A Work Project, presented as part of the requirements for the Award of a Master's Degree in Economics from the NOVA – School of Business and Economics.

The Big Four: discrete choice modelling to predict the four major Oscar categories

Pedro Neves Barriga Afonso

Student Number 3099

A thesis carried out under the supervision of:

Professor Paulo M. M. Rodrigues

Lisbon, January 3rd, 2018

The Big Four: discrete choice modelling to predict the four major Oscar categories

ABSTRACT

The present study formulates regression models that predict the four major Oscar categories (Picture, Director, Actor and Actress). A database was created, collecting publicly available information from 2005 to 2016. The approach taken was to apply discrete choice modelling. A remarkable predictive accuracy was achieved, as every single Oscar winner was correctly predicted. The study found evidence of the crucial role of directors, the predictive power of box office, gender discrepancies in the film industry and the Academy's biases in the selection of winners related to the film genre, nominees' body of work and the portrayal of actual events.

Keywords: Binary choice models, Prediction, Oscars, Cinema.

ACKNOWLEDGMENTS

I would like to thank my adviser, Professor Paulo M. M. Rodrigues, for granting my wish to pursue this particular thesis – which has become the achievement I am most proud of.

I dedicate this work to my best friends – my precious – who encouraged this idea ever since its inception, and to my family, especially my brother Miguel Allen who was key in awakening my love for cinema. Here's looking at you, kiddo.

May the force all of you have given me be an input on this thesis.

1. INTRODUCTION

Films, subjective as they may be, are more mathematical than one would think. From a film's budget to its earnings at the box office, from marketing costs to awards received, there is maths and statistics in every element of the business that is the film industry. As cinema is an art loved by everyone, there is always curiosity on which film is crowned the year's best – especially when renowned film awards are being discussed.

The Oscar is the most prestigious award in the American film industry and one of the most important in cinema, as Hollywood is the oldest film industry in the world. They are awarded once a year by the Academy of Motion Picture Arts and Sciences – usually referred to as the Academy – in February/March at a live ceremony, which is one of the most watched television broadcasts every year. From a total of 24 categories, the Best Picture Oscar is the most prized award, thus the evening's last one to be revealed.[1]

An Oscar win greatly impacts the recipient's future earnings, the quality of films they star in, their recognition, fame and creative power over their subsequent projects. Although virtually anyone can try and guess which film is going to reap these substantial benefits, there is very little empirical evidence on regression models for predicting the Oscars.

The purpose of this study is to formulate regression models that predict the four major Oscar categories – Picture, Director, Actor and Actress – henceforth named the Big Four. This is a reference to the commonly named Big Five, which are the five most important Oscar categories: the above four and Best Screenplay.

The Oscar ceremony takes place the year after the films were released in the US. For example, the 90th edition of the Oscars will take place in early 2018, but it will honour the films of 2017. Thus, a usual mix-up happens regarding the year being discussed. In this paper, whenever a year

[1] Past winners include *How Green Was My Valley* (1941), *All About Eve* (1950) and *The Godfather* (1972).

is mentioned, it refers to the year in which the films were released. For instance, the Best Picture winner of 2016 was *Moonlight* (not *Spotlight*, the 2015 film that won Best Picture in early 2016). The Academy members are the sole voters for the Oscars. Becoming a member is by invitation only. As of January 2017, the Academy was composed of 6687 members. Every year around the month of June, new members are invited. The year of 2017 saw a record 774 invitees.[2]

An outcome can be explained by different data sources: polling data, fundamental data, and prediction market data, to name the main ones. Polling is not an option, as Academy members cannot be polled on which films they are going to vote for. Hence, this study focused on fundamental data, as it has been recently disregarded and less preferred to prediction market data. Fundamental data relies on past results and variables considered possible indicators. Thus, Oscar winners were predicted using publicly available information that is believed to be representative of the Academy's preferences, such as film genre, a film's number of Oscar nominations, nominees' previous Oscar wins and nominations, and precursor awards (i.e. awards that precede the Oscar ceremony). There are mainly two types of film awards. There are critics' groups awards and there are industry awards, such as the Oscars, the BAFTA's, the Golden Globes (GG) and guild awards (such as the PGA, the DGA, and the SAG).

This paper aims at improving the predictive accuracy beyond that of the existing literature. By finding a way to accurately predict the members' vote preferences, one is able to identify any biases the Academy has when it is assessing outstanding achievement in cinema.

The structure of the study is as follows. Section 2 is the literature review. Section 3 describes how the data was collected, which variables were used, and the empirical approach. Section 4 presents the final results. Section 5 is the discussion in relation to the literature. Section 6 is the conclusion, including limitations of the paper and potential improvements.

[2] The invitees included actress Charlotte Gainsbourg, director Pedro Costa and cinematographer Linus Sandgren.

2. LITERATURE REVIEW

This paper builds on the existing literature. While there is little research on regression models using fundamental data for predicting the Oscars – the matter of this study – there is a great deal on factors associated with film awards and noms. Predictive modelling will be addressed first, in which two publications stand out: Pardoe & Simonton (2008) and Kaplan (2006).

Pardoe (2005) was the first to present an analysis whose purpose was to predict Oscar outcomes. Proceeding from his published work in 2005 and 2007, Pardoe & Simonton (2008) is the paper closest to the current study, as both create models aimed at predicting the Big Four. Their data ranged from 1938 to 2006, a span so large that allowed them to focus their analysis on the impact that each regressor has had on the likelihood of a film winning an Oscar over the decades. Their predictor variables included: total Oscar nominations, nominees' previous Oscar nominations and wins, genre, film length, release date, critics' ratings and a series of dummy variables (each guild award, GG Drama winner, GG Comedy winner and whether the film was nominated for the Best Director Oscar or not). Due to struggles in obtaining data on critics' ratings for the early years, this variable was excluded altogether. They pondered two different modelling methods – the multinomial logit model (MNL) and the mixed logit model – and two different estimation methods – Bayesian estimation and maximum likelihood. The MNL model was chosen and estimated using maximum likelihood. The study arrived at many interesting results: 1) "film length" and "release date" were found to worsen predictions; 2) receiving a Best Director nomination has become increasingly important for the winning chances of Best Picture nominees; 3) previous Oscar nominations have become less relevant for Best Actor nominees; 4) previous Oscar wins have been increasingly hurting the winning chances of Best Actor nominees since the 1970s; 5) previous Oscar wins have become less relevant for Best Actress nominees; 6) comedies and musicals prevailed over dramas as Best Picture winners in the 1960s and 1970s. In the

following decades however, tables turned and dramas have become predominant; 7) Best Actor wins have always favoured dramatic performances over humorous roles; 8) the DGA and PGA were more accurate predictors of Best Director than they were of Best Picture. Across all studied decades, Best Director was always the most predictable category, whereas Best Actress the hardest to forecast. In the earlier years, prediction accuracy was low, due to unavailable data. The success rate for the 1977-2006 period was 79% (95 out of 120 winners): 70% for Best Picture, 93% for Best Director, 77% for Best Actor and 77% for Best Actress.

Kaplan (2006) focused on predicting the Best Picture Oscar. The data ranged from 1965 to 2004, for a total of 200 nominated films. Most explanatory variables were divided into three categories: Personnel, Genre and Marketing. Personnel included nominees' previous Oscar nominations and wins, and the number of previous Best Actor Oscar wins between the cast. The dummy variables for a film's genre were drama, comedy, musical, biographic and epic. Epics are films characterised by ambitious production design and striking visual style.[3] Marketing comprised a film's length, release date (equal to 1 if released in the fourth quarter of the year), and screenplay (equal to 1 if it was adapted from another source). The other independent variables were previous awards dummies, such as the GG and DGA, and Oscars related variables, such as "MostNom" (equal to 1 if the film was the most Oscar nominated nominee). Kaplan built several models: the first one using every independent variable he had gathered, whereas the following ones were obtained dropping insignificant variables. His final Best Picture model had nine regressors. The most statistically significant ones were "MostNom", GG Drama winner, "EpicBiop" (equal to 1 if the film was an epic and biographical), and DGA winner. All four parameter estimates were positive, meaning the best candidate for winning Best Picture is a multiple-Oscar-nominated epic biographical drama film, helmed by a great director.

[3] Epics that have won Best Picture include *Gone with the Wind* (1939), *Lawrence of Arabia* (1962), *Titanic* (1997).

