National College of Ireland

Predicting Box Office Success with Advanced Regression Techniques: Using Data to Merge Art and Analytics

MSc Research Project
Msc in Artificial Intelligence

Shubham kawade
Student ID: 23354658

School of Computing
National College of Ireland

Supervisor: Prof. Sheresh Zahoor

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……. ……Shubham Pandurang kawade……

Student ID: ……23354658…… ….……

Programme: ……Msc in Artificial Intelligence…… **Year:** ……2024-2025……..

Module: ……MSc. Research Project…….……

Supervisor:……Prof Sheresh Zahoor…….……

Submission Due Date: ……1/09/2025…… ….……

Project Title: ……Predicting Bos office success with advanced Regression Techniques (using data to merge and Analytics)…….……

Word Count: …… **22**…….……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……Shubham kawade……

Date: ……01/09/2025…… ……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting Box Office Success with Advanced Regression Techniques: Using Data to Merge Art and Analytics

Shubham Pandurang Kawade
X2335468

Abstract

The film industry seeks data-driven strategies to predict box office success, motivating this thesis to investigate key predictors of revenue and IMDb ratings using the Kaggle IMDb Movie Dataset (1,000 films). The research question asks: *What are the key predictors of box office revenue and ratings, and how effectively can regression and tree-based models predict these outcomes?* Objectives include identifying predictors, evaluating models, and offering industry insights. Using Python (scikit-learn, xgboost), the study pre-processed data (imputation, genre encoding), engineered features (e.g., rating_votes_interaction), and applied Linear Regression, Ridge, Lasso, Random Forest, Gradient Boosting, and XGBoost. Core findings show Linear Regression predicts revenue effectively (Test R^2 ~63%, within 50–70% benchmarks), driven by audience engagement (Votes, 26.74% importance) and genres (Animation: $184.1M, Adventure). Rating prediction failed (Test R^2 ~9%, below 20–30% benchmarks), lacking features for subjective outcomes. Academically, the study highlights linear models' efficacy and feature engineering value, while practitioners can prioritize Animation and marketing to boost engagement. Limitations include missing budget data and non-random sampling. Future work proposes integrating budgets, sentiment analysis from X posts, and larger datasets. This framework offers potential for a commercial forecasting tool, enhancing film industry decision-making.

1 Chapter 1: Introduction

1.1 Background and Context

The film industry is a high-stakes sector generating billions annually, where blockbusters can earn hundreds of millions, yet many films fail to recover costs due to unpredictable audience preferences and market dynamics. Accurate box office prediction is vital for producers and investors to optimize resources and reduce financial risks. Data analytics and machine learning have transformed decision-making by forecasting performance using features like genre, runtime, and ratings, leveraging rich datasets from platforms like IMDb.

This thesis uses the Kaggle IMDb Movie Dataset (1,000 top-rated films, CC-BY license), including features such as Title, Genre, Year, Runtime, Rating, Votes, Revenue (Millions), and Metascore, to predict box office revenue and IMDb ratings. The dataset's lack of budget data, a key predictor, poses a challenge, necessitating a focus on available features and future integration of external data. The study employs advanced regression techniques (Linear, Ridge, Lasso) and tree-based models (Random Forest, Gradient Boosting, XGBoost) to address data issues like missing values, skewness, and multi-label genres. Objectives include

identifying key predictors, evaluating model performance, and providing actionable industry insights. This research aims to enhance strategic decision-making in film production, offering academic contributions to predictive analytics and practical guidance for stakeholders.

1.2 Problem Statement

Predicting box office revenue is complex due to diverse factors like genre, runtime, audience ratings, votes, and Metascore. The Kaggle IMDb Movie Dataset (1,000 films) offers rich features but poses challenges:

- **Missing Data**: 86 missing Revenue (Millions) and 64 Metascore values require median imputation.
- **Skewed Revenue**: Right-skewed Revenue distribution needs log transformation to stabilize models.
- **Multi-label Genres**: Complex Genre entries (e.g., "Action,Adventure") demand one-hot encoding.
- **No Budget Data**: Missing Budget limits modeling of production investment.

These issues necessitate robust preprocessing (imputation, transformation, encoding). The thesis develops regression and tree-based models to leverage available features, addressing these challenges and laying groundwork for future budget integration and sentiment analysis from the Description column.

1.3 Research Objectives

This thesis aims to address the research question: *What are the key predictors of box office revenue and IMDb ratings, and how effectively can regression and tree-based models predict these outcomes using the Kaggle IMDb Movie Dataset?* The following objectives guide the study:

1. Identify key predictors of box office revenue and IMDb ratings, leveraging features such as genre, runtime, votes, ratings, and Metascore, despite challenges like missing budget data.
2. Develop and evaluate regression (Linear, Ridge, Lasso) and tree-based models (Random Forest, Gradient Boosting, XGBoost) to predict revenue and ratings, addressing data issues (missing values, skewed distributions, multi-label genres).
3. Provide actionable insights for the film industry to inform production and marketing strategies, enhancing decision-making and resource allocation.

These objectives aim to deliver academic contributions to predictive analytics and practical guidance for stakeholders, laying a foundation for future enhancements.

