Exploratory Data Analysis on

Prediction of Oscar Winners

Contents

List of Figures

List of Tables

Abstract

The Academy Awards (Oscars) represent one of the most prestigious recognitions in the film industry. Predicting Oscar winners is an intriguing challenge due to the interplay of subjective preferences, artistic merit, and measurable factors like box office performance and critical acclaim. This project leverages machine learning techniques to analyze historical Oscar data and identify the key factors influencing nominations and wins across categories.

We utilized a comprehensive dataset derived from multiple sources, including IMDb, Rotten Tomatoes, Wikipedia, and award-specific archives, to extract a diverse range of features such as movie ratings, genres, release dates, production metrics, and social media sentiments. Additional features, like historical award trends and individual track records, were engineered to enhance the predictive capability of models. The collected data underwent rigorous cleaning, imputation, and transformation to ensure quality and compatibility with machine learning algorithms.

While the models achieved high accuracy, challenges such as limited historical data, evolving voting trends, and the unpredictability of subjective decisions remain. This project highlights the potential of data-driven insights in understanding and predicting outcomes in artistic domains. Future work aims to integrate real-time data, expand feature sets, and apply the framework to other prestigious award ceremonies to further refine the predictive methodology.

Chapter 1. Introduction

1.1 Project idea

The project focuses on predicting Oscar winners and nominations using machine learning techniques. The aim is to analyze various factors influencing the Academy Awards, such as genres, production houses, box office performance, critic reviews, and social sentiments. This predictive analysis could provide insights into trends and biases within the awards.

1.2 Data Collection

The data collection process is carried out in a structured, multi-phase approach, aimed at converting raw data from various sources into a well-organized, analysis-ready dataset. Initially, data is extracted simultaneously from IMDb, Rotten Tomatoes, and Wikipedia to gather movie metadata, ratings, and award details. Each dataset undergoes a thorough cleaning and standardization process, which includes addressing missing values, parsing text data, and formatting dates consistently.

During the feature engineering stage, the raw information is prepared for machine learning by creating genre-based dummy variables, categorizing age groups, and encoding award history as binary features. Finally, the processed features are merged into a single DataFrame, with numeric missing values imputed and all variables normalized, ensuring the dataset is ready for comprehensive analysis of cinematic performance.

1.3 Dataset Description

The dataset for this project encompasses a rich collection of features from various sources, tailored to analyze and predict the outcomes of the Academy Awards (Oscars). Below is a detailed breakdown of the dataset structure:

1. Movie Metadata

- **Title:** The name of the movie (e.g., The Godfather).
- **Release Year:** Year of the movie's release, used to analyze trends and eligibility for specific award years.
- **Genre:** Categories such as Drama, Action, Comedy, etc. A movie can belong to multiple genres.
- **Director:** Name of the director(s), as the director's reputation often influences award chances.

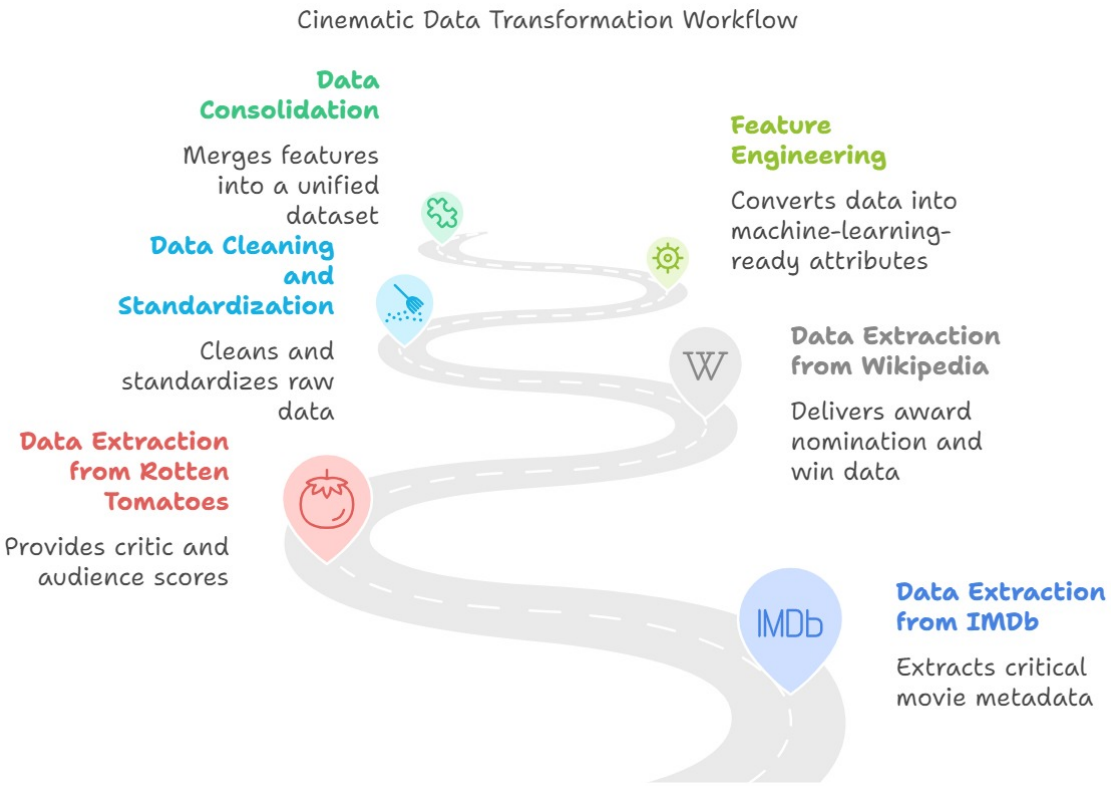

Figure 1.1: Data Transformation Workflow

- **Cast:** List of main actors, as star power and past nominations/wins are crucial factors.
- **Studio/Production House:** The studio backing the movie, as larger studios might have better campaigning resources.

