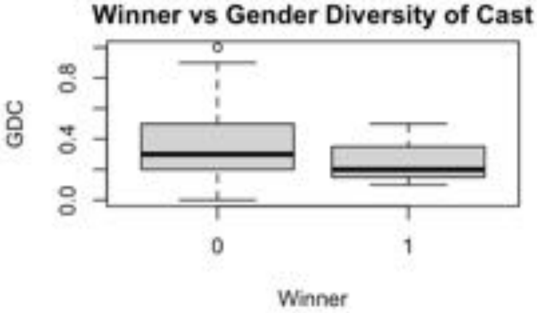


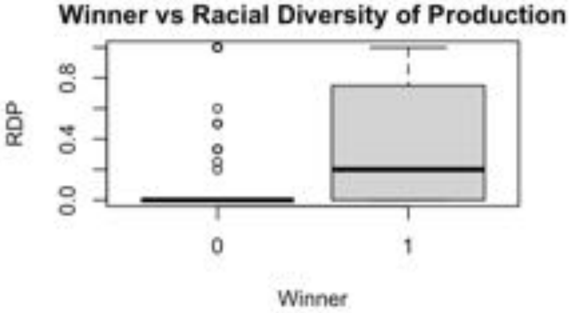


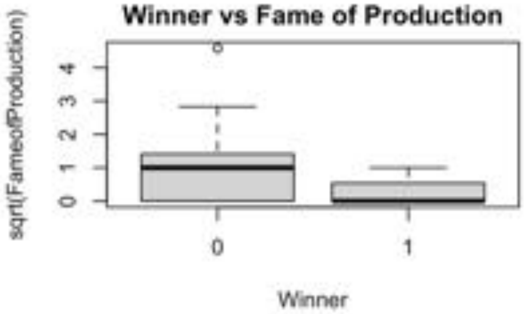
These graphs depict trends in the racial diversity of cast, gender diversity of cast, and passage of Bechdel Test of films nominated for Best Picture from 2010-2020. While the points are generally too variable to draw concrete inferences about how representation in nominated Best Picture films has changed over time, there seems to be a general trend of decreasing gender diversity of cast whereas racial diversity of cast appears approximately the same over time. Meanwhile, Best Picture film winners appear to be growing more racially and gender diverse over time.

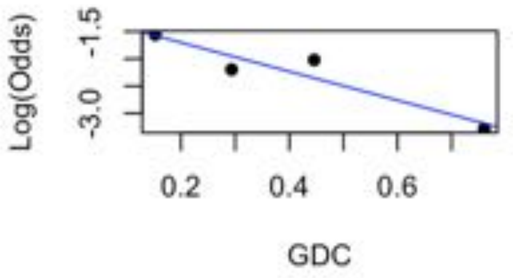
Appendix B

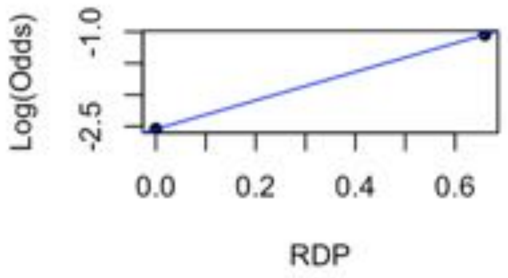
Initial model diagnostics

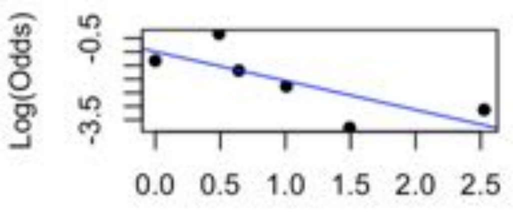












These graphs depict the Exploratory Data Analysis (EDA) and initial model diagnostics for the three predictors that were ultimately included in our model. For EDA, while the boxplots tend to show significant overlap, this can be attributable to the small sample size of winning films as compared to losing films. For the model diagnostics, the Empirical Logit Plots of the final variables show a pretty linear relationship, meaning that the final model variables seem to fulfill the linearity assumption.