Moving on to studies on the drivers of the success of a film, Dean K. Simonton is the leading authority. He is a pioneer and a prolific author in the field of cinematic success. First-rate directors were found to have a significant impact on a film's odds of winning Best Picture (Simonton, 2002), which is consistent with the finding that Best Director nominations correlate highly with Best Picture wins (Simonton, 2004b). One of the most interesting results is "the Best Actress Paradox" (Simonton, 2004a): there is a substantial gender-based discrepancy in the film industry, as a remarkable performance by a woman is less likely to be associated with high quality films than one by a man. A film led (or even supported by) a male performance is more likely to win Best Picture than one led by a woman. Simonton (2009) measured a film's success by awards received, critics' ratings and box office. The predictors for receiving film awards were divided into two categories: Production and Distribution. The former focussed on predictors related to the making of the film: if it was based on true events, the genre, and the budget. The latter focussed on the film's season of release and advertising costs. A film that portrays a true story is more likely to earn nominations and wins for the Big Four, and so is a drama film. Budget was found to have a zero correlation with nominations and wins for the Big Four. Films released during Christmas time are more likely to earn nominations in the Big Four than summer releases. Information on advertising costs is rarely available, hence, Simonton called attention to a unique study by Prag and Casavant (1994) that managed to acquire data on advertising for 195 films and concluded that film awards were positively correlated with marketing expenses.

3. METHODOLOGY

3.1. DATA

The initial idea was to look for film databases that would already include all the cross-sectional data necessary to estimate the four models. However, of the databases available online, none suited this study's aim. The databases found from reliable sources either had data on each year's

Oscar winners only (no data on the losing nominees), or did not contain most of the explanatory variables that had been envisioned. Therefore, building a new dataset from the ground up was the only method of fulfilling what this study had set out to do. Finding data was fairly simple, as all the necessary data was available on the Internet. What was most challenging was the tremendous amount of time that had to be put in to create the four datasets – one for each of the Big Four. For each dataset, information was collected for every previous winner and losing nominee from recent years. As this study's aim was to predict the eventual winners, only information that was available before each Oscar ceremony was eligible to be collected. Around eighteen independent variables were collected for each category, which summed up to a total of 4661 data points.

The data sources were Wikipedia, IMDb, Metacritic, Rotten Tomatoes and Box Office Mojo.[4]

Regarding the data range, a sufficient number of observations had to be reached so that the models yielded satisfactory predictions. Conversely, collecting too many years of data was not desirable for two reasons: first, some statistics have only been available in recent years, so it would be impossible to collect them for earlier years; second, the further away from the present date, the more irrelevant the data becomes, because the film industry changes a lot. One easy way to witness that is how international film festivals have recently transformed the industry.

The Telluride and the Venice Film Festivals have become main stages for world film premieres in the past decade: 8 out of the 9 most recent Best Picture winners premiered at one of these two festivals – the sole exception is *The Artist* which premiered at the 2011 Cannes Film Festival. Before 2008, no Best Picture winner had had its world premiere in either of them. In 2008, *Slumdog Millionaire* premiered at Telluride and would go on to win Best Picture at the Oscars. Due to this turning point for the industry, 2008 was chosen as the starting year, resulting in nine

[4] Peter Gloor, a research scientist at MIT Sloan School of Management, has conducted informal analyses of the Oscars and assured that Wikipedia and IMBb (where virtually anyone can go and post false information) are remarkably reliable online sources, as they are constantly monitored.

years of data that sum up to 77 observations (the category allows for 5 to 10 nominees per year). So as to follow the same logic, data for the other categories (Director, Actor and Actress) was collected also for the past nine years. However, as each of these categories only allows for 5 nominees, that summed up to 45 observations. As that was deemed not sufficient, the number of years was extended to twelve to reach a total of 60 observations. So, for Best Picture (2008-2016) we have 9 winners out of 77 nominees (68 zeros and 9 ones), whereas for Best Director, Actor and Actress (2005-2016) we have 12 winners out of 60 nominees (48 zeros and 12 ones).

As noted in the literature review, the following independent variables have been found to be relevant for predicting Oscar outcomes: a film's total Oscar nominations, genre[5], a Best Director nomination (for predicting Best Picture), nominees' previous Oscar wins and nominations, if based on true events or not, the guild awards (DGA, PGA and SAG) and the GG.

While Pardoe & Simonton (2008) were prevented from using critics' ratings as an explanatory variable due to unavailable data, that is no longer an issue, given the current paper's data range. Data was collected from the two most famous review aggregation websites, Rotten Tomatoes and Metacritic, so as to study which of the two is more predictive in each of the Big Four.

Besides collecting data for these predictor variables, this study proposes several new variables not present in the existing literature thus far. For each of the Big Four, data was collected on:

$BigFiveNoms_i$ – a film's number of Oscar nominations in the Big Five. A nominee might have better chances of winning the Oscar, if the Academy nominated the film in many of the Big Five.

$Wins_i$ – total competitive awards the nominee won for the nominated film. For Best Picture, every award the nominated film won regardless of the field (film editing, sound mixing, and so on). This regressor is a novelty, as in the literature only the most important awards have been

[5] Criterion for $Genre_i$: equal to 1 for standard dramas or dramatic films with humorous content (*The Martian*), while 0 for dark comedies (*Birdman*), musicals (*La La Land*), comedy-dramas (*American Hustle*) and animated films.

considered, such as the GG and the guild awards, but never the full amount of accolades received.

PremiereDate$_i$ – a dummy variable equal to 1 if the film premiered in the second half of the year (July-Dec), 0 otherwise. It is key to distinguish premiere (the date the film was first presented to an audience) from release (the date the film opened in theatres). Only the latter is used in the literature. For instance, 2016 Best Picture nominee *Manchester by the Sea* premiered at Sundance in January, while it was only released in the US in November – a ten-month span.

MajorStudio$_i$ – a dummy equal to 1 if the film was US distributed by one of Hollywood's six major film studios, 0 otherwise. As noted in the previous section, data on advertising costs is rarely available. Thus, "Major Studio" was proposed as a proxy, given that only the major film studios can afford spending millions of dollars on advertising to promote their films.

BoxOffice$_i$ – domestic total gross in millions of US dollars up to the week before the Oscar ceremony. There is considerable research on predicting box office using Oscar wins as regressor. However, it does not account that this only applies to films that are still in theatres at the time of the Oscar ceremony. In 2015, *Mad Max: Fury Road* was released in the US in March. Nearly a year later in February 2016, the film won six Oscars (the most awarded film of the evening), but no box office boost came with it, as the film had long left US screens in September 2015. Thus, it raises the question of whether box office influences Oscar awards, not the other way round.

3.2. PRELIMINARY DATA ANALYSIS

Since 2008, no film was awarded in the four categories. The films that got the closest were 2010's *The King's Speech* and 2011's *The Artist*, both winning Best Picture, Director and Actor.

89% of Best Picture winners were nominated for Best Director, while only 53% of Best Picture losing nominees were. Only one Best Picture winner did not receive a Best Director nom: *Argo*.

The average number of Oscar nominations for Best Picture winners is 8,89, whereas that for losing nominees is 6,22. *Spotlight* won Best Picture with only 6 Oscar nominations, while *La La Land* lost Best Picture after having received a record-tying 14 nominations.

Regarding critics' ratings, *Birdman* was the Best Picture winner with the lowest Rotten Tomatoes score (91), whereas *Toy Story 3* and *Selma* were the losing nominees with the highest score (99).

Argo and *Slumdog Millionaire* were the Best Picture winners with the lowest Metacritic score (86), while *Boyhood* was the losing nominee with the highest score (100).

Out of the last 77 Best Picture nominees, only two comedies have won: *Birdman* and *The Artist*.

3.3. MODEL

This study's goal was to predict the eventual Oscar winners in each of the four major categories, only making use of fundamental data that was available before each year's Oscar ceremony.

The approach taken was to use discrete choice modelling. In the case of predicting the Big Four, the discrete choice problem takes the form of a binomial choice model, as there are only two available alternatives: equal to 1 if the nominee won, 0 otherwise. This dichotomous outcome variable can be modelled using appropriate regression models, such as logit or probit models. The probit model was chosen.

Using the econometric software package Stata, the parameters of the probit model were estimated by the method of maximum likelihood. The explicitly estimated model followed from the estimation of a latent variable y_i^*, which is not observed but can be interpreted as a propensity to have outcome $y_i = 1$. The latent variable model is specified as: $y_i^* = x_i'\beta + \varepsilon_i$.