2. Awards Information

- **Award Year:** The year the movie was considered for an Oscar.
- **Categories:** Specific award categories (e.g., Best Picture, Best Actor).
- **Nominations:** Whether the movie was nominated for an award (Yes/No).
- **Wins:** Whether the movie won an Oscar (Yes/No).

3. Financial and Performance Metrics

- **Budget:** Production cost of the movie (in USD), indicative of scale and quality.
- **Box Office Revenue:** Total earnings (domestic and international), which can reflect popularity and impact.

- **Revenue-to-Budget Ratio:** Derived metric indicating profitability.

4. Ratings and Reviews

- **IMDb Rating:** Average user score from IMDb.
- **Critic Scores:** Ratings from platforms like Rotten Tomatoes and Metacritic, reflecting critical acclaim.
- **Audience Scores:** Ratings from the general audience, highlighting public opinion.

5. Textual Data

- **Plot Summary:** A textual synopsis of the movie, useful for sentiment analysis and keyword extraction.
- **Reviews:** Aggregated critic and audience reviews, providing detailed feedback on the movie.

6. Social Media and Sentiment

- **Social Media Mentions:** Frequency of mentions on platforms like Twitter and Facebook during the award season.
- **Sentiment Scores:** Derived from reviews and social media data using natural language processing (NLP) techniques, indicating overall perception.

7. Historical Trends and Derived Features

- **Past Success:** Track records of the director, actors, and studios in previous Oscar ceremonies.
- **Genre Success Rate:** The historical probability of a specific genre winning or being nominated.
- **Seasonality:** Analysis of release date impact, as movies released during award seasons (fall/winter) often fare better.

1.4 Packages required

Key Python libraries and tools:

- Data Manipulation: pandas, numpy
- Visualization: matplotlib, seaborn, plotly

- Feature Engineering: sklearn, nltk
- Web Scraping: BeautifulSoup, scrapy
- Modeling: scikit-learn, xgboost, tensorflow
- API Integration: requests, imdbpy

Chapter 2. Data Cleaning

2.1 Missing Data Analysis

In any dataset, missing data is a common issue that can significantly impact the results of statistical analysis and predictive modeling. Identifying and analyzing missing data patterns is crucial to ensure that the applied imputation methods are suitable and unbiased.

2.1.1 Types of Missing Data

Missing data can be broadly classified into the following categories:

1. **Missing Completely at Random (MCAR):** The probability of data being missing is independent of the observed and unobserved data. For example, a survey response might be missing because the participant accidentally skipped the question.
2. **Missing at Random (MAR):** The probability of data being missing is dependent only on the observed data and not the missing data. For instance, a participant's income information might be missing due to their reluctance, which can correlate with their level of education.
3. **Missing Not at Random (MNAR):** The probability of missing data depends on the value of the missing data itself. For example, individuals with very high or low incomes might intentionally leave income fields blank.

2.1.2 Exploratory Analysis of Missing Data

To understand the extent and nature of missing data in a dataset, the following steps are typically performed:

- **Visualization:** Heatmaps and bar plots can provide a clear view of missing values across features.
- **Summary Statistics:** Calculate the percentage of missing values for each feature.
- **Pattern Analysis:** Identify if missing data occurs in patterns, such as entire rows or specific columns being incomplete.

2.2 Imputation

Once missing data is identified and analyzed, appropriate imputation methods are used to handle the gaps. The choice of imputation technique depends on the type and extent of missing data.

2.2.1 Common Imputation Methods

- **Mean/Median Imputation:** Replace missing values with the mean or median of the non-missing data in the same feature. This method is simple but may introduce bias if the data distribution is skewed.
- **Mode Imputation:** For categorical features, missing values can be replaced with the mode (most frequent value). This is effective for non-numeric variables with a dominant category.
- **K-Nearest Neighbors (KNN) Imputation:** Missing values are imputed based on the values of the nearest neighbors. This method preserves relationships between features but can be computationally intensive.
- **Multivariate Imputation by Chained Equations (MICE):** This iterative approach models each feature with missing data as a function of other features, filling in values iteratively. It is suitable for datasets with complex interdependencies.
- **Regression Imputation:** Missing values are predicted using regression models based on observed data. This is useful when the relationships between variables are well-defined.

2.2.2 Imputation Strategies

The choice of imputation strategy depends on the missing data mechanism:

1. For MCAR data, simpler methods like mean or mode imputation may suffice.
2. For MAR data, advanced techniques like KNN or MICE are preferred as they account for dependencies between variables.
3. For MNAR data, additional external information or domain expertise may be required to model and impute the missing values effectively.

2.2.3 Impact of Imputation on Analysis

While imputation helps in preserving the dataset's size and usability, it can introduce bias or distort relationships between features. Therefore, post-imputation validation, such as comparing distributions before and after imputation, is crucial.

Chapter 3. Visualization

3.1 Univariate analysis

Univariate analysis involves examining each variable independently to understand its distribution, central tendency, and variability. Key insights are summarized below:

- **Award Variables:**
  - Dummy variables were created for certain awards (e.g., PGA, SAG, Critics' Choice) with three states: non-existent, nomination but no win, and win.
  - Total nominations are strong univariate predictors: median nominations for winners is 10 compared to 6 for losers.
- **Genre Variables:**
  - Films classified as *Biography*, *History*, *Music*, *Western*, or *War* are more likely to win compared to genres like *Horror* and *Mystery*.
  - *Drama* was excluded due to its overwhelming presence.
- **Critical and Popular Ratings:**
  - Winners have higher scores on average across *Rotten Tomatoes* (Audience and Critics' Scores) and *IMDB*.
- **Release Date:**
  - Films released in Q4 of the previous year (42%) or Q1 of the ceremony year (30%) are more likely to be nominated and win.
- **Age of Actors and Actresses:**
  - Actors above 75 have the highest probability of winning (23.7%), while actors below 25 have an 11% chance.

3.2 Multivariate analysis

Multivariate analysis examines relationships between multiple variables and their combined predictive power.

