*Box Office Revenue Prediction*

*[Torben Lund, Trym Gallefoss, Oskar Windelstad]*

Introduction

The global film industry represents a significant economic powerhouse, with box office revenues serving as a key performance metric. Predicting these revenues helps stakeholders make informed decisions about marketing, distribution, and strategic planning. This project aims to develop a predictive model for estimating worldwide box office revenues using historical data from The Movie Database (TMDB).

DESCRIBE THE PROBLEM

The film industry operates on slim margins where success or failure at the box office can significantly impact financial outcomes. Identifying potential revenue generators or duds before their release can optimize resource allocation and maximize profits.

SCOPE

Business Objective:

Our goal is to create a predictive model that can estimate a movie's box office revenue before its release, providing a strategic tool for production companies, distributors, and investors.

Usage of the Solution:

The model will assist in forecasting revenues, guiding promotional strategies, and aiding in decision-making regarding movie releases. Currently, such predictions are based on expert opinions and historical trends. Our solution aims to add a data-driven dimension to these predictions.

Performance Measurement:

The model's performance will be evaluated using the RMSLE metric, ensuring that large revenues do not disproportionately affect the model's accuracy.

System Integration:

This model could be integrated into a larger decision-support system, interfacing with marketing and financial planning tools. Changes in predictive features, such as budget adjustments, could directly feed into the model to reassess revenue projections.

Stakeholders:

Key stakeholders include production companies, marketing agencies, distribution networks, and financial investors within the film industry.

Timeline:

* Preparation (days 1 – 2)

The project started with some general preparations and goalsetting for what the project wanted to achieve. That included setting roles, setting up infrastructure like GitHub repository and Google Docs for this report.

* Data collection, cleaning, and exploration (days 3-5)

This part was the real start of the project. Here we collected the data from The Movie Database (TMDB). We cleaned up the data, like handling missing values. Then we explored the data to look for correlations, mostly by plotting different features.

* Model development (days 7-10)

Next, we started to look at different models suited for regression tasks.

* + Web-app(days 11-13)

Finally, we developed the Streamlit web-app

## *METRICS*

*Business Metric Performance*

*The minimal "business metric" performance required for this project to be deemed successful is the enhancement of profitability through improved forecasting accuracy. A specific benchmark could be set, such as improving the precision of revenue forecasts by at least [X]%, which is expected to lead to better budget allocation and increased revenue margins. The business objective is to utilize predictive insights to drive strategic decisions, reduce financial waste, and capitalize on market opportunities.*

*Machine Learning and Software Metrics*

*To measure the efficacy of our machine learning solution, we will rely on the following metrics:*

*Root Mean Squared Logarithmic Error (RMSLE): Given the continuous nature of revenue figures and the need to minimize the impact of large errors, RMSLE will serve as a primary performance metric. It will particularly help in cases where the overestimation of revenue is less tolerable than underestimation. This is also the metric used for the Kaggle leaderboard, so it is only natural to include it.*

*Mean Absolute Error (MAE): MAE will provide a straightforward average of error magnitudes, offering a clear view of prediction accuracy in the same units as the predicted variable, which in this case is the revenue.*

*Latency: In a live environment, the response time of our model is crucial. Low latency is a requirement for real-time applications where immediate decision-making is pivotal.*

*Throughput: This measures the number of units of work that can be handled per time unit by the system. For our purposes, it will be defined as the number of predictions made per second. A high throughput will be necessary to process large volumes of data in a scalable manner.*

*Connection to Business Objectives*

*These metrics are intrinsically tied to the business objectives:*

*RMSLE aligns with the goal of precise revenue prediction, as it emphasizes the proportionality of errors rather than their absolute value, which is more relevant for financial decisions.*

*MAE offers a direct indication of average prediction error, which is easily interpretable by business stakeholders. It reflects the average impact that prediction errors will have on revenue-related decisions.*

*Latency is directly linked to operational efficiency. The faster the model can provide predictions, the more swiftly the business can respond to emerging trends and make informed decisions.*

*Throughput ensures that the predictive model can scale with the business. As the volume of data and the demand for predictions grow, the system must maintain performance without bottlenecking the decision-making process.*

