Data Strategy Plan for AAA

Adrian Cuyugan

June 4, 2024

Project Overview and Business Objectives

Objectives

- Develop a data-driven project identifying the influential factors in making the next successful movie title.
- Assess available resources such as analytical infrastructure and licenses, data sources, personnel and budget requirements.
- Plan the data collection, preparation and exploration methodologies, perform relevant statistical tests and develop machine learning algorithms to prescribe a recommended movie title.
- Define a key metric that numerically score a successful movie title.
- Construct a project plan.

Funding

- Partner with movie-related websites and streaming services for databases available for non-commercial use.

Personnel Requirements

- Data Scientists
- Data Engineers
- Movie Consultants

Business Objectives

Project Timeline

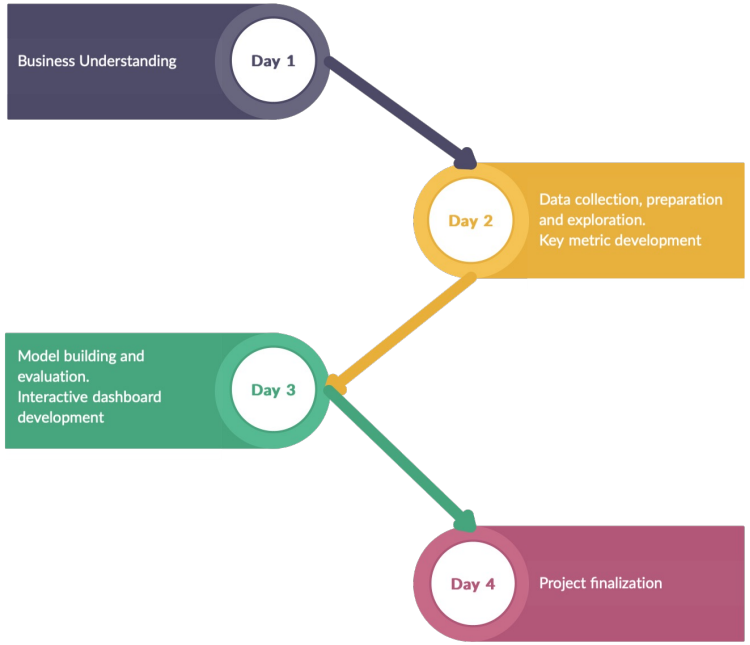

Figure 1 – 4-Day Project Timeline

Data Governance

- Public non-commercial datasets will be used for this project.
- Ensure that any web-scraped data complies with terms and conditions of the source websites.
- Compliance with international data privacy laws.

Bias Mitigation

- Check for fairness in any biases in the data and model predictions.
- Ethical consideration when prescribing movie title recommendations.

Data Understanding and Preparation

	Data Selection	Data Transformation	Data Exploration	Data Cleansing
Tasks	• Load IMDB database from IMDB website. • Look for external public data sources relevant to the study. • Web-scrape movie-related websites and databases. • Create data dictionary of datasets used in the analysis.	• Merge IMDB tables using index columns that join movie title and actor's names. • Filter only relevant rows and fields to minimize resource usage.	• Load technically correct data tables • Review data quality and completeness, properly attribute the datatypes	• Clean data • Construct data (derive attributes, normalize, standardize, etc.) • Extract features defining a key metric
Issues	• Loading big dataset on a 12GB RAM cloud infrastructure leads to runtime error.	• Filtering columns and rows require loading the whole Gzip table into the RAM. Workaround is to chunk size and load one by one. • External data sources from Kaggle and MovieLens are already filtered by top x rows, so the samples are already incomplete.		
Inputs	• title.basics.tsv.gz, title.ratings.tsv.gz • Kaggle datasets • MovieLens datasets	• Internal tables loaded into Python	• Technically correct data tables	• Exploratory report • Quality report • Enhanced data dictionary
Outputs	• Data dictionary • Correctly loaded filtered tables	• Technically correct and joined tables	• Exploratory report • Quality report • Enhanced data dictionary	• Cleansing documentation • Post-transformation quality report • Final data dictionary
Tools	• Google Colab • Python and other libraries • Amazon Sagemaker if needed more resources	• Google Colab • Python and other libraries	• Google Colab • Python and other libraries • Tableau Desktop Public	• Google Colab • Python and other libraries

Legend: Currently available, Proposed, Fallback/Contingency

Highest Rated Movie and Hit Movie Metrics

Highest Rated Movie:
Weighted Rating of Movie Rating and Votes

To balance the rating of movie titles with the number of votes, Bayesian statistics is used.

$$WR = \left(\frac{votes}{votes + top_{1000}}\right) * rating + \left(\frac{top_{1000}}{votes + top_{1000}}\right) * rating_{avg}$$

Figure 2 – Weighted Rating Formula

This provides a mathematical formula to update the calculated average movie rating when a new vote is accounted that particularly could affect the rating of movies with lower number of votes.

Metric

- The ratio of gross and budget was set as initial key metric in identifying the success of a hit-movie based on different factors such as Facebook likes of directors and cast, movie score, movie duration and critic and user reviews.

$$KeyMetric = \left(\frac{gross\_sales}{movie\_budget}\right)$$

Figure 3 – Key Metric Formula

- It has been found that this model has a predictive power of 0.7 when ran on a training model.
- Based on this key metric, the most influential factor is still the movie score.