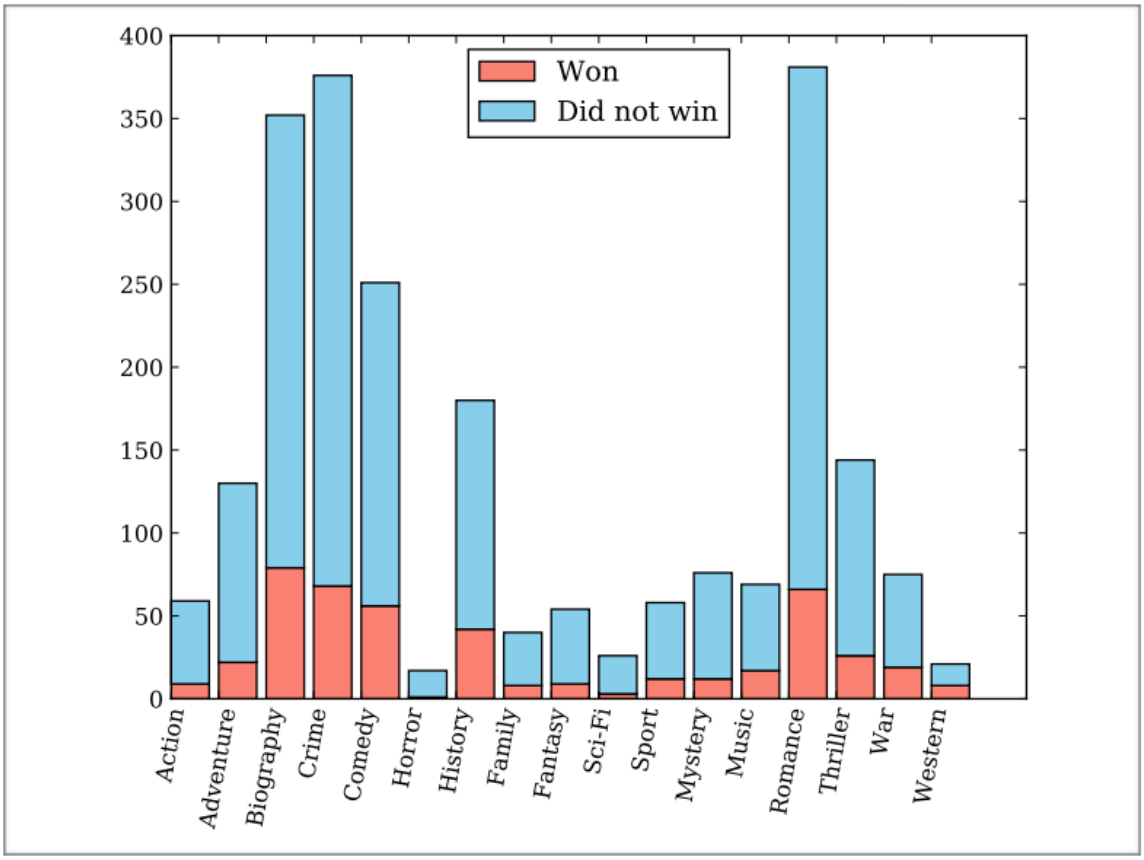

Figure 3.1: Genre Distribution of Oscar Nominated Films (1960–2018).

- **Correlation Between Variables:**
  - Figure 3.2 shows a positive correlation between *Rotten Tomatoes* and *IMDB* scores.
  - Awards such as DGA strongly correlate with Best Director and Best Picture categories.
- **Genre and Awards Interaction:**
  - Certain genres, like *Western*, increase the likelihood of winning in the Best Director category.
- **Award Overlaps:**
  - Significant overlaps exist among nominees and winners across award ceremonies.
- **Demographic Variables:**
  - Age segmentation interacts with variables like prior nominations and wins. Actors with previous wins have lower odds of winning, indicating diminishing returns from past accolades.

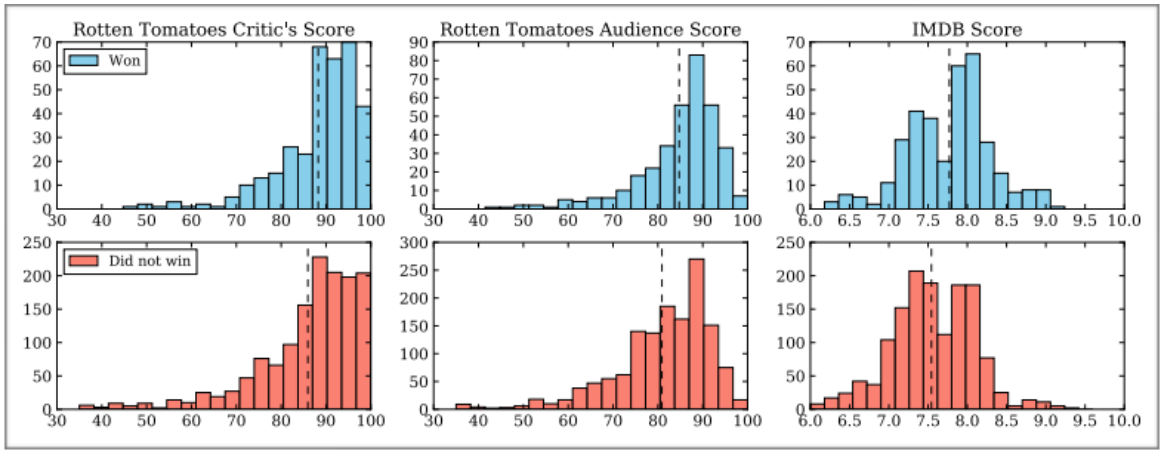

Figure 3.2: Distribution of Rotten Tomatoes and IMDB Scores.