By attributing values to the regressors, the equation will yield a value for the latent variable. In the probit model, the error term follows the standard normal distribution. Hence, the predicted probabilities are computed from the standard normal cumulative distribution function. For each

category, the predicted probability of winning the Oscar will be the area under the standard normal distribution curve that falls to the left of the latent variable's value.

In the probit model, the magnitude of its coefficients cannot be interpreted. Instead, one can interpret the marginal effects, which depend on the coefficient of the regressor in question and on the values of all regressors in the model. The sign of the coefficients can be interpreted though. A positive sign means that the probability of winning increases with added units of that variable.

Since what matters most in binary regression models are the signs of the regression coefficients and their statistical significance, the main concern was to ensure that all regressors were statistically significant. If the constant is not statistically different from zero and all the regressors in the model were to take the value zero, the latent variable would be zero. As the error term follows the normal distribution, the resulting probability of winning would be 50%. It would not be reasonable to report that a film, for instance, with a zero Metacritic score and zero wins in past award shows had a 50% probability of winning. Hence the importance of the constant to be statistically significant. The usual two-sided tests and a 5% significance level were used.

To ensure parsimony, a model was only allowed to have a maximum of six regressors. In order to correct for heteroskedasticity, robust standard errors were specified. To detect endogeneity, the residuals were taken and checked if they were correlated to any of the regressors in the model. Regarding possible omitted variable bias, a link test for model specification was performed.

Each model was tested for multicollinearity, by computing the variance inflation factor (VIF). As a rule of thumb, a variable whose VIF value is greater than 5 is worrisome. Informal methods of detecting multicollinearity were also considered.

By construction, there are several explanatory variables most likely correlated, such as $ActNoms_i$ and $ActWins_i$, $OscarNoms_i$ and $BigFiveNoms_i$, and $DirectorNom_i$ and $BigFiveNoms_i$. As such, it was ensured that no model included both variables from one of these pairs.

Additional pairs were also excluded from the model-building process, according to a chosen correlation threshold. It was decided that for values of correlation above 0,60 there was a worrisome correlation. For instance, the dummy variables "DGA" and "PGA" were highly correlated (r = 0,82). For that reason, they were never included in a model together.

Concerning goodness of fit, the McFadden's R^2 value was reported.

Finally, the model is sound and ready to be assessed for its predictive quality – the fundamental goal of this study. For each year, the nominees' predicted probabilities were compared to check whether the model was successful in attributing the highest probability to the actual winner.

This is how previous empirical research evaluates predictive accuracy, but it is also the simplest and most obvious way. It was decided to take it a step further, by using a more challenging way to assess predictive quality: command the model to determine if a nominated film is going to win, relying solely on its predicted probability, i.e., without comparing it to the other nominees' probabilities. For this purpose, a cut-off was established for each category. Only if a nominee's probability is above the cut-off, does the model pronounce it a winner. The difficulty now lies in the fact that if no nominee overcomes the cut-off, there is no winner. Likewise, if two nominees place above the cut-off, the model wrongly selects two winners. In a year when there is no clear frontrunner to win or when there are two favourites to win, only a great predictive model would be able to use this challenging method and still correctly predict the winner.

Another common way of evaluating predictive quality is to graph Receiver Operating Characteristic (ROC) curves. A ROC curve is a plot of the true positive rate versus the false positive rate, for every possible classification threshold. The better the classifier separates the two classes (winners from losing nominees), the bigger the area under the curve. The area is equal to 1 for a perfect classifier. Areas above 0,9 are considered excellent.

4. RESULTS

From the methodology that was described in section 3, the final models were obtained (Annex 1). There were several explanatory variables that were statistically significant when they were the single regressor in the model. In the Best Picture category, $DirectorNomDummy_i$ and $MajorStudio_i$ were statistically significant at a 5% significance level, whereas $OscarNoms_i$, PGA_i and $Metacritic_i$ were at a 1% level. In the Best Actor category, $RottenTomatoes_i$ and $Metacritic_i$ were statistically significant at a 5% level, while SAG_i and $Wins_i$ were at a 1% level. In the Best Actress category, $BigFiveNoms_i$ and $Semester_i$ were statistically significant at a 5% level, whereas SAG_i was at a 1% level. In the Best Director category, $OscarNoms_i$, $BAFTA_i$ and GG_i were statistically significant at a 1 % level. Thus, to a certain extent there is a level of correlation between each of these predictors and the corresponding outcome variable. However, when featured in the respective model alongside other more predictive regressors, they either lost statistical significance or worsened predictions.

4.1. BEST PICTURE

The final probit model was:

$$(1)\ Y_i^* = -41{,}15155 + 5{,}209567 * DGA_i + 0{,}4126603 * RottenTomatoes_i - \\ - 0{,}0319578 * BoxOffice_i + 0{,}0135294 * Wins_i$$

The regressors $RottenTomatoes_i$, $Wins_i$ and the constant term are statistically significant at a 5% significance level, while DGA_i and $BoxOffice_i$ are at a 1% level. The hypothesis that all coefficients are equal to zero was rejected at a 1% significance level (p-value equal to 0,0052).

The residuals were taken and they were not found to be correlated to any of the regressors (the highest value of correlation was 0,0603 with DGA_i). Hence, there is no evidence of endogeneity.

Regarding the link test, the variable of prediction *_hat* was statistically significant (p-value equal

to 0,03), but the variable of squared prediction *_hatsq* was not (p-value equal to 0,726), which are the desired results. Thus, the model is specified correctly and there is no omitted variable bias.

All variables have acceptable VIF values, as the highest VIF score is 4,00 ($Wins_i$). The mentioned informal methods were performed and none of them alerted for multicollinearity.

Joint significance was also checked for: all pairs of regressors were found to be jointly significant, obtaining test p-values ranging from 0,0011 to 0,0325.

DGA_i, $RottenTomatoes_i$ and $Wins_i$ have positive signs as expected, thus, the probability of winning Best Picture increases with added units of these variables. $BoxOffice_i$ has a negative sign, hence, the probability of winning Best Picture decreases with added units of this variable. This result may seem counter-intuitive, as one would expect a film that performed well at the box office to have an increased chance of winning. However, if one looks closely at the data, this result is no shocker: there is a vast split between top-grossing films and award-wining films. Since 2008, not a single film has won Best Picture while landing in the year-end US box-office Top 10. The film that got closest was *Slumdog Millionaire*, that ranked 16th in 2008, while the film that got furthest was *The Hurt Locker*, that ranked 116th in 2009 and is to date the lowest-grossing Best Picture winner ever. Analysing the box office of all Best Picture nominees from the past nine years, it becomes clear that box office has been looked at the wrong way (Annex 2).

According to the average marginal effects: winning the DGA award increases the probability of winning by 23,99 percentage points; one added unit to the Rotten Tomatoes score increases the probability by 1,90 percentage points; one added million of dollars earned in the US decreases the probability by 0,15 percentage points; one added award won increases the probability by 0,06 percentage points, ceteris paribus. Annex 3 is a comparison of Best Picture nominees that had similar values for two or three of the final model regressors. One notices how a slightly different value in one or two regressors can result in so disparate predicted probabilities.

Concerning goodness of fit, the McFadden's R^2 is 0,7591.

Regarding the predictive quality, the model was successful in attributing the highest probability to the actual winner, in every single year – an accuracy rate of 100% (Annex 4).

The predicted probabilities for the winning films range from 51,38% (*12 Years a Slave*, 2013) to 100,00% (*The Hurt Locker*, 2009), with the exception of one film: *Spotlight* in 2015 (10,93%). The average predicted probability for the winning films is 77,34%.

The predicted probabilities for the losing films range from 0,00% to 13,34%, with the exception of two films: *Boyhood* in 2014 (78,25%) and *La La Land* in 2016 (58,34%). The average predicted probability for the losing films is 2,97%.

Concerning the more challenging way to evaluate predictive accuracy, the cut-off that was best at separating the classes was 78,56%. As a result, 75 out of 77 Best Picture nominees were correctly identified – a prediction accuracy of 97,40%. The two wrong predictions were two Best Picture winners that scored below the cut-off, thus pronounced losers: *12 Years a Slave* and *Spotlight*.

The area under the ROC curve was 0,9918, which is excellent (Annex 5).

Interestingly enough, no variable specifically related to the film's cast is present in the model, whereas DGA_i – that concerns the film's director – is highly significant. This means that an acclaimed director is much more relevant for winning Best Picture than an acclaimed cast.

Concluding, the best candidate for winning the Best Picture Oscar is a low-grossing film, with a high Rotten Tomatoes score, that won the Director's Guild of America award, and also won numerous awards throughout the various categories of film, including any other honour received.

