LIGHTS, CAMERA, PREDICTION:

Predicting the Oscars with a nearest-neighbour classification model using the results of the BAFTAs, Golden Globes and SAG awards

Abstract

This report explores the question ‘how accurately can we predict the winners of the Academy Awards for Best Picture, Best Actor and Best Actress based on the winner’s of the Golden Globes, BAFTAs and SAG Awards from the same season’, utilising 25 years of nominee and winner data, we developed a nearest-neighbour classification model that predicts all three awards with an accuracy 49% per category and an overall accuracy of 63% per year. The report explores the predictive power of the three primary awards shows for predicting the outcomes of the Oscars, using a comparison between the classification model and a multiple regression model to potentially determine a more effective or worthwhile approach. These findings give insights into film industry trends and give an idea as to the extent to which statistical modelling and data can be used to predict such trends.

Introduction

Since its inception in 1929, the Academy Awards, also known as the Oscars, has been the most prestigious awards ceremony in the film industry (Miffin, 1995), a status earned partly through its history, and partly through its size and makeup. The Academy itself comprises just under 10,000 voting members, spanning nineteen professional branches, the largest bloc of which are actors (Pond, 2024).

The significance of the Oscars to industry professionals can not be understated. Studios have been known to spend hundreds of millions of dollars to promote their films during "Oscar season" in the months ahead of the awards ceremony in early spring, solely to win as many awards as possible (Reed, 2016). PricewaterhouseCoopers (PwC), one of the largest professional services firms in the world, has been employed for decades to verify each vote and thus the legitimacy of the awards themselves (D'Souza, 2025).

Other film awards ceremonies have emerged and found similar, although not equivalent, popularity and prestige. The BAFTAs are the UK's answer to the Oscars, the Golden Globes are voted for by industry journalists (rather than creatives), and in more recent times, the SAG awards give a voice to the 160,000 members of the SAG-AFTRA union (SAG-AFTRA, 2023).

Part of the spectacle is to predict the winners of the Academy Awards in each category, especially the "Big Three" – Best Picture, Best Actor, and Best Actress. Across the course of "Oscar season", the results of the Golden Globes (GGs), BAFTAs and SAG awards increasingly build a picture of what viewers might finally expect from the Academy Awards… yet the individual qualities, different demographics, and general volatility of each ceremony makes for a real challenge when it comes to accurately predicting the Oscars.

In this paper, we look at the last 25 years of results from these four film awards ceremonies and build a nearest neighbour classification model to determine how accurately the Academy Awards for Best Picture, Best Actor, and Best Actress can be predicted by the prior results in the current film awards season. We also use a multiple regression model to investigate which of the three other awards ceremonies is the best predictor when it comes to the Oscars.

The rest of this paper is structured as follows: Section 2 covers the data and models, Section 3 goes through the results, and Section 4 concludes with our findings and some limitations.

Data and model

The dataset we are using has been collated from various publically available datasets, including those from Wikipedia and Kaggle. We have opted to use nominee and winner data from 2000-2025 for all four awards shows, as the SAG awards only began in 1995, and not all datasets were complete prior to 2000. The Kaggle/Wikipedia data lends itself to our classification model as it is categorical and binary, and we can use the data we have to train and test the model. Consequently, we will be able to build our classification and multiple regression models using this data to analyse and predict the relationship between prior success in the award season and winning the relevant Oscar.

We had to clean up some discrepancies between the original public data sets – for example, Quentin Tarantino's "*Once Upon a Time… in Hollywood*" varied in how many dots made up its ellipsis, and two Best Actor nominees from BAFTAs/GGs/SAG, due to differences in screentime requirements, were nominated for Best Supporting Actor at the Oscars. Those two instances were discounted from our dataset. Finally, as the SAG awards celebrate acting in particular, their closest equivalent to Best Picture is the Outstanding Performance by a Cast in a Motion Picture award (OPCMP). For the purposes of our exploration, we treated these awards as being in the same category, and thus OPCMP is treated as a predictor of the Academy Award for Best Picture. Additional differences between awards will be explained in greater detail in the next section.

As explained in the introduction, we are using two models for this exploration, a nearest-neighbour classification model and a multiple regression model.