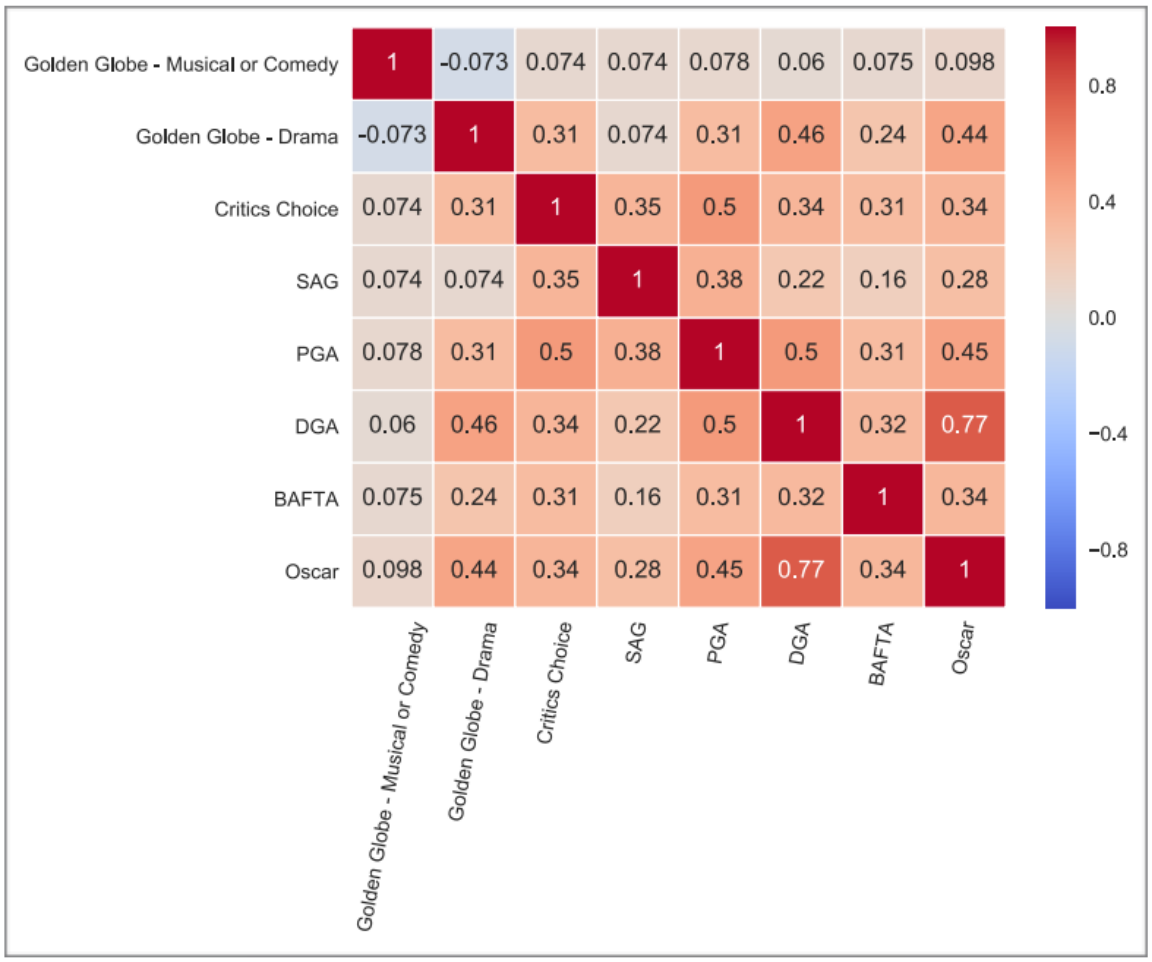

Figure 3.3: Correlation Between Awards.

Chapter 4. Feature Engineering

4.1 Feature extraction

Feature extraction involves transforming raw data into meaningful variables. Key features extracted are:

- **Award Wins and Nominations:**
  - Dummy-coded variables for each award ceremony, distinguishing between wins, nominations, and non-existence.
- **Critical and Popular Ratings:**
  - Numerical predictors based on scores from *Metacritic*, *Rotten Tomatoes*, and *IMDB*.
- **Release Date:**
  - Transformed into quarterly dummies (Q1–Q4) to capture seasonal trends.
- **Genre:**
  - Binary variables for each genre based on *IMDB*'s three-genre classification.
- **Age Segmentation:**
  - Actors' ages were categorized into seven segments for interpretability.

4.2 Feature selection

Feature selection identifies the most important predictors to optimize model performance. The key findings are:

- **Significant Predictors:**
  - *Award Data:* DGA, SAG, and PGA win strongly predict Best Picture and Best Director categories.
  - *Total Nominations:* Predictive across all categories.
  - *Genres:* Biography, History, and War genres are significant for Best Director.
  - *Ratings: Rotten Tomatoes* Audience Score predicts all categories except supporting roles, whereas Critics' Scores are more predictive.

  - *Age:* Actors over 75 are strong predictors for acting categories.

- **Excluded Variables:**
  - Budget and box office data were excluded due to low predictive power.
  - MPAA ratings did not improve accuracy.
  - Some genres, like *Horror* and *Mystery*, were not predictive.

Chapter 5. Model fitting

5.1 Regression

5.1.1 Logistic Regression

Logistic regression is a commonly used linear model for classification tasks. The model estimates the following function:

$$p(y = 1|x) = \frac{e^{\beta_0+\beta^T x}}{1 + e^{\beta_0+\beta^T x}}$$

Here, $x = (x_1, x_2, \ldots, x_p)$ represents the predictors, and $p(y = 1|x)$ is the conditional probability of the event occurring (in this case, an Oscar win) given the observed predictors. Python's Scikit-learn library employs L2-regularization as the default, which means that instead of directly maximizing the likelihood:

$$\min_{\beta} \sum_{i=1}^{n} - \log p(y_i|x_i; \beta)$$

it estimates the coefficients by solving the regularized version:

$$\min_{\beta} \sum_{i=1}^{n} - \log p(y_i|x_i; \beta) + \frac{\lambda}{2}\|\beta\|^2$$

Here, λ represents the regularization strength, which is controlled by the inverse parameter C. Smaller values of C indicate stronger regularization. Scikit-learn's `LogisticRegressionCV` is capable of finding the optimal C value using cross-validation .