*Each of these technical metrics provides insight into various aspects of model performance that, when optimized, drive us closer to achieving our overarching business goal of improved financial outcomes through enhanced predictive accuracy.*

# *DATA*

## *DATA*

### *Data Acquisition and Description*

*Dataset Overview:*  
*The dataset consists of metadata for 7,398 movies sourced from The Movie Database (TMDB). Each movie is uniquely identified by an id. The available metadata encompasses various attributes, including cast and crew details, plot keywords, budget, posters, release dates, spoken languages, production companies, and countries.*

*Test Set Details:*  
*The objective is to predict the worldwide revenue for a subset of 4,398 movies included in the test file. The provided data serves as both the training set for model development and the test set for performance evaluation.*

### *Data Uniqueness and Variability*

*Multiple Versions and Remakes:*  
*The dataset includes instances where multiple versions or remakes of a movie exist, each represented as a unique data point. For example, 'The Karate Kid' from 1986 (id: 5266) and its 2010 remake (id: 1987) are distinct entries.*

*Title Duplication:*  
*Some movies may share titles but are otherwise unrelated, such as 'Frozen' (id: 5295), the well-known Disney movie from 2013, and an earlier film with the same title (id: 139) from 2010 about skiers stranded on a chairlift. These instances are treated as separate entities.*

### *Data Preparation*

*Cleaning and Preprocessing:*  
*The preprocessing phase will involve handling missing values, ensuring that remakes and movies with duplicate titles are appropriately differentiated, and addressing any inconsistencies in the metadata.*

*Feature Engineering:*  
*New features may be engineered from the existing metadata, such as encoding the presence of notable cast or crew members, deriving the genre mix, and quantifying the breadth of a movie's international release.*

*Representation for Modeling:*  
*Categorical data such as languages and production companies will be transformed into numerical formats suitable for machine learning algorithms, potentially using one-hot encoding or embedding layers if deep learning techniques are employed.*

### *Ethical Considerations*

*Bias and Fairness:*  
*We will conduct analyses to ensure that the model does not inherently favor certain types of movies over others, particularly regarding budget, language, or country of production. This vigilance is to uphold fairness and avoid perpetuating any biases in movie revenue predictions.*

### *Ongoing Data Strategy*

*Data Monitoring:*  
*We will establish mechanisms to monitor the model's performance over time, ensuring it remains accurate as market trends evolve. Regular evaluations and updates to the model may be required to maintain its predictive power.*

# *MODELING*

*Describe which machine learning models you will explore. Describe how you plan to estimate baseline performance and baseline behavior. Remember that your first models should be simple. Baseline performance can typically be estimated using simple models or even non-machine learning-based solutions. You can also search for results obtained by others on the same or a similar task. You can also estimate “human-level performance” if relevant. Describe how you plan to investigate prediction mistakes and “feature importance” and how this will be used to improve your results.*

Model Exploration

Linear Regression Model:  
As a starting point, we employed a Linear Regression Model due to its simplicity and interpretability. It serves as a baseline to gauge the predictive power of our features without any complex interactions. Its performance will set a benchmark for subsequent, more sophisticated models.

XGBoost:  
Given its strong performance in various Kaggle competitions and projects, we integrated an XGBoost model. This gradient boosting framework is known for handling sparse data and providing a balance between prediction accuracy and computational efficiency. It also has the capability to manage the non-linearity in the data through its ensemble tree-based approach.

Random Forest Regressor:  
After reviewing solutions from similar problems, we implemented a Random Forest Regressor. This ensemble learning method is effective for regression tasks and offers several advantages, including handling overfitting better than decision trees and providing an internal method for feature importance evaluation. It ultimately outperformed the other models in our tests.

# *DEPLOYMENT*

The model will be deployed using Streamlit. Streamlit provides an easy way to implement web-apps. The web-app is very simple, only taking three input arguments. Budget, popularity, and runtime. After the input is provided, the model calculates the expected revenue of film of that nature. The model offers low complexity and makes it very easy to use. The drawback may be lack of performance due to few training features.

# *REFERENCES*

*-ChatGPT:*

-Hands-On Machine Learning with Scikit-Learn and TensorFlow

-Other references include various YouTube videos, forums and solutions on the Kaggle competitions