DSA 210 Term Project Final Report

Project Title: Predicting the Best Picture Oscar Winner

Student Name: Emir Vargör

Date: January 7, 2026

Abstract

The Academy Awards, particularly the Best Picture category, represent the pinnacle of cinematic achievement. However, anecdotal evidence suggests that winning is less about purely artistic quality and more about adhering to specific industry patterns. This project investigates this phenomenon by analyzing historical data from 1990 to 2020 to predict winners for the 2021–2025 period. Using statistical hypothesis testing with Bonferroni correction and a Random Forest Machine Learning model enriched with TF-IDF and RFE, this study achieved 92% accuracy in predicting winners. The results lead to the rejection of the null hypothesis, demonstrating that industry support (Guild Awards) is a far stronger predictor of success than genre, runtime, or critical acclaim alone.

1. Introduction & Motivation

The Academy Awards (Oscars) carry significant marketing value and cultural prestige. While nominally an award for the "best" film, industry observers have long suspected that winners follow a specific "pattern" driven by campaigning and industry consensus rather than objective film quality.

The primary motivation of this project is to investigate this "pattern" using a data-driven approach. By quantifying features such as genre, critical scores, and precursor awards, this study aims to:

1. Determine statistically significant predictors of a Best Picture win.
2. Build a Machine Learning model to predict future winners.
3. Analyze whether the Academy rewards specific "requirements" over artistic merit.

2. Data Collection and Preparation

To ensure a robust analysis, data was aggregated from multiple sources and enriched using Python (Pandas).

Data Sources:

- **Kaggle:** Historical Oscar nominees and winners (1929–Present).

- **IMDb (via Kaggle & Web Scraping):** Data for genre, runtime, and MetaCritic scores.
- **Golden Globes:** Historical winners data.
- **Producers Guild of America (PGA) (via Wikipedia):** A manually crafted dataset of PGA winners, as this is a key industry indicator.

Data Cleaning:

Missing values for genre, runtime, and critical scores were filled via web searching. Data from 1990–2020 was isolated for training, while data from 2021–2025 was reserved for testing and validation.

3. Exploratory Data Analysis and Hypothesis Testing

To validate feature significance, rigorous statistical testing was performed. A **Bonferroni Correction** was applied to avoid "p-hacking" across the 7 tested features.

- **Significance Threshold :** 0.05 / 7 = 0.0071
- **Null Hypothesis (H0):** The feature has no statistically significant predictive power.

3.1. Non-Significant Features (Fail to Reject H0)

Surprisingly, several features often associated with "Oscar Bait" proved statistically insignificant ($p > 0.0071$):

- **Genre:** Fisher's Exact Test yielded p-values ranging from 0.21 to 1.0.
- **Runtime:** Independent t-test ($p = 0.25$).
- **Release Date:** Chi-Square test ($p = 0.41$).
- **MetaScore (Critical Acclaim):** Independent t-test ($p = 0.024$).
    - *Insight:* While 0.024 is low, it did not meet the strict Bonferroni threshold. This suggests that critical acclaim is necessary but not sufficient; films with perfect scores (e.g., *The Power of the Dog*) often lose to films with lower scores but higher industry support (e.g., *CODA*).

3.2. Significant Features (Reject H0)

The analysis confirmed that industry consensus is the primary driver of success ($p < 0.0071$):

- **Best Director Nomination:** ($p = 0.00003$). A near-mandatory requirement, with *CODA* (2022) being a notable statistical outlier.
- **Golden Globe Win:** ($p = 0.00002$).
- **PGA Award Win:** ($p = 0.0000$). The strongest predictor, showing a 0.64 correlation with winning the Oscar.

4. Machine Learning Methodology

To operationalize these findings, a predictive model was built.

4.1. Model Selection: Random Forest

The Random Forest algorithm was selected to address specific challenges in the dataset:

- **Small Dataset:** With N = 193 samples, Random Forest prevents overfitting (high variance) better than complex neural networks.
- **Class Imbalance:** The model utilized class_weight='balanced' to penalize misclassification of the minority class ("Winner").

4.2. Advanced Refinements

A baseline model was initially created but yielded poor results. The final "Advanced Model" incorporated:

1. **TF-IDF (NLP):** Utilized on text descriptions to capture cross-genre characteristics.
2. **RFE (Recursive Feature Elimination):** To remove noise.
3. **GridSearchCV:** For hyperparameter optimization.

Feature Importance: The model identified the top three predictive features as:

1. PGA Winner (29%)
2. Director Nomination (17%)
3. Golden Globe Winner (13%)

5. Results and Evaluation

The Advanced ML Model demonstrated high reliability in predicting the Best Picture winner.

5.1. Performance Metrics

- **Accuracy: 92%** (An 8% increase over the baseline).
- **Recall: 80%** (Correctly identified 4 out of 5 winners in the test set).
- **AUC Score: 0.93** (Indicates high separability between winners and non-winners).
- **Precision: 0.57** (The model flags the top 2-3 contenders; "False Positives" are typically the runner-ups).

5.2. Generalization & Case Studies (2021–2025)

The model's probability outputs aligned closely with reality:

- **The Favorites (2021 & 2024):** Correctly predicted *Nomadland* (61%) and *Oppenheimer* (87%) as clear favorites.

- **The Complex Case (2023):** Using TF-IDF, the model correctly identified *Everything Everywhere All at Once* (53%) despite its unconventional genre.
- **The Outlier (2022):** The model ranked *CODA* 4th. This is consistent with *CODA* being a statistical anomaly (winning without a Director nomination and Golden Globe).
- **Future Prediction (2025):** The model identifies **Anora** as the clear favorite for the upcoming ceremony, as it shares the statistical profile of past winners (Director nomination + Precursor awards).

6. Limitations

- **Exclusion of Box Office Data:** To prevent data leakage and the "Oscar Bump" effect (where winning increases revenue), box office data was excluded.
- **Campaigning Politics:** Unquantifiable factors such as aggressive studio marketing (e.g., *Shakespeare in Love* vs. *Saving Private Ryan*) cannot be captured by the model.
- **Filmmaker Reputation:** The specific legacy of a director was not weighted, though Director Nomination status served as a proxy.

7. Conclusion and Future Work

This project successfully rejects the Null Hypothesis. The data proves that a Best Picture win is not a random event nor solely defined by artistic metrics like runtime or genre. Instead, it is a pattern of **Industry Consensus**. The strongest predictors are approval from the Producers Guild (PGA) and the Directors branch.

Future Work:

- Incorporating international award data (BAFTA, Cannes Film Festival, etc.) to better predict trends as the Academy globalizes.
- Testing Gradient Boosting models (XGBoost, LightGBM) to potentially improve precision.

8. AI Usage Disclosure

This project utilized Generative AI tools (specifically Google Gemini) to assist in the documentation, code debugging, and structuring of the analysis. All statistical logic and code implementation were verified by the student. Specific prompts are documented within the submitted .ipynb files.