Initially, the model included many explanatory variables, with as many as 60 features for some categories, such as lead acting. However, logistic regression struggles with high-dimensional feature spaces. To address this, feature selection was performed using Scikit-learn's `SelectKBest`, which evaluates linear dependencies between variables using F-tests. For most categories, the top 10 predictors were selected. However, for the Best Director category, limiting the predictors to 10 negatively impacted accuracy, so the 15 most relevant predictors were used instead. The coefficients and p-values of the selected predictors are shown in Tables 5.1 and 5.2.

Table 5.1: Applied Features for Lead and Supporting Acting Categories

Variable Name	Coefficient	p-value	Variable Name	Coefficient	p-value
rt_audience_score	0.0231	0.0001	rt_audience_score	-0.0030	0.0653
total_oscar_noms	0.0469	0.0000	rt_critic_score	0.0231	0.0422
best_film_nom [Yes]	0.5551	0.0004	total_oscar_noms	0.0497	0.0025
SAG_win_1 [Yes]	-1.2635	0.0000	SAG_win_1 [Yes]	-0.7086	0.0013
SAG_win_2 [Yes]	2.3804	0.0000	SAG_win_2 [Yes]	0.9888	0.0000
BAFTA_nom [Yes]	0.0831	0.0041	BAFTA_nom [Yes]	0.3434	0.0233
BAFTA_win [Yes]	1.3318	0.0000	BAFTA_win [Yes]	0.3147	0.0000
critics_choice_win_1	-0.2671	0.0072	critics_choice_win_1	-0.3880	0.0069
critics_choice_win_2	0.1336	0.0000	critics_choice_win_2	0.3179	0.0001
GG_drama_lead_win	1.8798	0.0000	GG_supporting_win	1.8449	0.0000

Table 5.2: Applied Features for Best Director and Best Picture Categories

Variable Name	Coefficient	p-value	Variable Name	Coefficient	p-value
rt_audience_score	0.0383	0.0074	rt_audience_score	0.0175	0.0098
total_oscar_noms	0.2132	0.0000	total_oscar_noms	0.2438	0.0000
best_film_nom [Yes]	0.9501	0.0687	SAG_nom_1 [Yes]	-5.0466	0.0020
DGA_nom [Yes]	0.9842	0.0618	PGA_win_1 [Yes]	-0.1940	0.0017
BAFTA_nom [Yes]	0.8979	0.0102	PGA_win_2 [Yes]	1.0347	0.0000

5.2 Classification

5.2.1 Random Forest Classifier

Random forests are ensemble models constructed from decision trees. A decision tree consists of nodes, starting with a "root" node at the top, which has no incoming edges. Nodes that have outgoing edges are termed internal nodes, while those without outgoing edges are called leaves. Internal nodes split the instance space into two or more sub-spaces based on the values of one of the input attributes. For classification tasks, the predicted class corresponds to the majority class in the terminal nodes.

One common metric used for creating splits in a decision tree is the Gini index, which measures node purity. It is defined as:

$$G_m = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}),$$

where $\hat{p}_{mk}$ is the proportion of training samples in the m-th region belonging to the k-th class. A lower Gini index indicates higher node purity, implying that most observations in a node belong to a single class.

Decision trees offer several benefits, such as interpretability, the ability to handle both qualitative and quantitative variables, and graphical representation. However, for this study, dummy variables were used universally across models, so the tree's ability to directly handle categorical variables was not leveraged. Trees are also considered closer to human decision-making processes. Figure 5.1 illustrates a decision tree fitted to the Best Picture training dataset, which consists of 227 films. The most influential predictors are positioned at the top, with their importance diminishing as one moves down

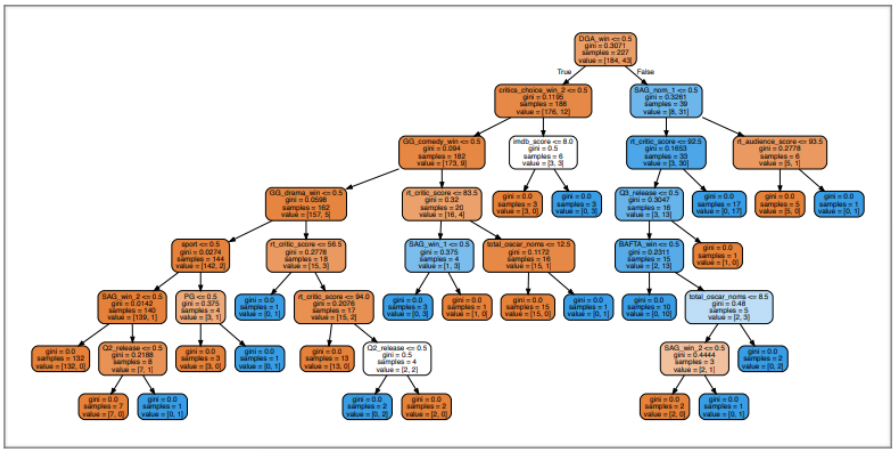

Figure 5.1: Decision tree for the Best Picture category

the tree. Nodes representing the 'winner' class are marked in blue, while those indicating a 'loser' majority are orange. Node purity is reflected by color intensity, with lighter shades (including white) representing higher impurity (Gini index of 0.5).

The tree shown is only for visualization; predictions were made using a random forest, an ensemble method introduced by L. Breiman. A random forest comprises multiple decision trees, where each tree is built on bootstrapped training samples. At every split, only a random subset of predictors (from the p available predictors) is considered, which prevents all trees from relying on the same strong predictors and thereby reduces correlation among the trees. This decorrelation improves model performance.

For this project, 250 estimators (trees) were used in the random forest. Each tree votes on whether a film is classified as a 'winner' or 'loser,' and the majority vote determines the final prediction. This voting mechanism mirrors real-life Oscar decisions, where winners are chosen based on aggregated votes. Another advantage of random forests is their ability to model variable interactions effectively. The importance of each feature in a random forest can be quantified by averaging the impurity reductions across all nodes where the feature is used, over all n trees in the forest. When using the Gini index as the impurity function, this metric is known as the Gini importance or Mean Decrease Gini.

