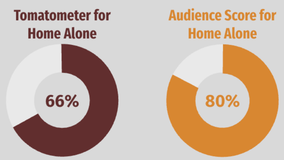
UNLOCKING MOVIE MAGIC: PREDICTIVE FACTORS TO FILM SUCCESS

WHAT ARE THE LEADING CONTRIBUTING FACTORS TO A MOVIE'S SUCCESS?

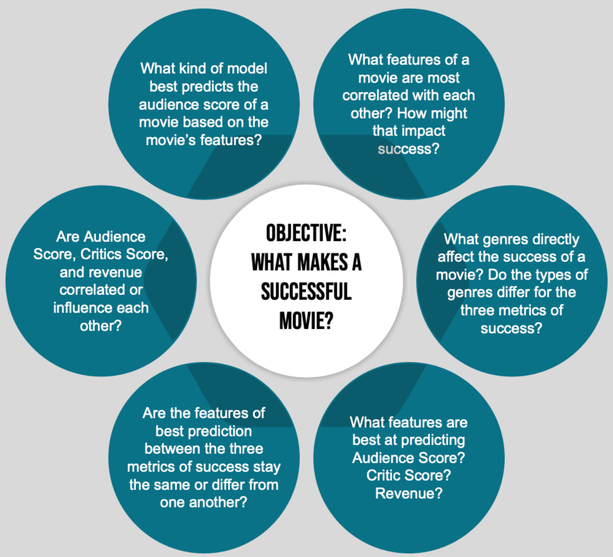
INTRODUCTION

Movies are inherently subjective and vary in popularity. Companies like Netflix, IMDb and Rotten Tomatoes rate movies and measure popularity using their own metrics. For example, IMDb has two different metrics, IMDB Score and Popularity Score, while Rotten Tomatoes rates movies using a Tomatometer and an Audience Score. For example, the Rotten Tomatoes ratings for the movie Home Alone (1990) are shown below.



Why are these scores different and what metrics are these companies using to rate these movies and their popularity? We wanted to see if there was a way to use the metrics to predict a movie's success. For our data, we pulled movie data from TMDb on April 11th, 2024 using an API. Additionally, we used movie data from Rotten Tomatoes for our analysis.

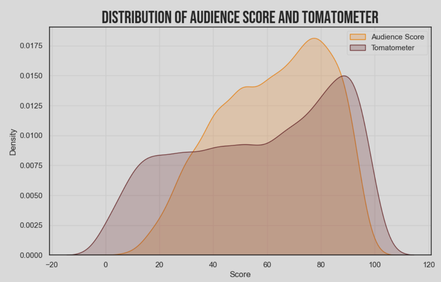
OBJECTIVE

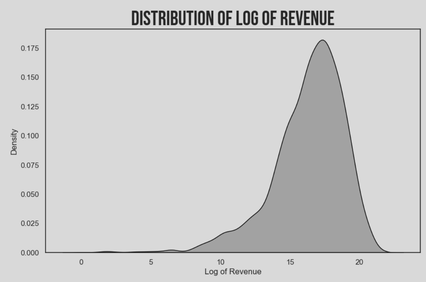


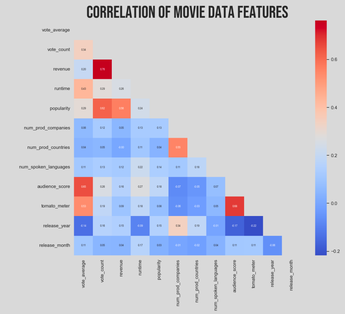
DATA EXPLORATION

Exploratory data analysis is a crucial initial step in uncovering insights from datasets, providing a foundational understanding of underlying patterns and relationships. In our analysis, EDA reveals valuable insights such as revenue distribution and correlations between various movie attributes.

- **Distributions**:
  - The tomatometer ratings seem to have lower scores compared to the audience scores.
  - Revenue seems to have a normal distribution when normalized to log form.
- **Correlations**:
  - Popularity and revenue are positively correlated with a value of 0.56.
  - Tomatometer and release year are negatively correlated with a value of -0.22.
  - Vote average (IMDb score) and runtime are positively correlated at 0.43.







FEATURE SELECTION

We used Generalized Linear Models to predict each target variable and performed feature selection based on the significant p-values derived from the summary model.

