PREDICTIVE MODELING ANALYSIS FOR MOVIE RATINGS HELPED FORECAST IMDB SCORES WITH 57% ACCURACY

THE CLIENT

A university team project aiming to predict IMDb ratings for 12 upcoming movies using historical data from over 2,000 IMDb-listed movies.

THE CHALLENGES

- Developing a regression model that predicts IMDb scores for unreleased films.
- Handling data complexity while avoiding issues like overfitting, multicollinearity, heteroskedasticity.
- Identifying the most significant predictors from a large dataset of movie characteristics, cast and crew information, and production details.

THE APPROACH

Phase 1: Data Exploration and Feature Engineering

- **Data Collection & Processing**: The IMDb dataset had over 20 variables therefore Initial preprocessing excluded irrelevant identifiers (movie titles, IDs, and IMDb links to avoid noise in the data)
- **Feature Engineering**:
  - New features were created to improve model performance (directors and cinematographers were labeled as "Good" based on whether their average IMDb score exceeded the median across the dataset)
  - Dummy variables were created for categorical features (genre, language, and production company size (e.g., "Big Distributor" for distributors with the most movie titles).
- **Numerical Data Analysis**: Boxplots and histograms revealed that most movies with moderate budgets (with a right-skewed distribution), and durations followed a close normal distribution.

Phase 2: Model Development

- **Regression Models**: 30+ polynomial regression models were run on numerical variables (budget, duration, and release year). The optimal polynomial degree (from 1 to 5) was selected using cross-validation, where the lowest Mean Squared Error (MSE) guided the final degree selection (duration was modelled with a 5th-degree polynomial, while budget required a linear model)
- **Backward Elimination**:
  - Backward elimination was used to remove statistically insignificant variables (those with p-values greater than 0.05) to refine the model and prevent overfitting.
  - Significant predictors retained: duration, budget, number of news articles, IMDbPro movie meter, and genre-based dummy variables (e.g., action, western, animation, drama).
- **Handling Heteroskedasticity**: Diagnostic tests for heteroskedasticity, including residual plots and the NCV (non-constant variance) test, were applied. Variables such as movie budget and duration showed signs of heteroskedasticity, which was corrected using robust standard errors in the final model.

Phase 3: Model Testing and Validation

- **Cross-Validation**: K-fold cross-validation ($K = 5$) was used to assess the model's out-of-sample performance. This process ensured that the model avoided overfitting by splitting the data into training and validation sets.
- **Final Model**: The final regression model included 15 variables that passed through feature selection and backward elimination.
  Final Formula: IMDb Score = β_0 + β_1 (release year polynomials) + β_2 (duration polynomials) + β_3 (budget) + β_4 (news articles) + β_5 (faces) + β_6 (IMDbPro movie meter) + … + β_n (genre & production dummies)

THE RESULTS

- **Model Performance**: The model achieved an R-squared value of 57%, meaning it explained 57% of the variance in IMDb scores for the training dataset and a Mean Squared Error (MSE) of 0.56, showing a relatively accurate model given the complexity of predicting IMDb scores.
- **Predictor Insights**: The most significant variables that contributed to a movie's predicted IMDb score included:
  - **Duration**: Longer movies generally scored higher, modeled with a 5th-degree polynomial.
  - **Movie Budget**: High-budget movies were associated with better ratings, although this effect weakened with extremely large budgets.
  - **Good Director/Cinematographer**: The presence of a well-rated director or cinematographer significantly boosted a movie's predicted IMDb score.
  - **Genres and Maturity Rating**: Certain genres (e.g., drama, animation, action) and an R-rating were associated with higher IMDb scores.
- **Predicted Scores**: Using the model, IMDb scores for the 12 upcoming blockbusters were predicted:
  - *The Holdovers*: Predicted score of 7.4, the highest among the movies, having a "Good Director" and "Good Cinematographer."

  - *The Marvels*: Predicted score of 4.7, the lowest, reflecting a lack of "Good Director" or "Good Cinematographer."

Movie Title	**Predicted IMDb Score**
The Holdovers	7.38
Napoleon	6.48
The Hunger Games: The Ballad of Songbirds	6.43
Dream Scenario	5.47
Next Goal Wins	5.64
The Marvels	4.67

Error Analysis: Some predictions were lower than anticipated. For example, *The Marvels* and *Wish* were predicted to score low due to missing key features like a well-rated director or cinematographer, which negatively affected the predicted scores.