Feature importance scores for the Best Director category are shown in Figure 5.2. The distribution of feature importances is highly skewed: only a few variables significantly influence predictions. Interestingly, some features deemed unimportant in the logistic regression model, such as `imdb_score` and `rt_critic_score`, emerged as important predictors in the random forest classifier.

For each award category, the top 20 features and their Mean Decrease Gini (MDI) scores are summarized in Table 5.3. This comparison highlights differences in variable importance across models and categories.

5.3 ML algorithms

We developed three distinct models to generate predictions for all six categories. Each model was trained using 70% of the dataset, with the remaining 30% reserved for testing. The training data was selected randomly, without considering the year, ensuring an unbiased sample for model development.

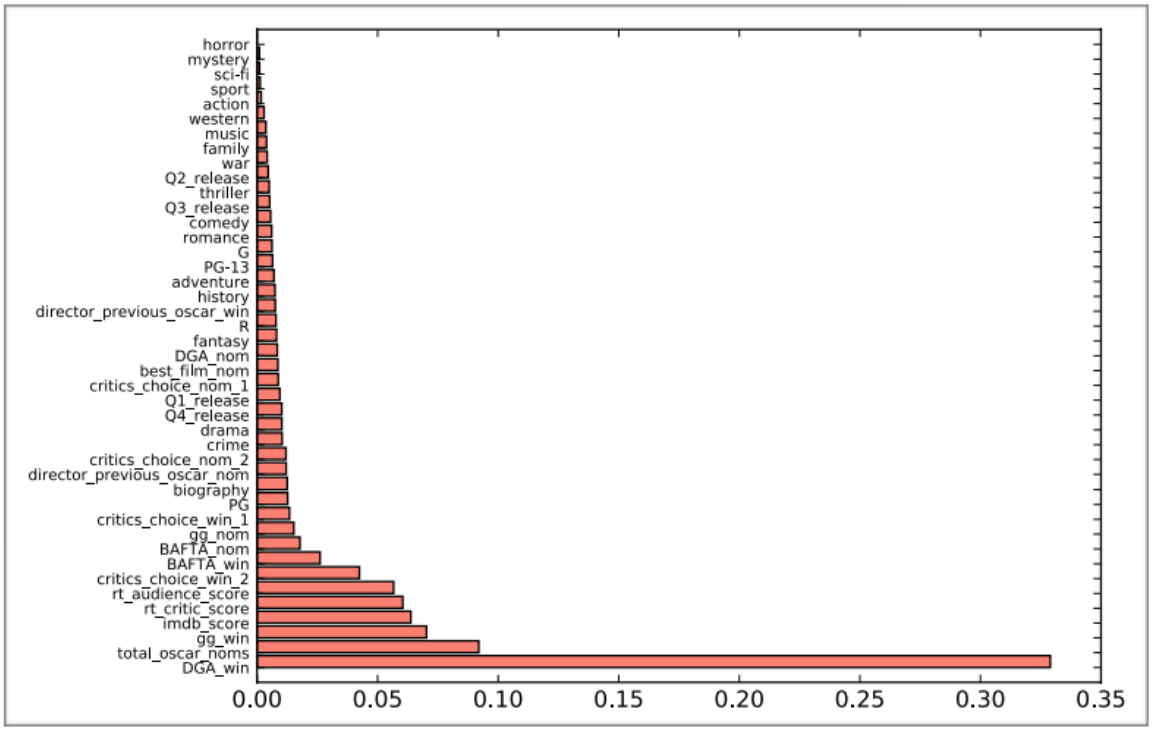

Figure 5.2: The Random Forest Classifier's feature importances for the Best Director category

Table 5.3: Top 20 Features and Mean Decrease Gini (MDI) Scores for Random Forest Classifier by Category

Best Director	MDI	Best Picture	MDI	Lead Acting	MDI	Supporting Acting	MDI
DGA_win	0.329	DGA_win	0.240	GG_drama_lead_win	0.089	GG_supporting_win	0.092
total_oscar_noms	0.092	total_oscar_noms	0.068	SAG_win_2	0.085	rt_critic_score	0.076
gg_win	0.070	rt_audience_score	0.063	rt_audience_score	0.069	imdb_score	0.073
imdb_score	0.064	imdb_score	0.056	rt_critic_score	0.062	rt_audience_score	0.072
rt_critic_score	0.060	PGA_win_2	0.053	imdb_score	0.060	total_oscar_noms	0.071
rt_audience_score	0.057	rt_critic_score	0.052	total_oscar_noms	0.058	BAFTA_nom	0.029
critics_choice_win_2	0.042	GG_drama_win	0.051	BAFTA_win	0.057	SAG_win_2	0.027
BAFTA_win	0.026	critics_choice_win_2	0.035	previous_oscar_noms	0.034	previous_oscar_noms	0.023
BAFTA_nom	0.018	BAFTA_win	0.023	SAG_win_1	0.029	BAFTA_win	0.023
gg_nom	0.015	PGA_win_1	0.022	critics_choice_win_2	0.028	GG_supporting_nom	0.022
critics_choice_win_1	0.013	BAFTA_nom	0.018	Q1_release	0.022	Q4_release	0.022
PG	0.013	SAG_win_2	0.018	best_film_nom	0.022	45-55	0.022
biography	0.013	critics_choice_win_1	0.018	BAFTA_nom	0.017	PG	0.020
director_previous_oscar_nom	0.012	SAG_nom_1	0.015	R	0.016	comedy	0.019
critics_choice_nom_2	0.012	SAG_win_1	0.015	Q4_release	0.016	SAG_win_1	0.019
crime	0.010	Q4_release	0.013	35-45	0.015	best_film_nom	0.018
drama	0.010	SAG_nom_2	0.013	critics_choice_win_1	0.015	R	0.018
Q4_release	0.010	best_dir_nom	0.012	previous_oscar_wins	0.015	Q1_release	0.017
Q1_release	0.009	GG_drama_nom	0.012	GG_comedy_lead_win	0.014	Q3_release	0.016
critics_choice_nom_1	0.009	GG_comedy_win	0.011	SAG_cast_win_1	0.014	critics_choice_win_2	0.016

5.3.1 Support Vector Machines

A Support Vector Classifier (SVC) is a binary classifier that constructs a hyperplane to separate two classes in high-dimensional space, as explained in Tibshirani et al. (2013). In a two-dimensional space, the hyperplane is simply a line. The equation of the hyperplane divides the space into two halves, where points on one side belong to one class, and points on the other side belong to the other class. In my case, the two classes represent Oscar winners and non-winners.

