Oscar Nominations

Team Members/ Responsibilities

- Jake Littman - Machine Learning model
- Oneil Anderson - Technologies/Presentation
- Breionna Turner- Database
- Elena Rivera - Presentation/ Dashboard
- Christine Ibrahim Puri - Github Setup/Dashboard

Overview

Our project will analyze movies throughout history. Using the datasets at our disposal, we will identify the variables that relate to oscar performance of actors/actresses/directors and writers. In our project, we will use the tools that we have learned throughout the bootcamp to create a full analysis on the data provided from the Oscar nominations. We will also address the questions with this analysis.

Questions to Address

- Why do certain genres outperform others?
- What feastures influence movie nominations?
- What genres and production companies have the highest probability of being nominated for an Oscar at the next academy award?

Technologies

Data Cleaning and Analysis

We will use Python to Extract, Transform and Load and perform meaning analysis over the data. We will be using the ImbalancedLearn, Pandas, Numpy, and SciKitLearn packages.

Database Storage

We will use PostgreSQL as our relational database storage.

Machine Learning

Our machine learning models will leverage SciKitLearn and ImbalancedLearn Python packages. Due to the oscar imbalance, we will use random over/undersampling, SMOTE, Cluster Centroid Undersampling, and/or SMOTEEN.

Dashboard for Presentation

We will leverage D3 Javascript/Tableau to create our dashboard to present our findings.

Database

Software used: Jupyter Notebook [Pandas], PostgreSQL

Cleansing the datasets:

The database has come together from bringing in the *movie_metadata.csv* and *oscars_movie*.csv datasets.

The data in both were cleaned via regex processing and taking out null values. In addition, *release_year* and *nominated* columns were added to the *movie_metadata*.

Joining the datasets:

From there, the data was exported into a Postgres database and then the datasets were combined into the combination_table.csv. The datasets were combined on the Oscar's '*film*' column with the Movie's '*title*' column.

The columns in the combined dataset are as follows: *budget, release_year, revenue, runtime, title, genres, nominated, winner, category.*

Machine Learning

Models:
We will use Logistic Regression and Random Forest Models. Random Forest Models are more likely to overfit the data, since they can split on multiple features whereas Logistic Regression handles better where the dimensionality is limited (single or limited variables).

Feature engineering/selection:
After data preprocessing, we selected genre (dummy), runtime, budget, and category as our selections.

As we continue to build on our model, we believe genres will have a significant correlation with movies getting nominated

Training and Testing Data Sets:

We imported the train_test_split model to split the dataset 50/50 by movie titles

Dashboard Integration

Software Used: Tableau

Budget, Revenue, Nominations

- We will use three bar charts side-by-side to display the budgets, revenues, and nominations for each genre in one view.

Average Budget | Oscar Nominations

- A scatter plot could be used to display the relationship between the average budget of films and the number of nominations films received to view if there is any significant positive or negative relationship between the two variables

- A bar chart could also be used along with Tableau Calculation field to form groups of "High", "Medium", and "Low" with regard to the average spending. For example, a budget of 300k - 500k would be Low, 500k - 1M would be medium, and >= 1M would be High.

Genre | Category

- Amongst the categories that receive Oscar awards (Actor, Actress, Art direction, Directing etc), we will display the genres that receive the most nominations in each category.

- The columns of the bar chart would be divided by the categories, and the bars would be color-coded according to genres, with the number of nominations on the y-axis

Interactivity

- Creating sets that group a range of years, for example 1990- 2000, 2001 - 2011, etc would create additional dimensions that would allow a user filter by years, and observe changes in relationships between genre and category over the years.

Dashboard Integration

Example of side-by-side bar charts showing the Budget, Revenue, and Nominations for each genre