Predicting Oscar Nominees: A Multi-class Labelling Approach Using Multiple Machine Learning Techniques

Victor Wu
wu.victo@northeastern.edu

Macarious Hui
hui.mac@northeastern.edu

Tianyi Zhang
zhang.tianyi9@northeastern.edu

- The Academy Awards, also known as the Oscars, represent one of the film industry's most prestigious honors. Each year, studios spend millions of dollars to craft marketing campaigns for a limited number of nomination slots across various categories. Identifying the key factors that contribute to a movie's Oscar nomination is a complex task.

- This project aims to not only predict Oscar nominations but also determine the most influential features contributing to these nominations. We will examine a range of factors associated with the movie, its cast and crew, and its reception by critics and audiences.

1. Introduction

The Oscars highlight a film's artistic and technical merits in the movie industry. Each year, the excitement and speculation around the nominations demonstrate the importance of the awards that go beyond the screen. Campaigning for nominations is a high-stakes effort for studios and filmmakers, given the prestige, opportunities, and lucrative media coverage the award show offers. However, the journey to securing a nomination is often opaque and subject to a great deal of uncertainty. As a result, forecasting Oscar nominations can be challenging and difficult due to the weight of cultural, economic, and political factors that contribute to them.

We believe that the use of machine learning can help us better understand the Oscar nomination process. Our project aims to use machine learning algorithms to predict Oscar nominations in different categories. We take a multi-label approach to analyze various factors that contribute to a movie's chances of receiving a nomination, aiming to make the Oscar nomination process more transparent.

Our approach focuses on using a wide range of movie-related information, starting from the development stage to the final Oscar nomination announcements. We've identified key factors that we believe play a significant role in determining a movie's nomination prospects.

These features include, but not limited to:

Figure 1. The epitome of cinematic excellence, the iconic Oscar trophy, gleaming with aspirations of filmmakers around the globe.

- Golden Globe and Screen Actors Guild (SAG) Nominations
- Genre and Motion Picture Rating (MPAA)
- Ratings from IMDb and Rotten Tomatoes
- Net worth of the studio and production company, adjusted for inflation
- Budget and box office performance
- Actor attributes such as sex, age, and previous nomination history

Additionally, as our research continues, we remain open to the possibility of including more features that may provide valuable insights into the Oscar nomination process.

During our research, we encountered numerous articles that tried to predict Oscar winners, each embracing a different set of features and methodologies[3, 1, 5]. Our project diverges from the aforementioned works by focusing on predicting nominations rather than winners. By doing that,

we will have more balanced dataset since only a few movie win Oscar, but more will be nominated. The goal is to identify the key features that contribute to a movie's likelihood of being nominated across the six major Oscar categories:

- Best Picture
- Best Director
- Best Actor
- Best Actress
- Best Supporting Actor
- Best Supporting Actress.

It is worth to note that one movie can win multiple nominations, so rather than a multi-class classification, it is more like a multi-labelling problem.

Specifically, we propose three methods to make the prediction, and we will evaluate the methods against each other.

Binary Relevance with Traditional Models:

- In this approach, we utilize the Binary Relevance for multi-label classification, treating each Oscar category as a separate label. One or more of the following traditional machine learning models can then be employed to predict the presence or absence of each label independently: logistic regression, K-Nearest Neighbors (KNN), and Principal Component Analysis (PCA).

Multi-Label Decision Trees and Random Forests:

- Multi-Label Decision Trees and Random Forests are extensions of decision trees and random forests to handle multi-label data. They make predictions for each label independently, allowing for the possibility of multiple labels being assigned to a single instance.

Neural Networks:

- Neural networks can be designed for multi-label classification by using a Sigmoid activation function in the output layer instead of Softmax. This allows for independent label predictions, enabling a single instance to be associated with multiple labels.

In our approach to predicting Oscar nominations across the six major categories, we adopt a strategy to ensure accurate predictions and predict exactly 5 nominations per year per category. After training our multi-label classification models, we obtain probability scores for each movie across the six Oscar categories. These probability scores represent the likelihood of each movie being nominated in a specific category, then we employ a two-step process. First, we group the movies and their associated probabilities by the release year to create year-specific datasets for each category. This step allows us to analyze nominations on a yearly basis. Second, within each category for a given year, we sort the movies based on their probability scores in descending order. This sorting places the most likely nominees at the top of the list, reflecting the strength of their nomination prospects. From this ranked list, we select the top 5 movies with the highest probability scores as our predictions for Oscar nominations in that category for that year.

This methodology ensures that our predictions align with the historical Oscar nomination process. By considering the top 5 movies based on probability scores, we emulate the actual nominations in each category. This process is particularly valuable when evaluating our models' performance, as it allows for a direct comparison between our predictions and the official Oscar nominations.