4.2. BEST ACTOR

$$(2)\ Y_i^* = -7{,}646065 + 0{,}4842008 * OscarNoms_i + 5{,}950575 * GG_i +$$

$$+ 3{,}396998 * True_i - 0{,}0517834 * BoxOffice_i + 0{,}771651 * ActNoms_i$$

All five regressors and the constant are statistically significant at a 1% significance level. The

hypothesis that all coefficients are equal to zero was rejected at a 1% level (p-value was 0,0000). The residuals were taken and they were not correlated to any of the regressors (the highest value of correlation was 0,0437 with $OscarNoms_i$). Hence, there is no evidence of endogeneity.

Regarding the link test, the variable of prediction *_hat* was statistically significant (p-value equal to 0,002), but the variable of squared prediction *_hatsq* was not (p-value of 0,303), which are the desired results. Thus, the model is specified correctly and there is no omitted variable bias.

All variables have acceptable VIF values, as the highest VIF score is 3,51 (for $OscarNoms_i$). The mentioned informal methods were performed and none of them alerted for multicollinearity.

Joint significance was also checked for: all pairs of regressors were found to be jointly significant, obtaining test p-values ranging from 0,0000 to 0,0002.

$OscarNoms_i$ and GG_i have the expected positive signs, thus, the probability of winning Best Actor increases with added units of these variables. However, $True_i$ and $ActNoms_i$ having positive signs merits remarks, as it denotes biases of the Academy. The former highlights the favouritism of voters towards actors portraying historical figures: Phillip Seymour Hoffman was Truman Capote, Sean Penn was Harvey Milk, Colin Firth was King George VI, Daniel Day-Lewis was Abraham Lincoln, Eddie Redmayne was Stephen Hawking, and the list goes on. The latter goes to show the Academy does not award Best Actor solely based on the quality of the performances, but also on how overdue of an Oscar an actor is. Most recently, on his fifth Oscar nomination, Leonardo DiCaprio won Best Actor for *The Revenant* – an award arguably more reflective of his body of work, than of his latest performance.

$BoxOffice_i$ has a negative sign, hence, the probability of winning Best Actor decreases with added units of this variable – the same result found in the Best Picture category.

According to the average marginal effects: one added Oscar nomination increases the probability

of winning by 2,08 percentage points; winning the Golden Globe (in either Drama or Comedy) increases the probability by 25,51 percentage points; portraying a true character increases the probability by 14,56 percentage points; one added million of dollars earned in the US decreases the probability by 0,22 percentage points; one added previous Oscar nom increases the probability of winning by 3,31 percentage points, ceteris paribus.

Concerning goodness of fit, the McFadden's R^2 is 0,8325.

Regarding the predictive quality, the model was successful in attributing the highest probability to the actual winner, in every single year – an accuracy rate of 100% (Annex 6).

The predicted probabilities for the winning films range from 83,88% (*The Revenant*, 2015) to 99,95% (*Dallas Buyers Club*, 2013), with the exception of one film: *Manchester by the Sea* in 2016 (11,17%). The average predicted probability for the winning films is 88,30%.

The predicted probabilities for the losing films range from 0,00% to 7,89%, with the exception of two films: *The Wolf of Wall Street* in 2013 (39,47%) and *Birdman* in 2014 (75,99%). The average probability for the losing films is 3,14%.

The cut-off that was best at separating the classes was 79,94%. As a result, 59 out of 60 Best Actor nominees were correctly identified – a prediction accuracy of 98,33%. A Best Actor winner scored below the cut-off, thus pronounced loser: Casey Affleck in *Manchester by the Sea*.

The area under the ROC curve was 0,9965, which is excellent (Annex 7).

Concluding, the best candidate for winning the Best Actor Oscar is a low-grossing film, that portrays a true story, was highly Oscar nominated, and whose leading actor won the Golden Globe (in Drama or Comedy) and has been nominated for the Oscar several times in the past.

4.3. BEST ACTRESS

$$(3)\ Y_i^* = 3{,}650643 + 0{,}8125794 * OscarNoms_i + 0{,}2709599 * Wins_i -$$

$$- 0{,}1747256 * Metacritic_i + 2{,}886029 * GGDrama_i + 0{,}0076872 * BoxOffice_i$$

The regressor $BoxOffice_i$ and the constant term are statistically significant at a 5% significance level, while $OscarNoms_i$, $Wins_i$, $Metacritic_i$ and $GGDrama_i$ are at a 1% level. The hypothesis that all coefficients are equal to zero was rejected at a 1% level (p-value equal to 0,0010).

The residuals were taken and they were not correlated to any of the regressors (the highest value of correlation was 0,0222 with $GGDrama_i$). Hence, there is no evidence of endogeneity.

Regarding the link test, the variable of prediction *_hat* was statistically significant (p-value equal to 0,002), but the variable of squared prediction *_hatsq* was not (p-value of 0,578), which are the desired results. Thus, the model is specified correctly and there is no omitted variable bias.

All variables have acceptable VIF values, as the highest VIF score is 4,26 ($OscarNoms_i$). The mentioned informal methods were all performed and none of them alerted for multicollinearity.

Joint significance was also checked for: all pairs of regressors were found to be jointly significant, obtaining test p-values ranging from 0,0001 to 0,0018.

$OscarNoms_i$ and $Wins_i$ have positive signs as expected, thus, the probability of winning Best Actress increases with added units of these variables. The remaining regressors merit comment.

GGDrama$_i$ has a positive sign, which means the probability of winning increases with winning the GG Drama, but not the GG Comedy. This drama-preferred-to-comedy result is akin to the portraying-true-characters bias we found in Best Actor, as biopics are typically dramas.

Metacritic$_i$ has a negative sign, thus, the probability of winning decreases with added units of this variable. This is not a bias of the Academy. Voters most certainly do not prefer awarding an actress that stars in a bad film over one in a good film – if anything, the opposite would make some sense. This is proof that women star in lower quality films compared to men: in the past twelve years, five Best Actress winners have won the Oscar for a film not nominated for Best Picture, whereas only two Best Actor winners have. It should also be noted that 2014 is the

most recent year this has happened for women (Julianne Moore for *Still Alice*), while for men one would need to go further back to 2009 (Jeff Bridges for *Crazy Heart*).

This gender discrepancy is further underlined when one notices that the Best Actress and the Best Actor Oscars not only have different explanatory variables ($Wins_i$ and $Metacritic_i$; $True_i$ and $ActNoms_i$), but one of the shared regressors has a different sign. $BoxOffice_i$ here has a positive sign, thus, the probability of winning increases with added units of this variable.

According to the average marginal effects: one added Oscar nomination increases the probability of winning by 3,69 percentage points; one added competitive acting award won increases the probability by 1,23 percentage points; one added unit to the Metacritic score decreases the probability by 0,79 percentage points; winning the GG Drama increases the probability by 13,10 percentage points; one added million of dollars earned in the US increases the probability by 0,03 percentage points, ceteris paribus.

Concerning goodness of fit, the McFadden's R^2 is 0,8259.

Regarding the predictive quality, the model was successful in attributing the highest probability to the actual winner, in every single year – an accuracy rate of 100% (Annex 8).

The predicted probabilities for the winning films range from 73,10% (*Silver Linings Playbook*, 2012) to 100,00% (*The Queen*, 2006), with the exception of one film: *Still Alice* in 2014 (15,33%). The average predicted probability for the winning films is 87,03%.

The predicted probabilities for the losing films range from 0,00% to 11,49%, with the exception of one film: *Transamerica* in 2005 (84,96%). The average predicted probability for the losing films is 3,10%.

The cut-off that was best at separating the classes was 13,41%. As a result, 59 out of 60 Best Actress nominees were correctly identified, corresponding to a prediction accuracy of 98,33%.

The single wrong prediction was a losing nominee that scored above the cut-off, thus pronounced winner: Felicity Huffman in *Transamerica*.

The area under the ROC curve was 0,9948, which is excellent (Annex 9).

Concluding, the best candidate for winning the Best Actress Oscar is a film that was highly Oscar nominated, with a low Metacritic score, a success at the box office, and whose leading actress won the Golden Globe Drama and many other awards for her performance.