We used a generalized linear model (GLM) as our baseline model and compare its performance to that of extreme gradient boosting and neural networks.

The table to the right shows the selected features of each model. We can see clear distinctions in different genres, number of spoken languages, as well as movie ratings. These features were used to train the predictive models.

- `y_audience = glm(audience_score ~ ., family = "gaussian")`
- `y_critics = glm(tomato_meter ~ ., family = "gaussian")`
- `y_revenue = glm(revenue ~ ., family = "gaussian")`

FEATURE SELECTION COMPARISON

	Audience	Critics	Revenue
vote_average	x	x	x
vote_count	x	x	x
runtime	x	x	x
popularity	x	x	x
action	x	x	x
science_fiction	x		
adventure	x		
drama	x	x	x
crime		x	
thriller	x	x	
fantasy	x	x	
comedy	x	x	x
romance	x	x	
western		x	x
mystery	x		
animation	x	x	x
family	x	x	x
horror	x	x	
music	x		
history	x		
documentary	x	x	x
num_prod_companies	x		x
num_spoken_languages		x	
pg_13	x	x	
r	x		x
pg_13		x	x
g		x	x
release_year	x	x	x
release_month		x	

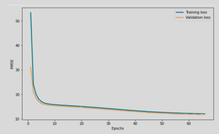
PREDICTIVE MODELS

To answer the questions regarding prediction, we created a feed-forward fully connected neural network for each success metric. Subsetting the features based on the feature selection of the GLMs, we ran these models and hypertuned on a normalized training and validation set.

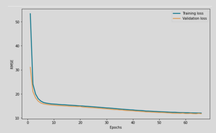
Using a grid search, we hypertuned based on the following parameters; number of hidden layers, size of layers, optimizer, activation function, dropout rate, L1 and L2 regularization, and learning rate.

Additionally, we also used the xgb package to train and test XGBoost Regression models for each success metric. Next, we will review the results of the prediction models.

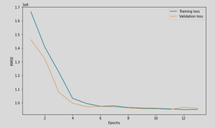
AUDIENCE NEURAL NETS MODEL



CRITICS NEURAL NETS MODEL



REVENUE NEURAL NETS MODEL



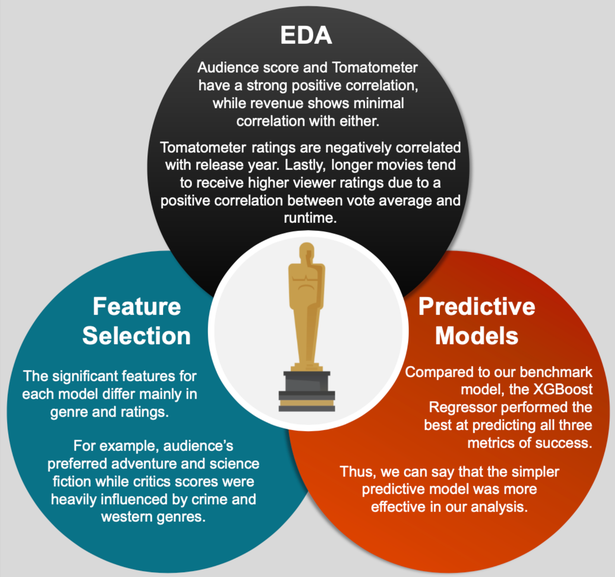
RESULTS

We used the GLM as a baseline model to evaluate the performance of XGBoost and Neural Network models. This allowed us to measure the variation across the models using the RMSE metric. Based on the results we observed that the XGBoost algorithm outperformed both Neural Network and GLM models. In contrast, we observed that the Neural Network model for revenue performed worse in the train and test set compared to the GLM model.

NORMALIZED RMSE COMPARISON

	GLM	XGBoost	Neural Network
Audience - Train	1	0.686	0.853
Audience - Validation	1	0.668	0.779
Audience - Test	1	0.682	0.776
Critics - Train	1	0.849	0.944
Critics - Validation	1	0.820	0.902
Critics - Test	1	0.850	0.923
Revenue - Train	1	0.794	1.090
Revenue - Validation	1	0.802	0.990
Revenue - Test	1	0.750	1.307

CONCLUSIONS



Agustina Zuckerberg, Lindsay Strong, Patricia Schenfeld, Shriya Chinthak

April 2024