Appendix C

Summary of Data	Estimate	Std. error	Z-value	P-value
Intercept	0.276	0.819	0.337	0.736
Gender Diversity of Cast	-5.347	2.273	-2.353	0.019
Racial Diversity of Production	2.348	1.052	2.233	0.026
Fame of Production$^{1/2}$	-2.349	0.835	0.005	0.005

The summary of our final model shows the significance of each of the variables as well as associations for each variable: a negative association between gender diversity of cast and log(odds) of winning, a positive association between racial diversity of production and log(odds) of winning, and a negative association between sqrt(fame of production) and log(odds) of winning.

Prediction Table	Actual Loss	Actual Win	Total
Predicted Loss	86	10	96
Predicted Win	1	1	2
Total	87	11	98

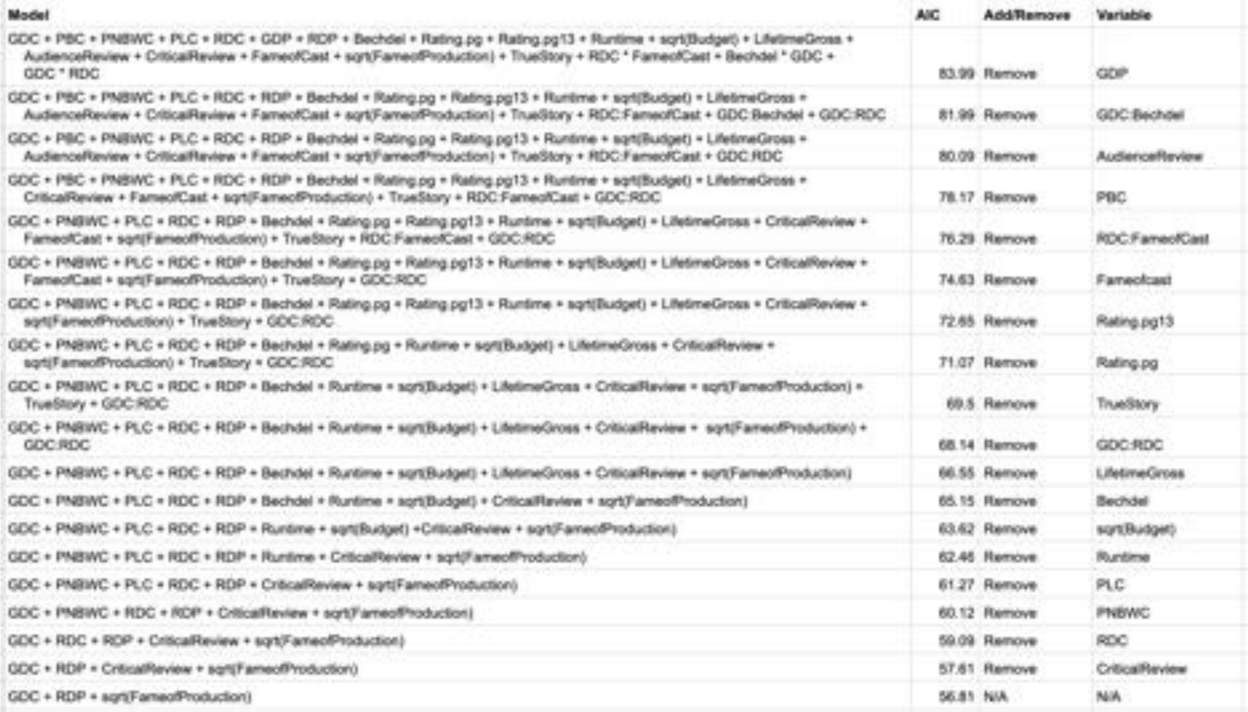
This table shows the ability of our model to correctly predict the outcome of the films in our dataset 88.78% of the time.

Exponential 95% Confidence Intervals	2.5%	97.5%
Gender Diversity of Cast	.00002862	0.254
Racial Diversity of Production	1.397	98.908
Fame of Production$^{0.5}$	.01368	0.387

The above confidence intervals confirm the negative and positive associations drawn from the estimates, with none of the intervals containing 0.