Also, this approach significantly influences the process of splitting our data into training, validation, and testing sets. By creating year-specific datasets, we ensure that our predictions can be meaningfully compared and evaluated against the actual Oscar nominations.

2. Motivation

The consumption and recognition of cinema has long been at the heart of the zeitgeist and cultural production. Identifying the important features that correlate with or even influence the Academy Awards can serve as a litmus test for social consciousness. For example, actors tend to only be rewarded when they are older, whereas actresses are oftentimes rewarded when they are younger. This could demonstrate some of the ways in which social attitudes and systems of power play a role in award bodies and online discourse. What does it mean, for example, if genre is an important feature that determines Oscar nominations? Or what does it mean if box office is an important feature for cultural prestige? These questions are but a few of the problems we seek to tackle in our work through machine learning. In addition, understanding what sort of features influence the success of the films at the Oscars can help to understand audience tastes and the important economic structures that belie the supposed artistic meritocracy of the award. Given the increasing fears of the misuse of emerging technologies such as machine learning and artificial intelligence (AI), we believe this project can serve to show how machine learning can be used to diagnose larger cultural issues and patterns and has potential to lead to progressive change.

3. Evaluation

We identified three key metrics and outcomes that would signify a successful project:

Model Performance:

- High accuracy in predicting Oscar nominations across

various categories. Accuracy will measure the proportion of the top 5 predicted nominations that match the actual nominations for a specific category and year.

- Precision, recall, and F1-score for evaluating performance especially in the presence of class imbalances. These metrics will be calculated for each label independently, but instead of treating it as a binary classification problem, these metrics assess the fraction of the predicted top 5 nominations that are correct.

Feature Importance:

- Identification and understanding of key features contributing to a movie's likelihood of being nominated.
- Clear ranking or scoring of feature importance validated through techniques like permutation importance.

Model Interpretability:

- Model interpretability through visualizations, decision trees, and diagrams generated using tools like Matplotlib.

4. Resources

Our analysis will leverage a rich compendium of datasets sourced from reputable platforms like Kaggle and IMDb, which provide a comprehensive view of the movies, their performance metrics, and awards history. Specifically, the datasets we intend to utilize include, but not limited to:

- The Movies Dataset[2]
- IMDb datasets[4]
- Golden Globe nominations dataset[6]
- Oscar nominations dataset[7]
- SAG Award nominations dataset[8]

A high-performance personal computer will be the primary hardware resource; if the computational requirements exceed the capabilities of the personal computer, NEU discovery cluster will be utilized which provides access to high-performance computing resources including GPUs.

The project will be coded in python. Libraries such as Pandas for data manipulation, Scikit-learn and TensorFlow for machine learning, and Matplotlib for data visualization will be employed.

5. Contributions

Our project involves several common tasks that all team members will collaborate on. These include brainstorming and gathering relevant datasets, actively participating in the project's discussions and decision-making, and collectively contributing to the proposal. All members will contribute to the data preprocessing and exploratory data analysis.

- **Tianyi Zhang** will have a primary focus on implementing and evaluating the machine learning models using decision trees and random forests. This includes data preprocessing, model implementation, and performance evaluation.
- **Victor Wu** will explore and evaluate traditional machine learning models, including logistic regression, K-Nearest Neighbors, and Principal Component Analysis. He will determine which is the most relevant traditional machine learning approach based on the data's characteristics and problem requirements. Subsequently, he will make an informed decision on which approach or approaches to implement.
- **Macarious Hui** will contribute to data preprocessing, model implementation, and evaluation. His specific focus will be on the Neural Networks method. He will design the network architecture, including the choice of layers, nodes, and activation functions.

All three team members will collaboratively contribute to code comments, document valuable insights, and jointly compile the final project report, ensuring comprehensive documentation and clear communication of the project's progress and findings.

References

[1] Predicting the oscars using preferential machine learning. *Towards Data Science*.

[2] R. Banik. The movies dataset. https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset, 2023.

[3] S. Deregowski. Oscar-predictions, 2022.

[4] IMDb. Imdb data files available for download. https://datasets.imdbws.com/, 2023. Documentation can be found at http://www.imdb.com/interfaces/.

[5] J. Kim, S. Hwang, and E. Park. Can we predict the oscar winner? a machine learning approach with social network services. *Entertainment Computing*, 39:100441, 2021.

[6] UNIMAD. Golden globe awards. https://www.kaggle.com/datasets/unanimad/golden-globe-awards, 2023.

[7] UNIMAD. The oscar award, 1927 - 2023. https://www.kaggle.com/datasets/unanimad/the-oscar-award, 2023.

[8] UNIMAD. Screen actors guild awards, 1994 - 2020. https://www.kaggle.com/datasets/unanimad/screen-actors-guild-awards, 2023.