4.4. BEST DIRECTOR

The Best Director model building-process was the hardest. The best predictor, DGA_i, was dropped by Stata, as DGA_i equal to 1 predicted success perfectly. Since 2008, if the director that won the DGA award was also nominated for the Best Director Oscar, that director would always win the Oscar. No exceptions. The models that resulted from working only with the remaining explanatory variables were not good enough. The best one, which had as predictor an interaction between $Wins_i$ and $DGAnom_i$ (equal to 1 if nominated for the DGA award), had an accuracy rate of 75%. It became clear that excluding DGA_i from the final model would never produce great results, thus, an index was created as the solution. Several indices containing two or more independent variables were tested, until a three-variable index was found that managed to assign the highest probability to the actual winner, in every single year – an accuracy rate of 100% (Annex 10). Those three variables were DGA_i, $Wins_i$ and $DGAnom_i$. Regarding each variable's weight on the index, Microsoft Office Excel's Goal Seek feature created several sets of weights that achieved a prediction accuracy of 100%. Each set was run on Stata and the one that yielded the best ROC curve was chosen:

$$(4)\ Index_i = 0{,}96 * DGA_i + 0{,}03 * Wins_i + 0{,}01 * DGAnom_i$$

The final probit model was:

$$(5)\ Y_i^* = -2{,}511019 + 2{,}845288 * Index_i$$

The regressor $Index_i$ and the constant term are statistically significant at a 1% significance level. The hypothesis that all coefficients are equal to zero is rejected at a 1% significance level (p-value equal to 0,0000).

The residuals were taken and they were not correlated to $Index_i$ (r = 0,1560).

Regarding the link test for model specification, the variable of prediction *_hat* was statistically significant (p-value equal to 0,05), but the variable of squared prediction *_hatsq* was not (p-value of 0,152), which are the desired results.

Concerning multicollinearity, no pair of independent variables was highly correlated.

$Index_i$ and all the regressors it is composed of (DGA_i, $DGAnom_i$ and $Wins_i$) have positive signs, thus, the probability of winning Best Director increases with added units of these variables.

Concerning goodness of fit, the McFadden's R^2 is 0,7327.

The predicted probabilities for the winning films range from 63,09% (*The King's Speech*, 2010) to 99,96% (*The Hurt Locker*, 2009), with the exception of one film: *Life of Pi* in 2012 (1,30%). The average predicted probability for the winning films is 84,13%.

The predicted probabilities for the losing films range from 0,60% to 53,12% (*Boyhood*, 2014). The average predicted probability for the losing films is 4,58%.

The cut-off that was best at separating the classes was 58,11%. As a result, 59 out of 60 Best Director nominees were correctly identified, corresponding to a prediction accuracy of 98,33%. The single wrong prediction was a Best Director winner that scored below the cut-off, thus pronounced loser: Ang Lee for directing *Life of Pi*.

The area under the ROC curve was 0,9601, which is excellent (Annex 11).

Concluding, the best candidate for winning the Best Director Oscar is a film whose director won the Director's Guild of America award and also many other awards for directing the film.

5. DISCUSSION

It is imperative to discuss the present study's results in relation to the reviewed literature.

Pardoe & Simonton (2008) found that receiving a Best Director nomination had become extremely important for Best Picture nominees. The current study found it to be statistically significant at a 5% level, but it was not part of the final Best Picture model. They established previous Oscar wins lost relevance for Best Actress nominees, which is supported by this paper, as no significant relationship was found. They found that for an actor the number of Oscar nominations he had had no longer helped his chances of winning as much, while his Oscar wins more and more hurt his chances. Here lies the first major difference: this study found no relationship between past Oscar wins and chances of winning, but also that previous Oscar noms are of help (the average marginal effect is 3,31%) and statistically significant at a 1% level. Another interesting disparity is that they found the Academy has always favoured actors in dramatic roles over actors in comedic roles, whereas this paper found that drama-preferred-to-comedy relationship not for actors, but for actresses at a 1% significance level.

Every year, Pardoe updates his models with the new data. Then he runs the updated model, before posting the new predictions on his website. He correctly predicted 6 of the last 9 Best Picture winners. In the last 12 years, he correctly predicted 10 Best Actress winners, 11 Best Actor winners and 11 Best Director winners. Hence, a total of 38 out of 45 were correct.

Kaplan (2006) found that the likelihood of winning Best Picture is greatest for the most Oscar nominated nominee, for the winner of GG Drama, for the winner of the DGA award and for an epic biographical film. The current study found the number of Oscar nominations to be statistically significant at a 1% level, but it was not included in the final model. No relationship was found regarding the GG. The relevance of the DGA award was supported by this paper, as it was found to be statistically significant at a 1% level and reported an average marginal effect

equal to 23,99%. Regarding genre, no relationship was found. It is a highly subjective predictor, as it reflects the researcher's personal assessment. For instance, some of Kaplan's classifications were not adequate in this paper's view, such as stating 2000's *Gladiator* is a biographical film – a classification that improved the significance of his "EpicBiop" regressor, as *Gladiator* won the Best Picture Oscar.

Simonton (2002) found that first-rate directors have a significant impact on a film's odds of winning Best Picture. The current study reached the same finding. In fact, the director's proven weight in winning both Best Picture and Best Director could allow directors to demand higher wages and bonuses for significant wins such as the DGA award. The author also established that a film that portrays a true story is more likely to earn nominations and wins for the Big Four (2009). That relationship was tested in each of the Big Four, but was only found relevant for Best Actor, at a 1% significance level. It was also found evidence for his "Best Actress Paradox" in this paper, as Best Actress winners continue to star in lower quality films compared to men – proving the gender-based discrepancy is still very much significant in the film industry.

Prag and Casavant (1994) concluded that film awards were positively correlated with marketing expenses. "Major Studio" was proposed as a proxy for advertising costs, and it was only found relevant for Best Picture. It was statistically significant at a 5% level and, surprisingly enough, its coefficient was negative. This means a film distributed in the US by one of Hollywood's six major film studios has a decreased chance of winning Best Picture. A possible (and arguably the most likely) explanation is related to the themes depicted in smaller scale films. Indie films frequently address real life issues, play with society's conventions and break taboos. This serious content – which has been preferred by the Academy in recent years – is usually not as present in bigger scale films. Thus, the negative sign of "Major Studio" might not be measuring the impact of advertising costs, but the impact of themes portrayed on the chances of winning Best Picture.

6. CONCLUSION

While great uncertainty outlines the film industry, this study was able to reveal that the Oscars, the pinnacle of film awards, have a large degree of predictability. A strength of the paper is that it adds an entry to the existing literature on regression models using fundamental data for predicting the Oscars. Moreover, it analyses a time frame that had yet to be examined, while achieving an extraordinary predictive accuracy. All four models correctly attributed the highest probability of winning to the actual winner, in every single year. That's 45 out of 45 winners correctly forecasted. Even introducing a more challenging way to assess predictive quality – the cut-off, the accuracy rates were 97,40% for Picture, and 98,33% for Actress, Actor and Director.

This study's results were able to pick notable upsets, i.e., a victory over a favoured competitor. The most recent upset (and arguably the most astonishing in recent memory) was *Moonlight* winning Best Picture in 2016. The Best Picture model correctly predicted it by quite a margin.

The nominees that scored the highest probabilities and yet lost were *Boyhood* for 2014 Best Picture, Michael Keaton for 2014 Best Actor, Felicity Huffman for 2005 Best Actress and Richard Linklater for 2014 Best Director. What a competitive year 2014 was.

This paper found proof of gender discrepancies – a flaw of the film industry and of the studios running it, for shutting women out of high quality films – and proof of the criticism the Academy so often takes: portraying true characters wins acting Oscars, drama is preferred over comedy, and voters award for lifetime achievement rather than the actual best performance of the year.

A major finding was the role of Box Office in predicting the Oscars. It is statistically significant in three of the Big Four. Box office as a measure of a film's Oscar-worthiness might have been true for the 1900's. However, we are in a new Best Picture era, in which commercial acclaim does not imply Academy acclaim. Also, being helmed by a great director was found to be a bigger asset for a film coveting the Best Picture Oscar than having an ensemble cast.

Just like any study, this paper is subject to limitations. The data range was rather small, as the number of observations was only 60 in three categories. Also, as all predicted probabilities were for films that were part of the data sample, there were no out-of-sample predictions. Empirical evidence would have been trustworthier if an out-of-sample period had been used to assess the predictive quality. Lastly, the same logic that was applied to restricting the data range for a maximum of 12 years (data too old is obsolete data) can be applied to this own study. Perhaps in 12 years the final models proposed will be outdated. Eventually they will lose their prediction accuracy – though forever be a solid representation of a specific period in cinema history.

The following can be extended in future research: collect more years of data; extend the study to the remaining Oscar categories; find a way to measure variables that are yet hard to quantify, such as buzz, late surges and backlashes; create a model capable of accurately predicting the Oscars as early on in the awards season as possible.