We will run a variety of tests using the nearest-neighbour classification model. Each test will train the model on previous data from the results of each award show, and use it to predict whether it is likely that a film / actor / actress in a certain year will be nominated for their Academy Award, based on their results in the other award ceremonies. We can also use this model to provide a probability for the likelihood of an entry winning the award, based on the number of winner / non-winner neighbours it has. The data lends itself to this type of model as it is categorical and binary in nature, and we can also use the dataset we have created to both train and test the model.

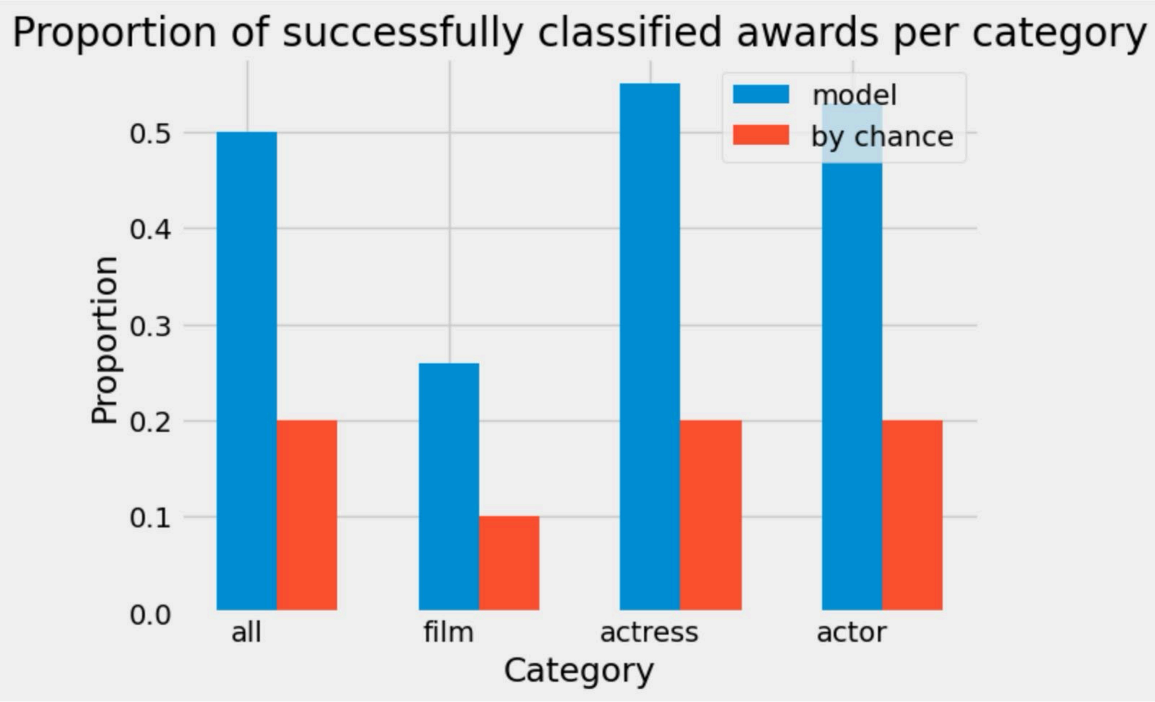
Additionally, our multiple regression model should give an indicator of the influence of each 'predictor' award show on the Academy Awards, allowing for comparison with the results of the classification model.

One downside to using the same dataset for both of these models is that if the data itself is incorrect or skewed in some capacity, this will skew both models, but due to the scope of the project at this stage (just 25 years), we are limited in what we can do. Another downside of the dataset we are using is that the Golden Globes have split all their awards into 'Comedy/Musical' and 'Drama' categories. We decided to include these categories as separate variables in our models, but realise that this is a limitation as the models consider all variables equally.