The goal of classification is to assign a test observation, represented as a vector of observed features, to one of these two classes. We classify a test point based on the sign of a function that evaluates its position relative to the hyperplane. If the result is positive, the test observation is classified as a winner, and if negative, as a non-winner. A larger value (in absolute terms) means greater confidence in the classification.

To improve classification accuracy, we aim to find the hyperplane with the maximum margin, which maximizes the distance between the hyperplane and the closest training data points. These points, lying on the edges of the margins, are called support vectors. The maximal margin classifier seeks to maximize the margin while minimizing the number of misclassified points.

In some cases, it is impossible to perfectly separate the two classes with a hyperplane. To avoid overfitting and improve generalization, a support vector classifier allows some misclassification using a "soft margin." The solution to this optimization problem involves slack variables that permit violations of the margin. The tuning parameter C controls the trade-off between margin width and misclassification tolerance. A small C allows more violations, while a large C creates a narrower margin with fewer violations.

For non-linear decision boundaries, the Support Vector Machine (SVM) extends the support vector classifier by using a kernel function to map the data into a higher-dimensional space. The kernel function measures the similarity between pairs of observations and enables SVM to handle more complex decision surfaces. The linear kernel computes a standard inner product, while the Gaussian Radial Basis Function (RBF) kernel can handle non-linear boundaries.

The decision function in the transformed feature space takes the form:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b$$

Here, $K(x_i, x)$ is the kernel function, y_i is the class label, and α_i are the Lagrange multipliers. The predicted class for a test observation is determined by the sign of this function.

To select appropriate values for the hyperparameters C and γ, I used GridSearchCV, a Python tool that evaluates various combinations of parameters by fitting the model to the data. Table 5.4 shows the best parameters found for each category.

Table 5.4: SVM Parameters by Category

Category	**C**	γ
Best Picture	100	0.001
Best Director	100	0.001
Lead Acting	10	0.01
Supporting Acting	0.1	1

SVMs are a powerful classification tool, particularly in high-dimensional spaces, and they can provide accurate predictions. However, one limitation is the lack of transparency in how they make predictions, as they do not easily provide human-interpretable results (Auria, 2008). Additionally, when using SVMs for classification, predicting probabilities requires the use of Platt scaling, which calibrates the output to provide probabilities instead of just binary labels (Platt, 1999).

Chapter 6. Conclusion & future scope

6.1 Findings/observations

Predicting Academy Award winners with high accuracy is achievable using statistical learning techniques. Among the models tested, the random forest classifier emerged as the most effective, likely due to its ensemble nature and ability to simulate human-like decision-making. This aligns well with the voting process inherent in the Oscars. Between 1960 and 2018, the random forest model achieved an impressive 91.5% accuracy in predicting winners across the six main categories. In comparison, the logistic regression model reached 70% accuracy, and the support vector machine (SVM) model achieved 86%. Of all the categories analyzed, the Best Director award was the easiest to predict, while the supporting acting categories proved to be the most challenging.

A key objective of this project was to forecast Oscar winners for a specific year. For the 2018 Oscars, the predictions were highly accurate, aligning closely with forecasts made by experts in mathematics and economics within the film industry. Through this analysis, it became evident that when models fail to predict the winner, it is often because the outcome surprises even industry insiders. Notably, the 2018 Oscars featured very few unexpected winners.

Another goal was to determine the most significant factors influencing Oscar predictions. Results from prior award ceremonies, such as the Directors Guild of America (DGA) Awards, BAFTA, Golden Globe Awards (Drama category), and SAG Awards, emerged as the most critical predictors. Additionally, the total number of Oscar nominations and nominations for Best Picture or Best Director serve as strong indicators of a film's likelihood of winning. Measures of critical and audience acclaim, such as Rotten Tomatoes scores, and in certain cases, genre classifications, also contribute to predictive success. While trends like release dates offer interesting insights into Oscar dynamics, they hold less predictive power overall.

6.2 Challenges

The task of predicting Academy Award winners involves several challenges stemming from the complexity of the process and the dynamic nature of the entertainment industry. These challenges can be broadly categorized into data-related, modeling, and external factors. Below is a detailed description of these challenges:

1. Data-Related Challenges

- Data Availability and Quality
  - **Limited Historical Data:** While data on past Oscars is available, the number of observations (winners and nominees) is relatively small, especially when divided into categories.

This can limit the effectiveness of machine learning models that require large datasets.

- **Incomplete or Missing Data:** Information on older movies or certain award categories might be incomplete or missing, such as revenue figures, ratings, or preceding award results. Handling these gaps without introducing bias is a significant challenge.
- **Inconsistent Formats:** Data collected from multiple sources often requires extensive cleaning and formatting to ensure consistency. For example, budget data may be recorded in different currencies or formats.

- Subjective Data
  - **Ambiguity in Sentiment Analysis:** Critics' reviews and audience opinions can vary widely. Assigning numerical values to subjective opinions is inherently challenging and might not capture the nuances of public and critical reception.
  - **Genre Overlap:** Movies often span multiple genres, making it difficult to analyze the influence of a single genre on award outcomes.

2. Modeling Challenges

- Complex Interactions
  - **Non-Linear Relationships:** The relationship between features like box office revenue, critical acclaim, and nominations is often non-linear, requiring advanced models to capture these patterns effectively.
  - **Overfitting:** Given the limited size of the dataset, there is a risk that complex models like random forests or neural networks might overfit to the training data, reducing their generalizability.
- Feature Selection
  - **Identifying Relevant Features:** Deciding which features to include (e.g., preceding award results, audience sentiment, or release date) and how to weight them can be difficult. Irrelevant or redundant features may dilute the predictive power of the model.
  - **Dynamic Importance of Features:** Factors influencing Oscar wins can change over time. For example, genres or trends popular in the 1970s may not carry the same weight in modern times.
- Multi-Category Predictions
  - **Diverse Award Categories:** The factors influencing awards like Best Picture, Best Actor, and Best Cinematography vary significantly, making it hard to create a unified model applicable to all categories.

