Title: *What Makes A Movie?* How Features of a Movie Can Predict its Revenue

Group Members: Jayden Booker, Mavis Gao, Natassia Lowe, Opeyemi Ogundele

Central Hypothesis: What features of a movie contribute the most to its revenue?

Problem Description: *What is the problem the project will seek to address? What would a solution look like? Who would stand to benefit from a solution to this problem? How will you know if you've solved it?*

The project wants to see what factors will contribute to the revenue of films. The solution to the problem will be a list of features that contribute to explain the variance in film revenue, which is listed in decreasing order of their coefficient value. To check if we answer the question, i.e., if we find the features that explain most of the variance of the film revenue, we check the accuracy of the model. To be more specific, for the models we train, if the test set accuracy is high, which means the models successfully predict the revenue, we pick the features from the model and say these features are the features contributing to the variance of the film revenue. And we will continue to tune the model until the test set accuracy is high and obtain the features. This will benefit those in the film industry to determine different aspects of their film making process and specifically the budget that can be used for the film to make profit.

Algorithms(to consider):

- Lasso Regression (Use Library)
    - We prefer Lasso regression because it would 1) automatically shrink unnecessary features to zero to ease feature selection process compared to logistic regression, 2) avoid multicollinearity in film features)
- Random Forests (Use Library)
- Neural Networks (Implement) - MAVIS
- KNN (Implement)

Data:

We are using the IMDB and TMDB movie metadata from kaggle, which is a standard repository dataset with 921371 samples (films, including features such as cast, revenue, etc. Here is the link: https://www.kaggle.com/datasets/shubhamchandra235/imdb-and-tmdb-movie-metadata-big-dataset-1m

Experiments:

1. **Data pre-processing:**
    a. Result: Uncleaned data → cleaned, normalized train/test set
    b. Methods:
        i. clean and standardize the inputs; exclude samples with missing features from the train or test set.
        ii. splitting the data into train/test sets: 70% training and 30% testing
2. **Model selection and training:**
    a. Lasso regression:
        i. Objective: use a Lasso regression to predict continuous revenue, capturing linear relationships
        ii. Experimental design: We want to see whether 1) computer-selected features or 2) human-selected features will allow Lasso regression to yield higher predictive ability (i.e., higher prediction accuracy).
            1. Use Lasso regression to find the features with non-zero coefficients that explain the variance in the data.
            2. Manually select features that humans believe to contribute to prediction. Create and feed the features-selected dataset to the Lasso regression model.
        iii. Hyperparameter tuning: use k-fold cross-validation for tuning hyperparameters and use *k*-fold validation to assess model performance consistently across different subsets: 1) regularization strength (alpha)
    b. Random Forest:
        i. Objective: use a random forest regressor to predict continuous revenue, capturing non-linear relationships
        ii. Hyperparameters: 1) number of trees, 2) maximum depth, 3) minimum sample split
    c. Neural network:
        i. Objective: use a neural network to predict continuous revenue, capturing complex, non-linear relationship
        ii. Network: use a simple feedforward, full-connected architecture initially
        iii. Hyperparameters:
            1. Number of layers, number of neurons by layer
            2. Learning rate, batch size, number of epochs

3. **Result:**
    a. Expected result:
        i. The top features that explains most variance of the dataset (by Lasso regression)
        ii. Prediction accuracy of test set for Lasso regression, RF, and neural network
        iii. Confidence intervals of the three models for the robustness of the model
    b. Validation:
        i. Evaluation metrics: use R-squared that measures the proportion of variance in the revenue that is explained by the model.
        ii. Statistical validation: compute the confidence intervals for predictions through bootstrapping for all three models
        iii. Use cross-validation result (R-squared) and test set accuracy to measure the generalizability and performance of each model to determine what the algorithm deems the most important features that helps it predict the revenue for a particular movie. Because we have a total sample size of 1 million entries, we would like to experiment with training our algorithms on a sample size of 5,000, 10,000, 50,000, and 100,000 entries.

Impacts: *What do you plan to explore in terms of the potential societal impacts of this work? In your final paper, you will describe who stands to benefit from this work. Are there potential negative impacts on populations or other ethical/social concerns with respect to this work? You do not need to flesh out all of the details for the proposal, but instead provide ideas for what topics you will explore for your final paper.*

In the final paper, our group plans to explore how our results could affect the film industry, specifically the way movies are produced. If we find that certain features of movies such as a high budget, certain actors or directors being involved, low or high runtime, etc. could lead to higher revenue at the box office, then that could influence what features are emphasized in the movie making process or what movies get picked up by production companies. There could also be potential concerns with this project if certain demographics in casting are underrepresented. It may very well be the case that movies that have a higher number of white and/or male cast members could make more in revenue due to the prevalence of misogyny and racism in the movie and marketing industry. In general, we believe that if producers were to rely on our model to decide which movies to fund, it could lead to a lack of variety in the film industry, create a negative feedback loop of less diverse actors/actresses on screen, and may influence the creation of films that were made with the priority of maximizing revenue instead of creativity and new and fresh ideas.

References:

Kaggle.com