Appendix D

Stepwise Selection done by R



Model	AIC	Add/Remove	Variable
GDC + PBC + PNBWC + PLC + RDC + GDP + RDP + Bechdel + Rating.pg + Rating.pg13 + Runtime + sqrt(Budget) + LifetimeGross + AudienceReview + CriticalReview + FameofCast + sqrt(FameofProduction) + TrueStory + RDC * FameofCast + Bechdel * GDC + GDC * RDC	83.99	Remove	GDP
GDC + PBC + PNBWC + PLC + RDC + RDP + Bechdel + Rating.pg + Rating.pg13 + Runtime + sqrt(Budget) + LifetimeGross + AudienceReview + CriticalReview + FameofCast + sqrt(FameofProduction) + TrueStory + RDC:FameofCast + GDC:Bechdel + GDC:RDC	81.99	Remove	GDC:Bechdel
GDC + PBC + PNBWC + PLC + RDC + RDP + Bechdel + Rating.pg + Rating.pg13 + Runtime + sqrt(Budget) + LifetimeGross + AudienceReview + CriticalReview + FameofCast + sqrt(FameofProduction) + TrueStory + RDC:FameofCast + GDC:RDC	80.09	Remove	AudienceReview
GDC + PBC + PNBWC + PLC + RDC + RDP + Bechdel + Rating.pg + Rating.pg13 + Runtime + sqrt(Budget) + LifetimeGross + CriticalReview + FameofCast + sqrt(FameofProduction) + TrueStory + RDC:FameofCast + GDC:RDC	78.17	Remove	PBC
GDC + PNBWC + PLC + RDC + RDP + Bechdel + Rating.pg + Rating.pg13 + Runtime + sqrt(Budget) + LifetimeGross + CriticalReview + FameofCast + sqrt(FameofProduction) + TrueStory + RDC:FameofCast + GDC:RDC	76.29	Remove	RDC:FameofCast
GDC + PNBWC + PLC + RDC + RDP + Bechdel + Rating.pg + Rating.pg13 + Runtime + sqrt(Budget) + LifetimeGross + CriticalReview + FameofCast + sqrt(FameofProduction) + TrueStory + GDC:RDC	74.63	Remove	Fameofcast
GDC + PNBWC + PLC + RDC + RDP + Bechdel + Rating.pg + Rating.pg13 + Runtime + sqrt(Budget) + LifetimeGross + CriticalReview + sqrt(FameofProduction) + TrueStory + GDC:RDC	72.65	Remove	Rating.pg13
GDC + PNBWC + PLC + RDC + RDP + Bechdel + Rating.pg + Runtime + sqrt(Budget) + LifetimeGross + CriticalReview + sqrt(FameofProduction) + TrueStory + GDC:RDC	71.07	Remove	Rating.pg
GDC + PNBWC + PLC + RDC + RDP + Bechdel + Runtime + sqrt(Budget) + LifetimeGross + CriticalReview + sqrt(FameofProduction) + TrueStory + GDC:RDC	69.5	Remove	TrueStory
GDC + PNBWC + PLC + RDC + RDP + Bechdel + Runtime + sqrt(Budget) + LifetimeGross + CriticalReview + sqrt(FameofProduction) + GDC:RDC	68.14	Remove	GDC:RDC
GDC + PNBWC + PLC + RDC + RDP + Bechdel + Runtime + sqrt(Budget) + LifetimeGross + CriticalReview + sqrt(FameofProduction)	66.55	Remove	LifetimeGross
GDC + PNBWC + PLC + RDC + RDP + Bechdel + Runtime + sqrt(Budget) + CriticalReview + sqrt(FameofProduction)	65.15	Remove	Bechdel
GDC + PNBWC + PLC + RDC + RDP + Runtime + sqrt(Budget) +CriticalReview + sqrt(FameofProduction)	63.62	Remove	sqrt(Budget)
GDC + PNBWC + PLC + RDC + RDP + Runtime + CriticalReview + sqrt(FameofProduction)	62.46	Remove	Runtime
GDC + PNBWC + PLC + RDC + RDP + CriticalReview + sqrt(FameofProduction)	61.27	Remove	PLC
GDC + PNBWC + RDC + RDP + CriticalReview + sqrt(FameofProduction)	60.12	Remove	PNBWC
GDC + RDC + RDP + CriticalReview + sqrt(FameofProduction)	59.09	Remove	RDC
GDC + RDP + CriticalReview + sqrt(FameofProduction)	57.61	Remove	CriticalReview
GDC + RDP + sqrt(FameofProduction)	56.81	N/A	N/A

This shows the stepwise selection done by R to get the final model. It removed variables at each step to end up with the final model.