Results

Classification test 1: predicting per category

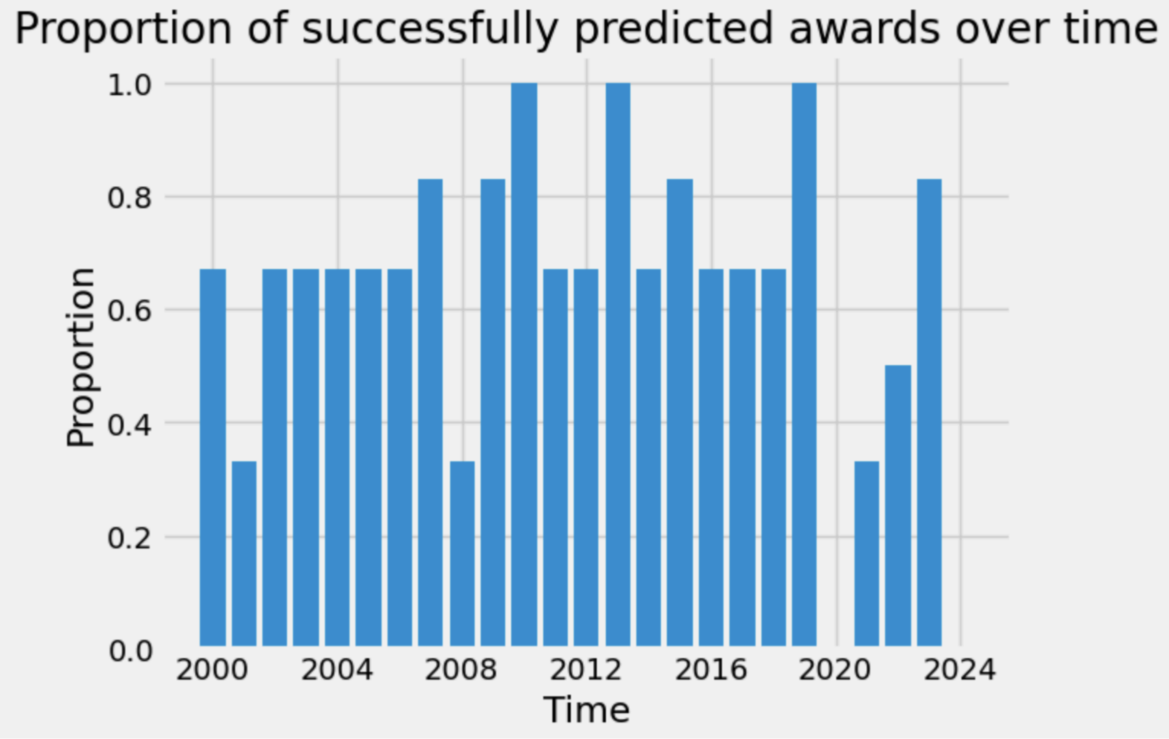
The first test for classification was calculating the accuracy of the predictor over each category. We ran 1000 trials, randomly shuffling the data each time. We then trained the model on 11/16 of the data from the category of interest and tested the model on the Oscar winners from the remaining part of the data, recording the proportion of successful trials and finding the mean average at the end of the process. On average, our model correctly classified 24% of Best Picture winners, 57% of Best Actor winners, and 56% of Best Actress winners, meaning it was able to predict each category with over twice the accuracy of random chance (as the award for Best Picture has 10 nominees, and the awards for Best Actor and Best Actress have five nominees each). Considering all the data, the model accuracy was 49%. The graph below shows this in more detail, and in comparison to random chance.



Classification test 2: predicting per year

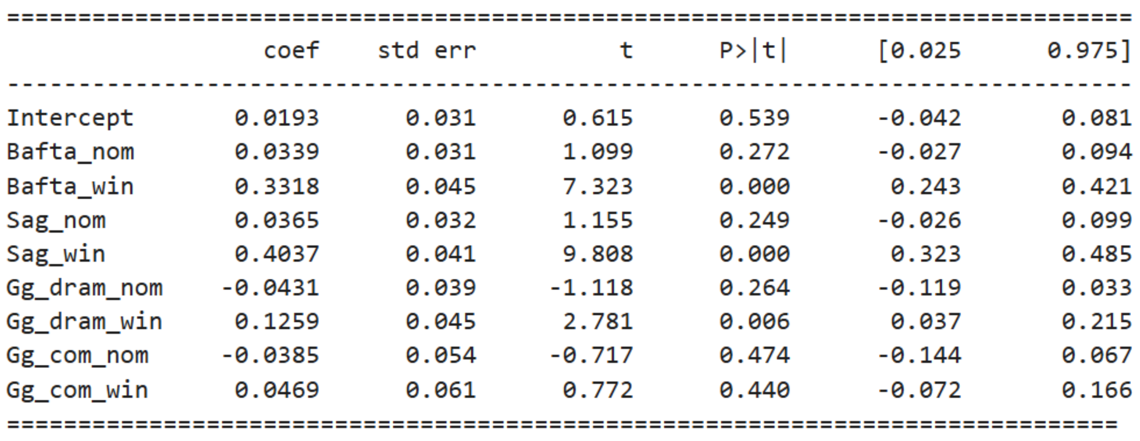
We then ran the model over every year in our dataset. This test took the nominees for each year, trained the model on all the data excluding that year, outputted the most likely winners for each category, and then checked whether they were correct. Using a probability measure of our own design[1], we averaged the results of this to give us an accuracy of 63%. The graph on the next page shows the variation of this each year.