7. REFERENCES

Bothos, Efthimios, and Dimitris Apostolou, and Gregoris Mentzas. 2010. "Using Social Media to Predict Future Events with Agent-Based Markets." *IEEE Intelligent Systems*, 25(6): 50-58.

Kaplan, David. 2006. "And the Oscar Goes to… A Logistic Regression Model for Predicting Academy Award Results." *Journal of Applied Economics & Policy*, 25(1): 23-41.

Krauss, Jonas, and Stefan Nann, and Daniel Simon, and Kai Fischbach, and Peter Gloor. 2008. "Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis." *16th European Conference on Information Systems, ECIS* (2008).

Manski, Charles F. 2006. "Interpreting the Predictions of Prediction Markets." Economic Letters, 91(3): 425-429.

Pardoe, Iain, and D. K. Simonton. 2008. "Applying Discrete Choice Models to Predict Academy Award Winners." *Journal of the Royal Statistical Society*, 171(2): 375-394.

Pardoe, Iain. 2005. "Just how predictable are the Oscars?" *Chance*, 18(4): 32–39.

Pardoe, Iain. 2007. "Predicting Oscar winners". *Significance*, 4(4): 168–173.

Prag, Jay, and James Casavant. 1994. "An empirical study of the determinants of revenues and marketing expenditures in the motion picture industry." *Journal of Cultural Economics*, 18(3): 217–235.

Simonton, D. K. 2002. "Collaborative aesthetics in the feature film: Cinematic components predicting the differential impact of 2,323 Oscar-nominated movies." *Empirical Studies of the Arts*, 20(2): 115–125.

Simonton, D. K. 2004a. "The "Best Actress" paradox: Outstanding feature films versus exceptional performances by women." *Sex Roles*, 50 (11-12): 781–795.

Simonton, D. K. 2004b. "Film awards as indicators of cinematic creativity and achievement: A quantitative comparison of the Oscars and six alternatives." *Creativity Research Journal*, 16: 163–172.

Simonton, D. K. 2009. "Cinematic success criteria and their predictors: The art and business of the film industry." *Psychology & Marketing*, 26(5): 400-420.

Surowiecki, James. 2004. *The Wisdom of Crowds*. New York: Doubleday.

Wolfers, Justin, and Eric Zitzewitz. 2004. "Prediction Markets." *Journal of Economic Perspectives*, 18(2): 107-126.

Simonton, D. K. 2004. "Group artistic creativity: Creative clusters and cinematic success in 1,327 feature films." *Journal of Applied Social Psychology*, 34(7): 1494–1520.

Simonton, D. K. 2005. "Cinematic creativity and production budgets: Does money make the movie?" *Journal of Creative Behavior*, 39(1): 1–15.

Simonton, D. K. 2007. "Film music: Are award-winning scores and songs heard in successful motion pictures?" *Psychology of Aesthetics, Creativity, and the Arts*, 1(2): 53–60.

Simonton, D. K. 2007. "Is bad art the opposite of good art? Positive versus negative cinematic assessments of 877 feature films." *Empirical Studies of the Arts*, 25(2): 143–161.

Box Office Mojo. 2017. http://www.boxofficemojo.com/.

IMDb. 2017. http://www.imdb.com/.

Metacritic. 2017. http://www.metacritic.com/.

Rotten Tomatoes. 2017. https://www.rottentomatoes.com/.

Wayback Machine. 2017. https://web.archive.org/.

Wikipedia. 2017. https://en.wikipedia.org/wiki/Main_Page.

8. ANNEX

Annex 1 – Estimation results for the final four models.

	Best Picture b/se
WINNER	
DGA	5.210***
	(1.450)
Rotten Tomatoes	0.413**
	(0.205)
Box Office	-0.032***
	(0.010)
Wins	0.014**
	(0.006)
_cons	-41.152**
	(19.583)
Number_of_obs	77
Wald_chi2	14.75
Prob>chi2	0.0052
Pseudo_R2	0.7591

* p<0.10, ** p<0.05, *** p<0.01

	Best Actor b/se
WINNER	
OscarNoms	0.484***
	(0.145)
GG	5.951***
	(1.018)
True	3.397***
	(0.870)
Box Office	-0.052***
	(0.012)
ActNoms	0.772***
	(0.187)
_cons	-7.646***
	(1.592)
Number_of_obs	60
Wald_chi2	39.46
Prob>chi2	0.0000
Pseudo_R2	0.8325

* p<0.10, ** p<0.05, *** p<0.01

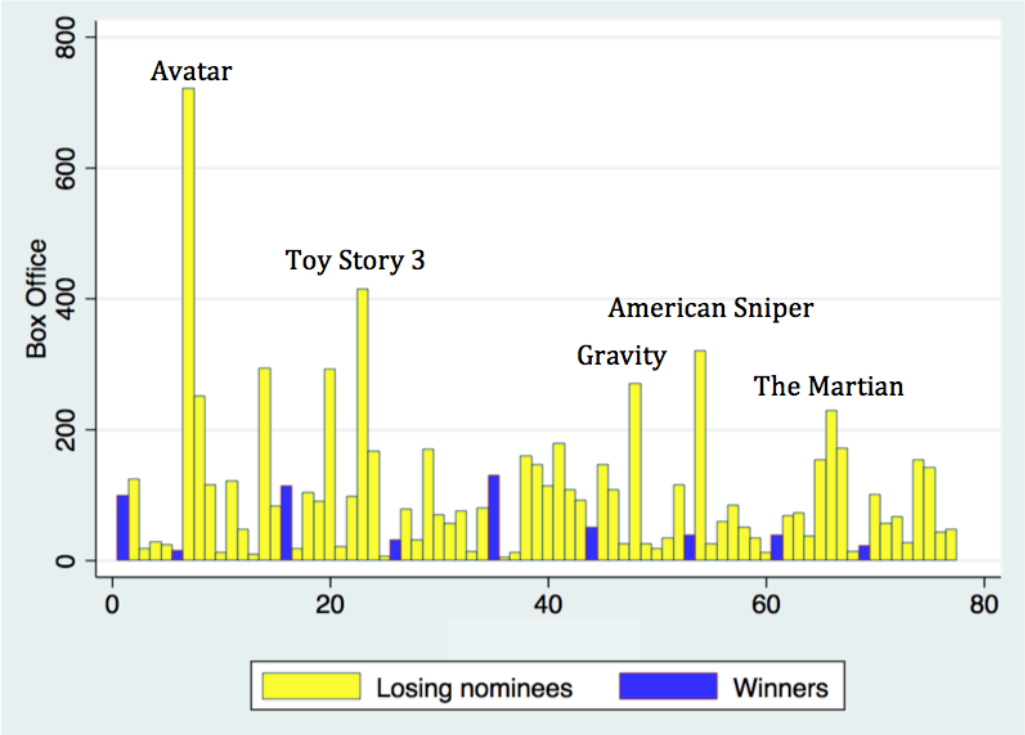
	Best Actress b/se
WINNER	
OscarNoms	0.813***
	(0.234)
Wins	0.271***
	(0.067)
Meta	-0.175***
	(0.049)
GGDrama	2.886***
	(0.923)
Box Office	0.008**
	(0.003)
_cons	3.651**
	(1.704)
Number_of_obs	60
Wald_chi2	20.55
Prob>chi2	0.0010
Pseudo_R2	0.8259

* p<0.10, ** p<0.05, *** p<0.01

	Best Director b/se
WINNER	
Index	2.845***
	(0.654)
_cons	-2.511***
	(0.618)
Number_of_obs	60
Wald_chi2	18.94
Prob>chi2	0.0000
Pseudo_R2	0.7327