3. External Challenges

- Industry Bias and Politics
  - **Campaigning Efforts:** Studios often invest heavily in award campaigns, which can influence the outcomes in ways not reflected in the data. This introduces a layer of unpredictability that models cannot account for.

  - **Personal and Industry Bias:** Voter preferences may be influenced by factors like diversity, inclusion efforts, or industry politics, which are challenging to quantify or predict.

- Surprise Winners
  - **Unpredictable Outcomes:** Occasionally, award winners defy predictions and expectations, possibly due to political, cultural, or emotional factors. These "surprise winners" are nearly impossible to forecast.

- Changing Trends
  - **Evolution of Criteria:** The criteria and preferences for award winners evolve over time. For instance, the increasing importance of diversity and representation in recent years may affect how movies are judged.
  - **Shifting Genres:** Certain genres may gain or lose favor among voters over time, reflecting broader cultural or industry shifts.

4. Computational and Resource Challenges

- Model Training
  - **Computational Costs:** Training multiple models, especially complex ones like random forests or SVMs, can be computationally intensive, particularly when optimizing hyperparameters.
  - **Balancing Performance and Interpretability:** Simpler models like logistic regression are easy to interpret but may lack accuracy, while more accurate models like random forests may not provide easily interpretable insights.

- Cross-Validation and Testing
  - **Limited Testing Data:** Splitting the already small dataset for training and testing can lead to insufficient data for robust validation.
  - **Temporal Bias:** Using data from older Oscars to predict recent winners might introduce bias due to the evolution of voting trends and preferences.

5. Future Uncertainties

- **Unforeseen Events:** External events, such as controversies surrounding nominees or socio-political movements, can dramatically affect award outcomes but are nearly impossible to model.
- **Lack of Predictive Power in Some Categories:** Certain categories, such as Supporting Actor/Actress, are highly subjective and influenced by a broader set of intangible factors, making them less predictable.

6.3 Future plan

The future plan focuses on building upon the existing project framework to improve predictive accuracy, expand scope, and derive deeper insights into the factors influencing Academy Award outcomes. Below is a detailed outline of the steps for future development:

1. Expanding the Dataset

- Including More Years
  - Extend the dataset to include recent Oscar results to account for evolving trends, such as diversity and representation.
  - Incorporate older historical data (e.g., from the 1920s-1950s) to examine long-term patterns.
- Integrating Additional Data Sources
  - **Social Media Data:** Gather public sentiment from platforms like Twitter and Reddit to analyze audience opinions leading up to the awards.
  - **Streaming Platforms:** Include viewership statistics from streaming platforms to assess the impact of accessibility on award outcomes.
  - **Industry Campaign Data:** Attempt to quantify the effects of "For Your Consideration" campaigns by collecting marketing and promotional data.
- Enhancing Feature Set
  - Add new features, such as cinematography styles, technical awards, and representation of marginalized groups, to better capture voting patterns.
  - Include macroeconomic or cultural indicators (e.g., the impact of global events) to evaluate their influence on voters.

2. Refining Models

- Model Tuning and Optimization
  - **Hyperparameter Tuning:** Use techniques like grid search or Bayesian optimization to fine-tune models for better performance.
  - **Ensemble Learning:** Combine predictions from multiple models (e.g., random forest, SVM, and neural networks) to leverage the strengths of each.
- Exploring Advanced Techniques
  - **Deep Learning Models:** Implement neural networks, such as LSTMs, for sequence analysis, particularly in text reviews or historical trends.
  - **Natural Language Processing (NLP):** Use NLP techniques to analyze critic reviews, acceptance speeches, and news articles for deeper insights.

- Handling Imbalanced Data Use techniques like SMOTE (Synthetic Minority Oversampling Technique) or weighted loss functions to address class imbalances, especially for less predictable categories.
- Real-Time Prediction System Build a real-time prediction system that updates predictions as new data, such as results from preceding award ceremonies, becomes available.

3. Enhanced Visualization

- Interactive Dashboards
  - Develop interactive dashboards using tools like Tableau, Power BI, or Plotly for dynamic visualization of trends and predictions.
  - Allow users to explore how different features (e.g., genre, critical acclaim, or budget) affect predictions.
- Trend Analysis Visualize long-term trends, such as the influence of diversity on nominations or the rise of independent films in winning major awards.

4. Addressing Challenges

- Improving Data Quality
  - Collaborate with film industry databases or academic institutions to access high-quality, structured data.
  - Automate data cleaning processes to reduce manual effort and ensure consistency.
- Incorporating Human Judgment Combine statistical predictions with insights from film critics or industry experts to create hybrid models that account for qualitative factors.

5. Broader Applications

- Expanding to Other Award Ceremonies
  - Apply the predictive framework to other prestigious film awards, such as the Golden Globes, BAFTAs, or Cannes Film Festival.
  - Compare the factors influencing different awards to understand unique voting behaviors.
- Recommendations for Film Studios Use insights from the model to advise studios on strategies to increase their chances of winning Oscars, such as timing their releases or targeting specific categories.
- Public Engagement
  - Create an open platform where the public can input their predictions and compare them against the model's forecasts.
  - Publish annual prediction reports for major award categories, providing transparency and fostering interest in the project.

6. Academic and Industry Collaboration

- Publishing Research Publish findings in academic journals or at conferences to share methodologies and insights with the broader data science and film communities.
- Collaborating with Industry Stakeholders Partner with film production companies, marketing agencies, or streaming platforms to gain deeper insights into industry practices.

7. Ethical Considerations

- Ensuring Fairness Investigate potential biases in the dataset (e.g., gender, race, or language biases) and implement methods to mitigate them.
- Transparency Make the predictive process transparent by providing explanations for model decisions, ensuring users understand how predictions are generated.