[1] This assigned 1 to a success, 0 to a fail and 0.5 to a partial success (where the award was successfully predicted as a winner but wasn't the only one predicted to win)



Regression on all award types

A p-values test on all award ceremonies against the Oscar winners highlights that only Bafta winners, SAG winners, and GG drama winners have p-values < 0.05 (0.000, 0.000, & 0.006 respectively, as shown below).

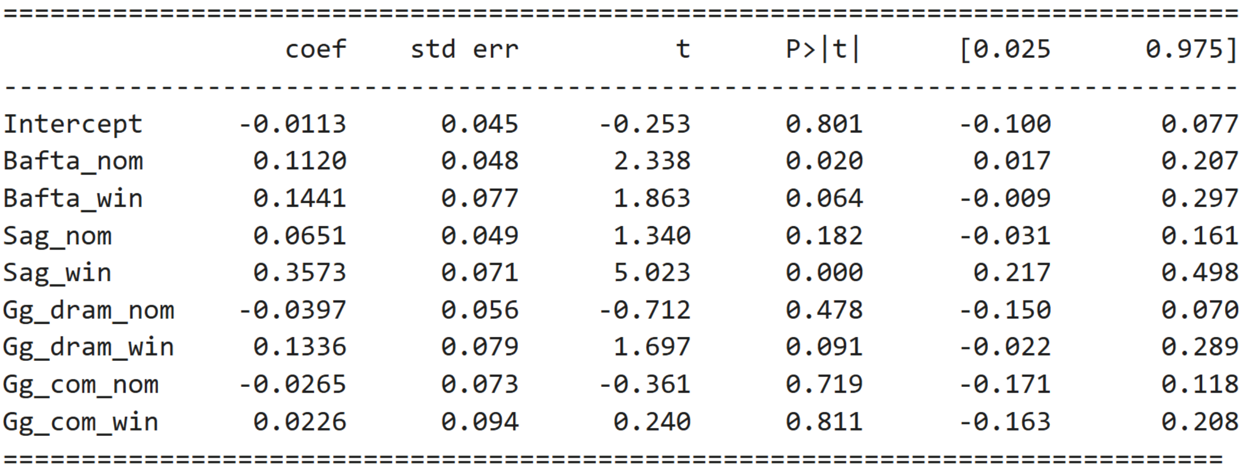


```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.0193      0.031      0.615      0.539      -0.042       0.081
Bafta_nom      0.0339      0.031      1.099      0.272      -0.027       0.094
Bafta_win      0.3318      0.045      7.323      0.000       0.243       0.421
Sag_nom        0.0365      0.032      1.155      0.249      -0.026       0.099
Sag_win        0.4037      0.041      9.808      0.000       0.323       0.485
Gg_dram_nom   -0.0431      0.039     -1.118      0.264      -0.119       0.033
Gg_dram_win    0.1259      0.045      2.781      0.006       0.037       0.215
Gg_com_nom    -0.0385      0.054     -0.717      0.474      -0.144       0.067
Gg_com_win     0.0469      0.061      0.772      0.440      -0.072       0.166
=============================================================================
```

The multiple regression of these variables (Bafta_win, Sag_win, and Gg_win) gives an r-squared value of 0.459, meaning 45.9% of the variance in the model can be explained by the Bafta, SAG, and GG winners.