* p<0.10, ** p<0.05, *** p<0.01

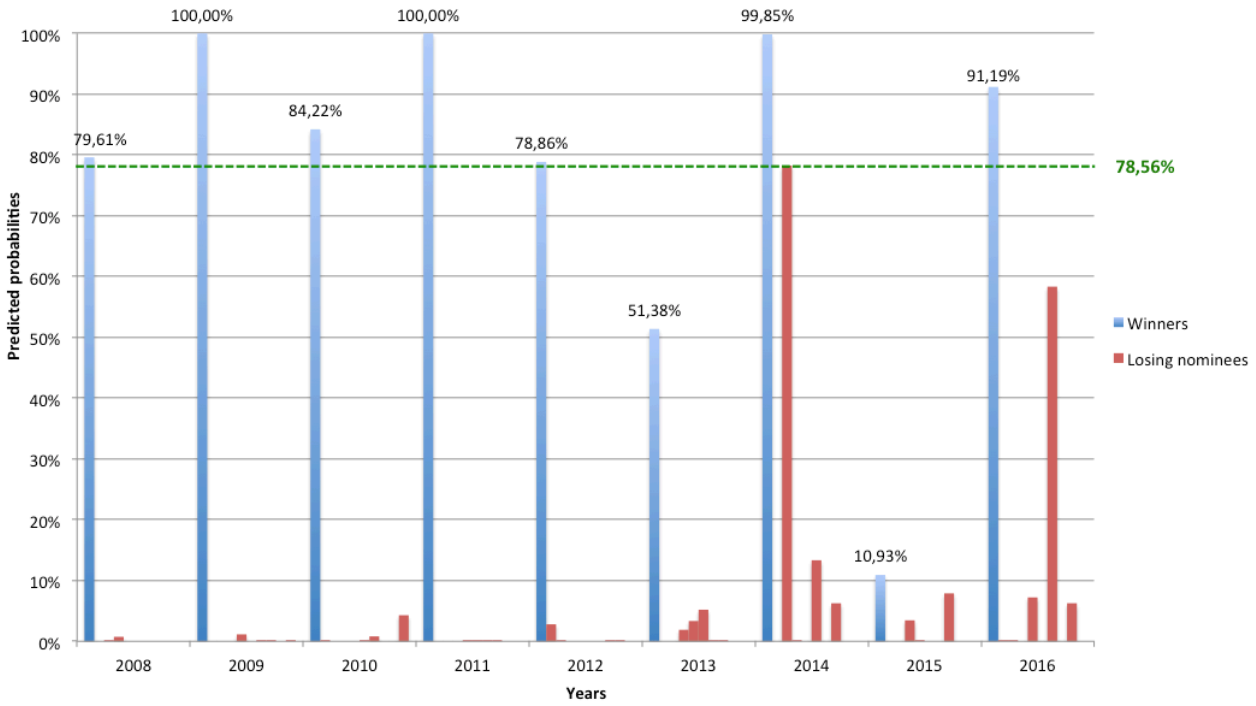
Annex 2 – Domestic total gross (in millions of US dollars) of every Best Picture nominee from 2008 to 2016. Some of the highest-grossing films are highlighted.



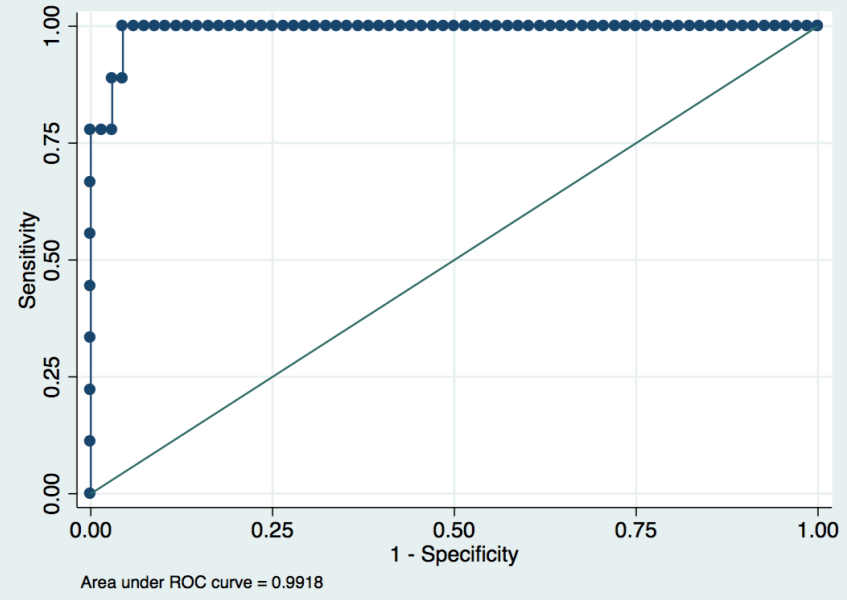
Annex 3 – Comparison of six Best Picture nominees.

Year	Film	Best Picture Oscar	Probability	DGA	Rotten Tomatoes	Box Office	Wins
2016	Moonlight	Won	91%	0	98	22,223633	205
2016	Hell or High Water	Lost	7%	0	97	27,007844	39
2014	Birdman	Won	100%	1	91	37,780892	189
2009	Precious	Lost	0%	0	91	47,395661	111
2013	Gravity	Lost	3%	1	96	270,47883	232
2012	Argo	Won	79%	1	96	129,6535	94

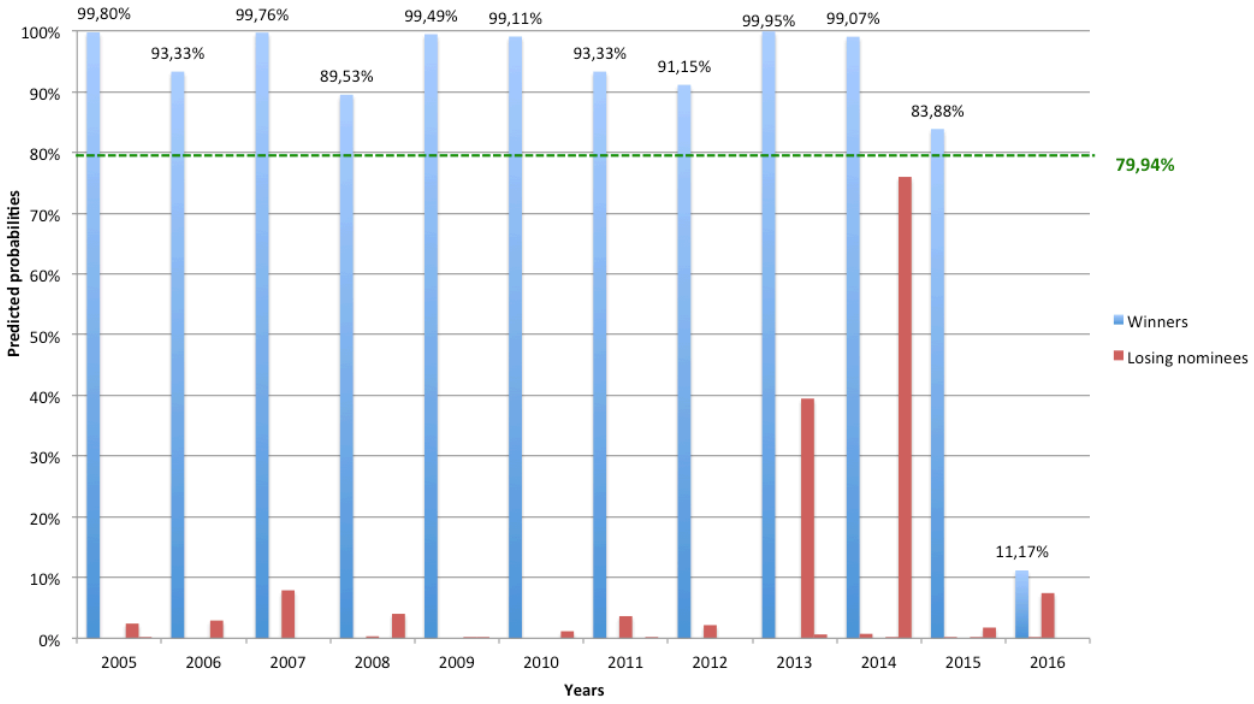
Annex 4 – Best Picture model: nominees' predicted probabilities and cut-off (green dashed line).



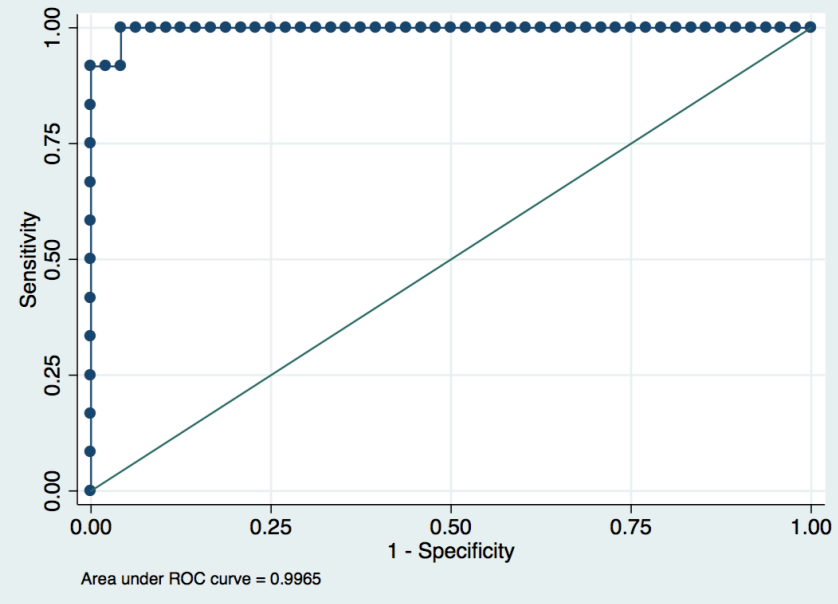
Annex 5 – Best Picture model: ROC curve.



