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 Oscar nominated movies throughout history. This topic was chosen for our analysis because of the large amount of data available and the common interest in movies between our team members.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 features influence movie nominations?
- What genres have the highest probability of being nominated for an Oscar at the next academy award?

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.

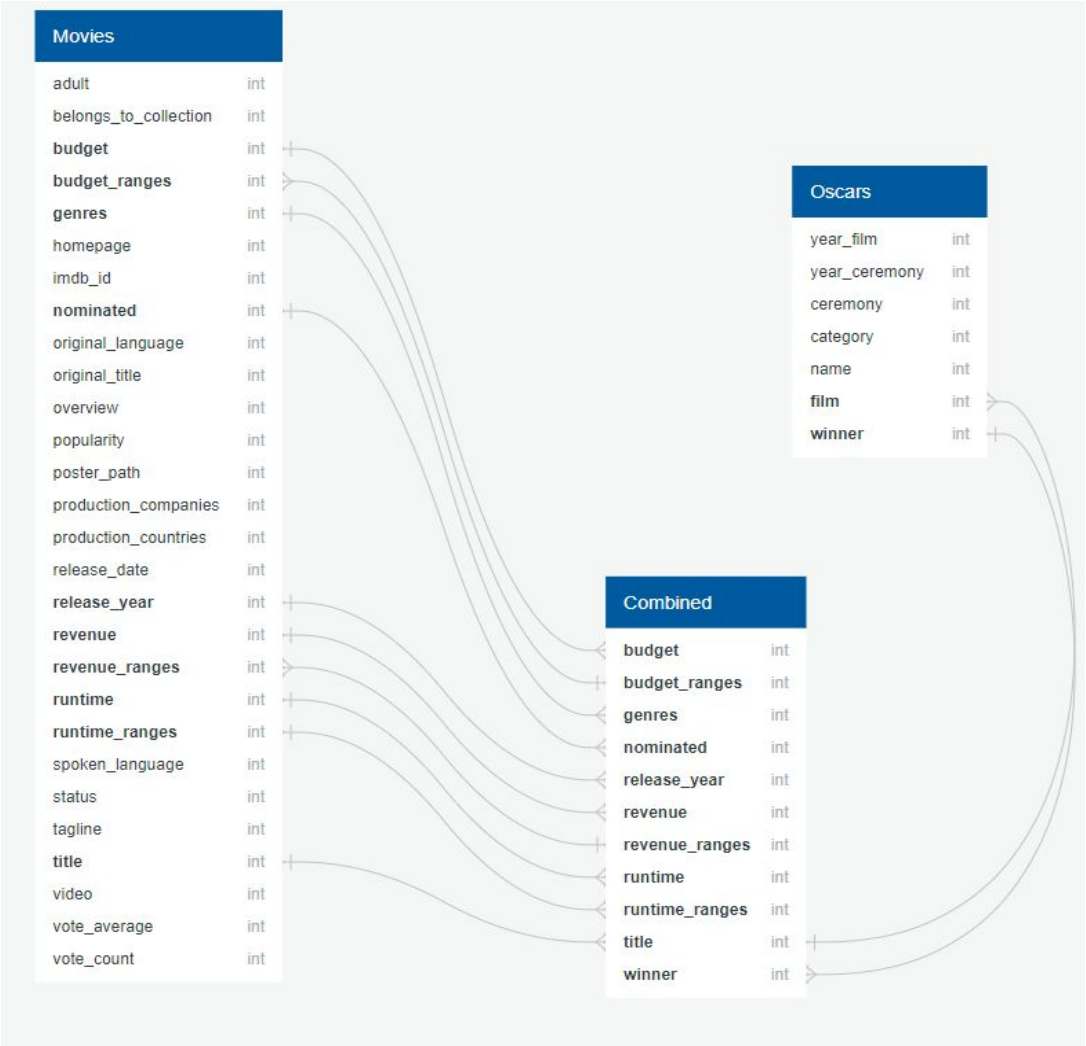
The data in both were cleaned via regex processing and taking out null values. In addition, *release_year* and *nominated* columns were added to the *movie_metadata*.

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.*

Kaggle Movie Dataset

Kaggle Oscar Movie Dataset



Entity Relationship Diagram (ERD) of tables

Machine Learning

Models: Logistic Regression and Random Forest

Observations: We imported the train_test_split model to split the dataset 50/50 by movie titles. There are 4,264 movie titles in the population that were used.

Feature Selection:

Independent Variables: Genre, Budget_Ranges, Revenue_Ranges, Runtime_Ranges (Dummy)

Dependent Variable: “Nominated”

- 0 = Was not nominated for an Oscar
- 1 = Was nominated for an Oscar

Model Summaries and Feature Importance

Sampling Methodology	Model	BA_Score	Confusion Matrix	pre	rec	spe	f1
Naive Random Oversampling	Logistic Regression	0.67	[[18 11] [290 747]]	0.99	0.72	0.62	0.83
Cluster Centroids Undersampling	Logistic Regression	0.69	[[28 1] [612 425]]	1	0.41	0.97	0.58
SMOTE	Logistic Regression	0.57	[[7 22] [113 924]]	0.98	0.89	0.24	0.93
SMOTEENN	Logistic Regression	0.63	[[10 19] [88 949]]	0.98	0.92	0.34	0.95
Random Forest	Random Forest	0.78	[[22 7] [210 827]]	0.99	0.8	0.76	0.88
Random Forest w/ Easy Ensemble Classifier	Random Forest	0.67	[[20 9] [370 667]]	0.99	0.64	0.69	0.78

Feature Importance	Feature
0.081591422	budget_ranges_High
0.074069895	runtime_ranges_High
0.069631364	genres_Drama
0.068413632	budget_ranges_Medium
0.063683912	revenue_ranges_Medium
0.062838783	genres_Comedy

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