Regression on film awards

Isolating the data to only include film awards changes which variables have p-values < 0.05; now it is Bafta nominees (0.020) and SAG winners (0.000).

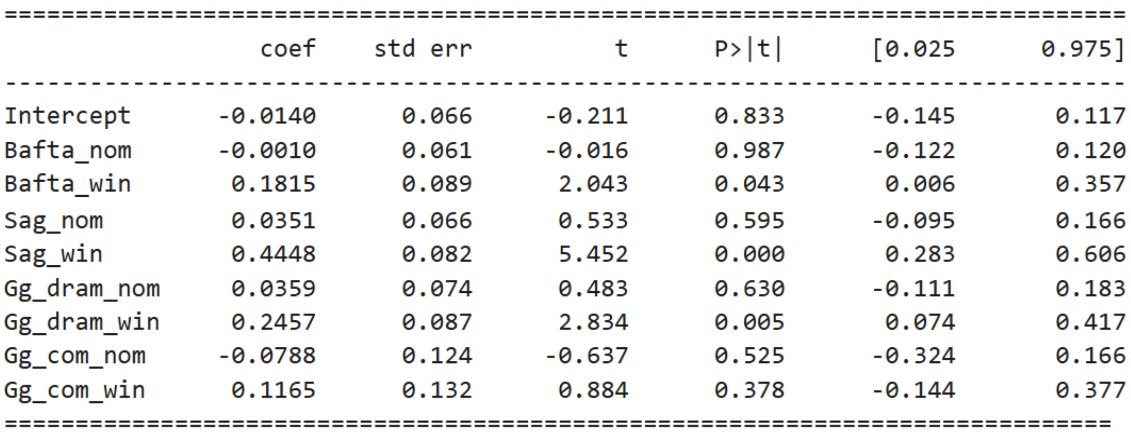


```
================================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept     -0.0113      0.045     -0.253      0.801      -0.100       0.077
Bafta_nom      0.1120      0.048      2.338      0.020       0.017       0.207
Bafta_win      0.1441      0.077      1.863      0.064      -0.009       0.297
Sag_nom        0.0651      0.049      1.340      0.182      -0.031       0.161
Sag_win        0.3573      0.071      5.023      0.000       0.217       0.498
Gg_dram_nom   -0.0397      0.056     -0.712      0.478      -0.150       0.070
Gg_dram_win    0.1336      0.079      1.697      0.091      -0.022       0.289
Gg_com_nom    -0.0265      0.073     -0.361      0.719      -0.171       0.118
Gg_com_win     0.0226      0.094      0.240      0.811      -0.163       0.208
===============================================================================
```

This multiple regression of Oscar winners against Bafta nominees and SAG winners gives an r-squared value of 0.247.

Regression on actor awards

Now taking the actor awards data, again there are differences in which variables have p-values < 0.05. These are Bafta winners, SAG winners, and GG comedy winners (0.043, 0.000, & 0.005 respectively).

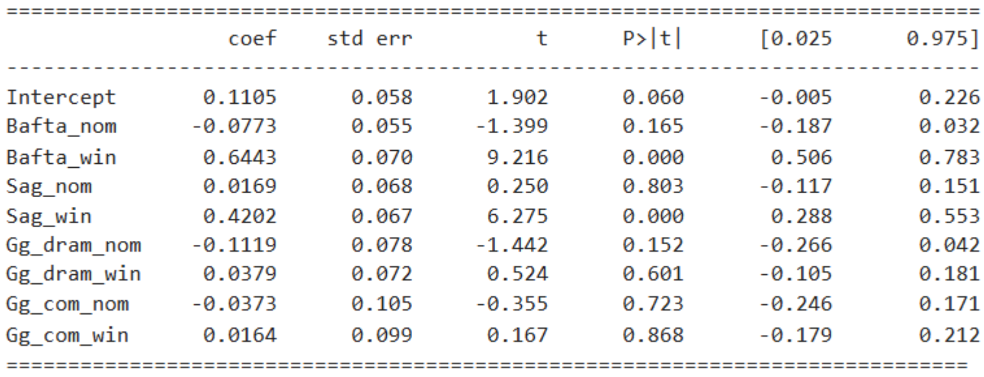


```
===============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept     -0.0140      0.066     -0.211      0.833      -0.145       0.117
Bafta_nom     -0.0010      0.061     -0.016      0.987      -0.122       0.120
Bafta_win      0.1815      0.089      2.043      0.043       0.006       0.357
Sag_nom        0.0351      0.066      0.533      0.595      -0.095       0.166
Sag_win        0.4448      0.082      5.452      0.000       0.283       0.606
Gg_dram_nom    0.0359      0.074      0.483      0.630      -0.111       0.183
Gg_dram_win    0.2457      0.087      2.834      0.005       0.074       0.417
Gg_com_nom    -0.0788      0.124     -0.637      0.525      -0.324       0.166
Gg_com_win     0.1165      0.132      0.884      0.378      -0.144       0.377
==============================================================================
```