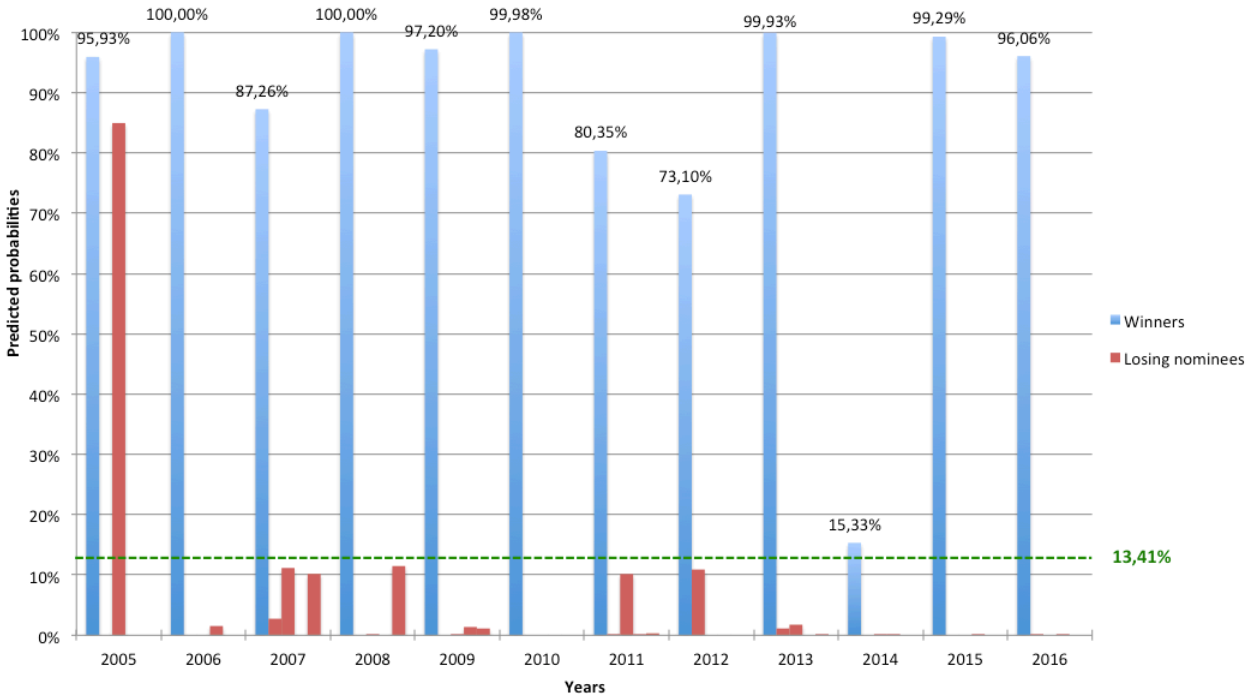
Annex 6 – Best Actor model: nominees' predicted probabilities and cut-off (green dashed line).



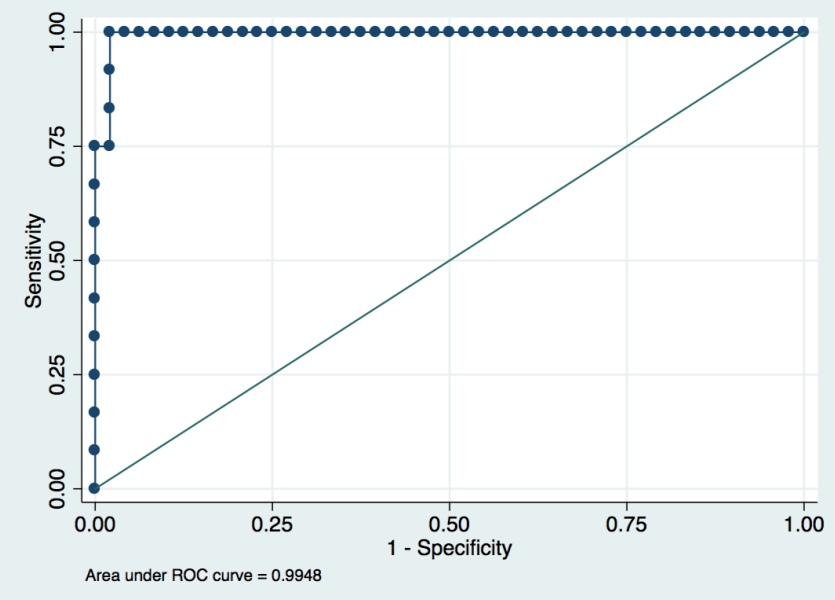
Annex 7 – Best Actor model: ROC curve.



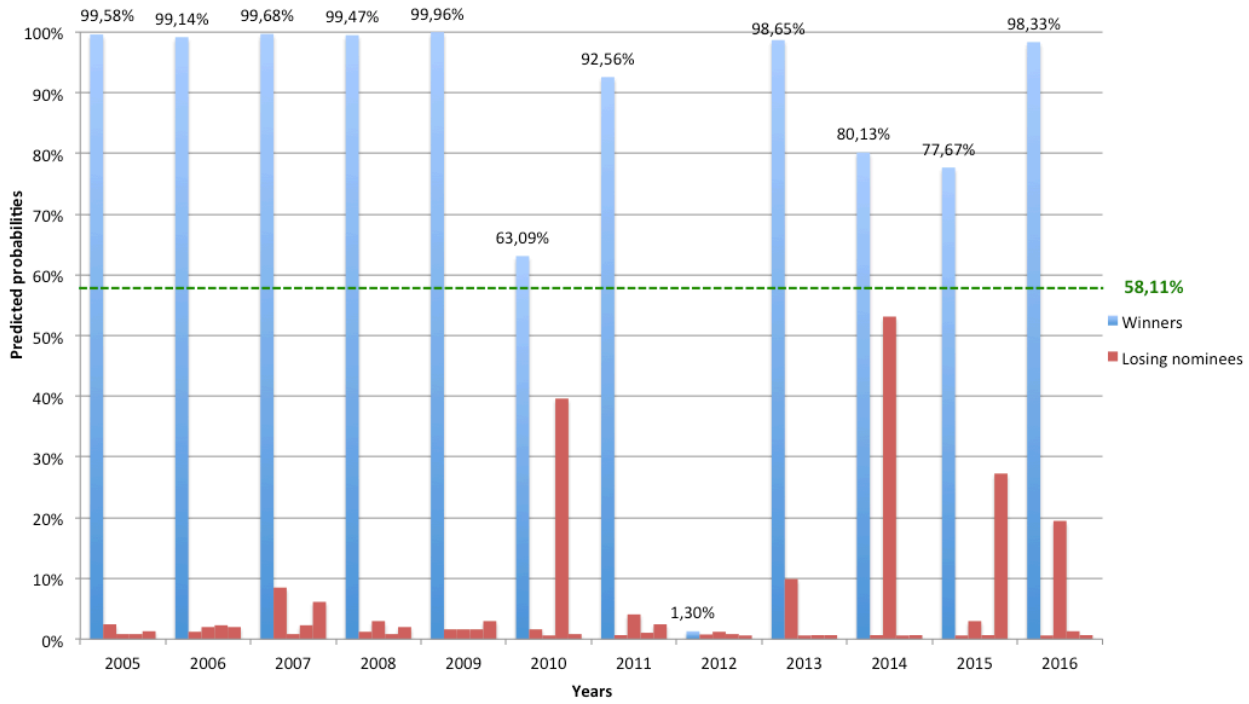
Annex 8 – Best Actress model: nominees' predicted probabilities and cut-off (green dashed line).



Annex 9 – Best Actress model: ROC curve.



Annex 10 – Best Director model: nominees' predicted probabilities and cut-off (green dashed line).



Annex 11 – Best Director model: ROC curve.