Movie Title

Using the popularity based on number of votes of movies belonging in the most profitable genres (Drama, Adventure, Action, Comedy and Sci-Fi), the recommended movie title that is predicted to be popular:

Oscar and Jane Austen:
An Idiotic and Pointless Story of Awesome Teenagers

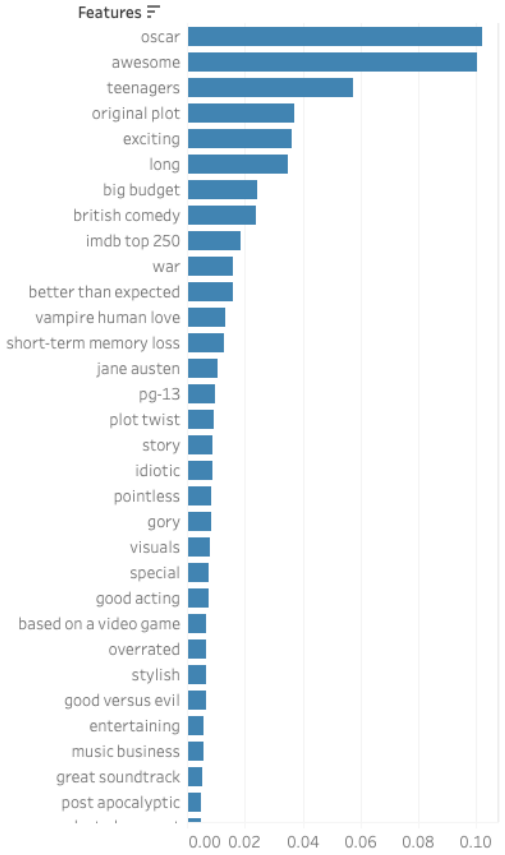

Figure 4 - Feature Importance Plot

Correlation between movie popularity and gross collection in Millions
$R^2 = 0.45$

Evaluated model performance
$R^2 = 0.80$

Appendix

Exploratory Data Analysis

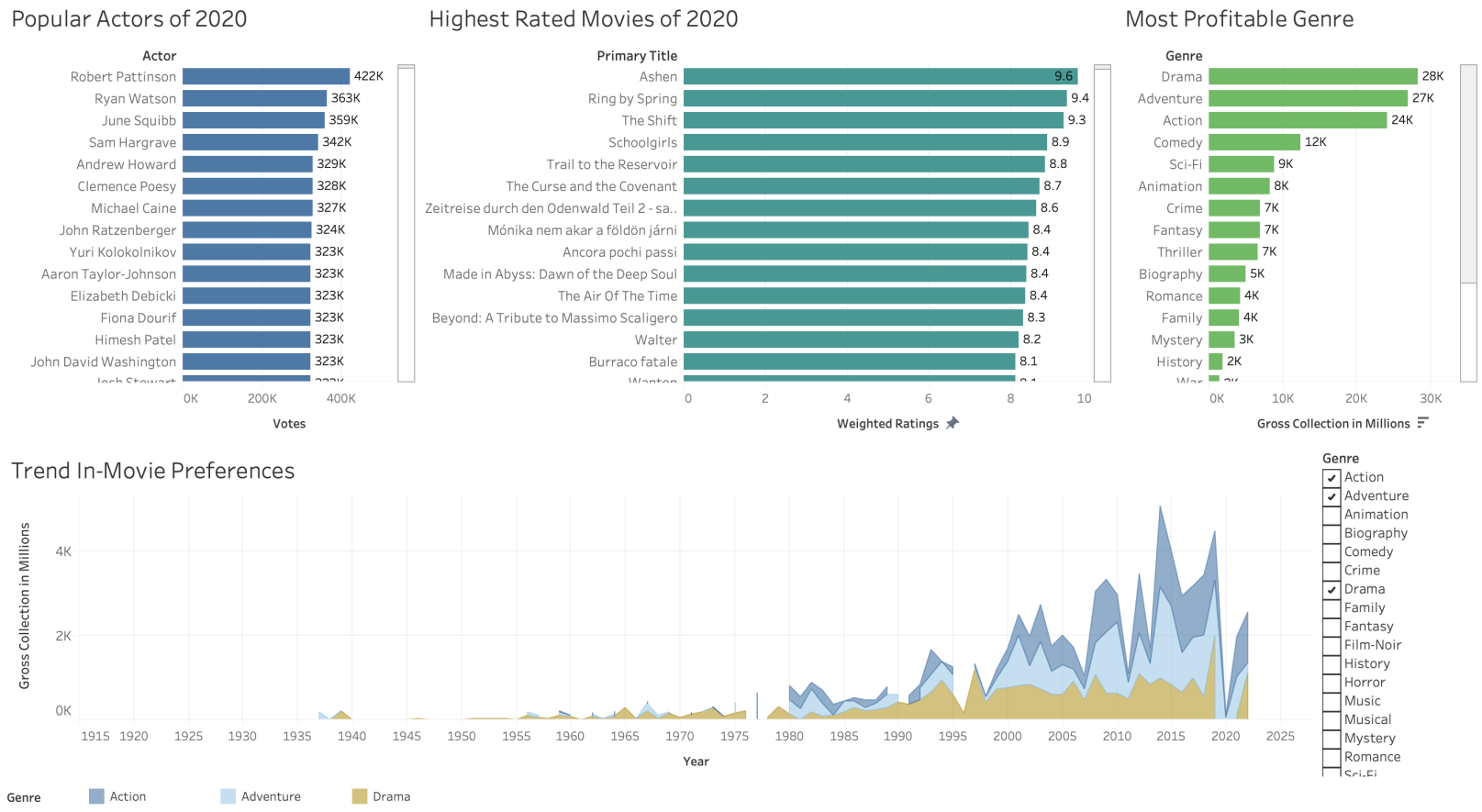

Figure 5 – Interactive Dashboard of EDA

Interactive dashboard can be access here: https://public.tableau.com/app/profile/foxyreign/viz/AAAMovies/Summary

Thank you