The multiple regression of Oscar winners against Bafta winners, SAG winners, and GG comedy winners on actor awards data gives an r-squared value of 0.543.

Regression on actress awards

Finally, with the actress awards data it’s just BAFTA winners (0.000) and SAG winners (0.000) that have a p-value < 0.05, as demonstrated below:



```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.1105      0.058      1.902      0.060      -0.005       0.226
Bafta_nom     -0.0773      0.055     -1.399      0.165      -0.187       0.032
Bafta_win      0.6443      0.070      9.216      0.000       0.506       0.783
Sag_nom        0.0169      0.068      0.250      0.803      -0.117       0.151
Sag_win        0.4202      0.067      6.275      0.000       0.288       0.553
Gg_dram_nom   -0.1119      0.078     -1.442      0.152      -0.266       0.042
Gg_dram_win    0.0379      0.072      0.524      0.601      -0.105       0.181
Gg_com_nom    -0.0373      0.105     -0.355      0.723      -0.246       0.171
Gg_com_win     0.0164      0.099      0.167      0.868      -0.179       0.212
==============================================================================
```

The multiple regression of Oscar winners against BAFTA winners and SAG winners on actress awards data gives an r-squared value of 0.628.

Multiple regression coefficient values and adjusted r squared for full model only considering statistically significant predictors (look at p-values)

Discussion

Our classification model successfully predicted the winners of the Academy Awards in the Best Picture, Best Actor, and Best Actress categories with an accuracy that more than doubled the baseline of random chance. Best Actor and Best Actress predictions were particularly strong, with over 55% accuracy, compared to the 20% baseline. Although Best Picture predictions were less accurate at 24%, they still exceeded the 10% random baseline, suggesting that even in broader categories with more variability, prior award results provide meaningful predictive value.

The results of the multiple regression models provide further insight into which award ceremonies carry the most predictive weight. Across all categories, the BAFTAs and SAG Awards consistently produced statistically significant p-values (< 0.05), especially in the acting categories. This aligns with the fact that the Academy has a large acting branch, increasing the overlap with the SAG voting base. Interestingly, the Golden Globes, which are decided by journalists rather than industry professionals, showed less consistent significance across award categories, highlighting their more inconsistent influence on Oscar outcomes.

Our strongest regression result was found in the Best Actress category, where BAFTA and SAG wins together explained 62.8% of the variance in Oscar winners, as measured by the r-squared value. This suggests that in this category, prior award season wins are particularly indicative of Oscar success. Meanwhile, the Best Picture regression was the weakest, presumably due to the increased number of nominees.

Whilst our analysis is limited to 25 years of data and a relatively small number of award categories, future iterations could expand into technical categories or look at shifts over time to explore how predictive relationships have evolved (e.g. since the increase from 5 to 10 nominees for Best Picture). Moreover, differences between awards — such as juries, voting blocs, or cultural leanings — could be explored in greater depth to build upon our findings. Overlap and differences in voter base between awards shows provide a broader ‘opinion’ and wider scope for the films/actors/actresses that win, leading to more variation in the dataset.

In summary, our models demonstrate that while Oscar winners are far from fully predictable, there is a clear and quantifiable relationship between preceding major awards and eventual Academy Award success. This supports the idea that the Oscars are not decided in a vacuum, but rather reflect an aggregation of industry opinion formed and reinforced throughout the award season.

References

D’Souza, D. (2025, January 4). ‘*Why an Accounting Firm Still Plays a Starring Role in the Oscars*’.

