ProjectProposal

March 21, 2025

1 As Movie Budgets are Increased, Do Review Scores and Movie Popularity Correlate?

Authors: James Allen & Maitland Huffman

1.1 Project Summary

In this project, our goal is to determine if there is a positive, or any correlation with regards to movie budgets, popularity and their review scores. Using the full tmbd movies dataset, this will allow us to use classification to determine any correlation between the three variables of movie budgets and movie review scores.

1.2 Problem Statement

Does increasing the budget of a movie increase said movie's review score and/or popularity, or is there no significant correlation between the three aspects?

1.3 Dataset

The dataset that we will use for our machine learning project is the full tmbd movies dataset located here.

Instances: The dataset contains 1,142,342 rows, allowing for in-depth analysis across a wide range of films. Attributes: The dataset features a variety of attributes, including:

- Budget
- Revenue
- Release date
- Genres
- Production companies
- Cast and crew information
- User ratings
- Critical reviews
- Box office performance

1.4 Exploratory Data Analysis

<Complete for **Project Progress**> * What EDA graphs you are planning to use? * Why? - Add figures if any

<Expand and complete for the **Project Submission**> * Describe the methods you explored (usually algorithms, or data wrangling approaches). * Include images. * Justify methods for feature normalization selection and the modeling approach you are planning to use.

1.5 Data Preprocessing

<Complete for *Project Progress*> * Have you considered Dimensionality Reduction or Scaling? * If yes, include steps here.
* What did you consider but *not* use? Why?

<Expand and complete for **Project Submission**>

1.6 Machine Learning Approaches

<Complete for **Project Progress**>

- What is your baseline evaluation setup? Why?
- Describe the ML methods that you consider using and what is the reason for their choice?
  - What is the family of machine learning algorithms you are using and why?

<Expand and complete for **Project Submission**>

- Describe the methods/datasets (you can have unscaled, selected, scaled version, multiple data farmes) that you ended up using for modeling.
- Justify the selection of machine learning tools you have used
  - How they informed the next steps?
- Make sure to include at least twp models: (1) baseline model, and (2) improvement model(s).
  - The baseline model is typically the simplest model that's applicable to that data problem, something we have learned in the class.
  - Improvement model(s) are available on Kaggle challenge site, and you can research github.com and papers with code for approaches.

1.7 Experiments

< **Project Progress** should include experiments you have completed thus far.>

<**Project Submission** should only contain final version of the experiments. Please use visualizations whenever possible.> * Describe how did you evaluate your solution * What evaluation metrics did you use? * Describe a baseline model. * How much did your model outperform the baseline? * Were there other models evaluated on the same dataset(s)? * How did your model do in comparison to theirs? * Show graphs/tables with results * Present error analysis and suggestions for future improvement.

1.8 Conclusion

<Complete for the **Project Submission**> * What did not work? * What do you think why? * What were approaches, tuning model parameters you have tried? * What features worked well and what didn't? * When describing methods that didn't work, make clear how they failed and any evaluation metrics you used to decide so. * How was that a data-driven decision? Be consise, all details can be left in .ipynb