Investopedia.

https://www.investopedia.com/what-does-an-accounting-firm-for-the-oscars-do-4586515

Mifflin, L. (1995, May 22). ‘*More Awards Programs, More Winners, More Money*’. The New York Times.

https://www.nytimes.com/1995/05/22/business/more-awards-programs-more-winners-more-money.html

Pond, S. (2024, December 12). ‘*How Many Votes Will It Take to Get an Oscar Nomination in 2025?*’. TheWrap.

https://www.thewrap.com/how-many-votes-to-get-an-oscar-nomination-2025/

Reed, J. (2016, February 23). ‘*How much does it cost to win an Oscar?*’. BBC News.

https://www.bbc.co.uk/news/entertainment-arts-35613630

SAG-AFTRA. (2023). ‘*About SAG-AFTRA | SAG-AFTRA*’. Sagaftra.org.

https://www.sagaftra.org/spanish-language-television/about-sag-aftra

Kaggle. (2025). ‘*The Oscar Award, 1927 - 2025*’.
https://www.kaggle.com/datasets/unanimad/the-oscar-award

Wikipedia. (2025, January 27). ‘*Screen Actors Guild Award for Outstanding Performance by a Cast in a Motion Picture*’.
https://en.wikipedia.org/wiki/Screen_Actors_Guild_Award_for_Outstanding_Performance_by_a_Cast_in_a_Motion_Picture

Wikipedia. (2025, March 3). ‘*Screen Actors Guild Award for Outstanding Performance by a Female Actor in a Leading Role*’.
https://en.wikipedia.org/wiki/Screen_Actors_Guild_Award_for_Outstanding_Performance_by_a_Female_Actor_in_a_Leading_Role

Wikipedia. (2025, April 3). ‘*Screen Actors Guild Award for Outstanding Performance by a Male Actor in a Leading Role*’.
https://en.wikipedia.org/wiki/Screen_Actors_Guild_Award_for_Outstanding_Performance_by_a_Male_Actor_in_a_Leading_Role

Wikipedia. (2025, March 27). ‘*Golden Globe Award for Best Motion Picture – Musical or Comedy*’.
https://en.wikipedia.org/wiki/Golden_Globe_Award_for_Best_Motion_Picture_–_Musical_or_Comedy

Wikipedia. (2025, March 25). '*Golden Globe Award for Best Actress in a Motion Picture – Musical or Comedy*'. https://en.wikipedia.org/wiki/Golden_Globe_Award_for_Best_Actress_in_a_Motion_Picture_–_Musical_or_Comedy

Wikipedia. (2025, March 15). '*Golden Globe Award for Best Actor in a Motion Picture – Musical or Comedy*'. https://en.wikipedia.org/wiki/Golden_Globe_Award_for_Best_Actor_in_a_Motion_Picture_–_Musical_or_Comedy

Wikipedia. (2025, March 4). '*Golden Globe Award for Best Motion Picture – Drama*'. https://en.wikipedia.org/wiki/Golden_Globe_Award_for_Best_Motion_Picture_–_Drama

Wikipedia. (2025, March 30). ‘*Golden Globe Award for Best Actress in a Motion Picture – Drama*’. https://en.wikipedia.org/wiki/Golden_Globe_Award_for_Best_Actress_in_a_Motion_Picture_–_Drama

Wikipedia. (2025, March 4). ‘*Golden Globe Award for Best Actor in a Motion Picture – Drama*’. https://en.wikipedia.org/wiki/Golden_Globe_Award_for_Best_Actor_in_a_Motion_Picture_–_Drama

Wikipedia. (2025, April 2). ‘*BAFTA Award for Best Film*’. https://en.wikipedia.org/wiki/BAFTA_Award_for_Best_Film

Wikipedia. (2025, March 22). ‘*BAFTA Award for Best Actress in a Leading Role*’. https://en.wikipedia.org/wiki/BAFTA_Award_for_Best_Actress_in_a_Leading_Role

Wikipedia. (2025, April 3). ‘*BAFTA Award for Best Actor in a Leading Role*’. https://en.wikipedia.org/wiki/BAFTA_Award_for_Best_Actor_in_a_Leading_Role