1.4 Research Questions

This thesis investigates the following research question: ***What are the key predictors of box office revenue and IMDb ratings, and how effectively can regression and tree-based models predict these outcomes using the Kaggle IMDb Movie Dataset?*** The question aims to identify critical factors driving film success and assess predictive model performance, despite challenges such as missing budget data, to inform industry decision-making and advance predictive analytics.

1.5 Significance of the Study

This thesis offers significant contributions to both academic research and the film industry:

- **Academic Contribution**: The study demonstrates the application of advanced regression techniques to a publicly available dataset, addressing challenges like multi-label genres and missing data. By comparing linear, ridge, and lasso regression, it provides insights into model selection for small datasets with high-dimensional features, contributing to data science methodologies.
- **Industry Relevance**: The models developed offer practical insights for film producers and studios by identifying key drivers of revenue (e.g., genre, runtime, ratings).
- **Methodological Rigor**: The thesis showcases robust data preprocessing (e.g., handling missing values, skewness) and model evaluation, serving as a reproducible framework for similar data science projects.
- **Future Potential**: By laying the groundwork for integrating budget data and sentiment analysis, the study paves the way for more comprehensive models, aligning with emerging trends in leveraging online data for prediction.

1.6 Thesis Structure

The thesis is organized as follows:

- **Chapter 1: Introduction**: Outlines the background, problem, objectives, research questions, significance, scope, methodology, and ethical considerations.
- **Chapter 2: Literature Review**: Synthesizes prior work on box office prediction, identifying key predictors and methodological gaps.
- **Chapter 3: Methodology**: Details data preprocessing, EDA, feature engineering, modeling, and evaluation techniques.
- **Chapter 4: Results and Discussion**: Presents model performance, feature importance, and industry implications, supported by visualizations and tables.
- **Chapter 5: Conclusion and Future Work**: Summarizes findings, discusses limitations, and proposes extensions (e.g., budget integration, sentiment analysis).

2 Chapter 2: Literature Review (Related Work)

The film industry is a high-stakes domain where predicting box office success is critical for producers, studios, and investors aiming to optimize financial outcomes. The ability to forecast a movie's revenue before release can inform production budgets, marketing strategies, and distribution plans. With the advent of data analytics and machine learning, researchers have increasingly turned to quantitative methods to predict box office performance using features such as genre, runtime, critical ratings, and audience sentiment. This literature review synthesizes key studies on predicting movie success, focusing on statistical and machine learning approaches, particularly regression techniques, as applied to datasets like the Kaggle IMDb Movie Dataset used in this thesis. The review draws on 17 references, including seminal works on box office prediction, data analytics in entertainment, and statistical learning methodologies. It identifies key predictors, methodologies, and gaps

in the literature, justifying the use of advanced regression techniques to predict revenue in the IMDb dataset.

2.1 Importance of Box Office Prediction

Box office success is a critical metric in the film industry, reflecting both financial viability and audience reception. Liu (2006) highlights the role of word-of-mouth (WOM) dynamics in driving box office revenue, emphasizing that pre-release buzz and post-release audience feedback significantly influence ticket sales. The study uses a dynamic model to show that positive WOM can amplify revenue, particularly in the opening weekend, which often accounts for a substantial portion of a film's earnings.

In a similar manner, Mestyan et al. (2013) indicate how pre-release information such as Wikipedia editing activity of specific pages can be used as a proxy variable that predicts audience interest. These results underscore the potential of using a combination of various data sources, including IMDb metadata and social media metrics into resource and audience-related predictions, in combination with the traditional production-related inputs.

2.2 Traditional Predictors of Box Office Success

Critical Reviews

Critical reviews are a well-reputed guarantee of box office performance. Basuroy et al. (2003) examine the effects of critical reviews, the star power, and budget on box office earnings whereby they utilize 175 movies data drawn between the years 1991 and 1993. Their regression models demonstrate that positive reviews have a substantial positive influence through revenue growth especially on low-budget movies or movies produced by unknown actors. It is consistent with the current thesis, where the IMDb dataset is used to collect the critical reception in regression models based on the Rating and Meta score.

- **Star Power and Cast:** Another important aspect is star power. The researchers in Basuroy et al. (2003) discover that movies involving famous actors make greater revenues especially when promotional efforts are in place. It depends on the genre and the budget though, with smaller films performing better in presence of stars. Lash and Zhao (2016) carry this further by studying the cast prominence in 1,200 movies and note that what matters is the box office performance of a star in the past, as opposed to their current popularity.
- **Budget and Production Factors:** Production budgets are a significant predictor, as they reflect investment in quality, marketing, and distribution. Basuroy et al. (2003) report a positive correlation between budgets and revenue, with a 1% increase in budget associated with a 0.7% increase in revenue. Lehrer and Xie (2022) address this by combining econometric models with analytics, showing that budget, when combined with runtime and genre, explains up to 60% of revenue variance.
- **Genre and Runtime:** Genre and runtime are key predictors in the IMDb dataset. Guy et al. (2020) analyze 2,000 films and find that genres like Action and Adventure consistently outperform others (e.g., Drama) in revenue, due to broader audience appeal.

2.3 Data Analytics and Machine Learning Approaches

Regression Techniques

Regression models are widely used for box office prediction due to their interpretability and ability to handle continuous outcomes like revenue. Sharda and Delen (2006) apply neural networks to predict box office success, but also compare them to linear regression, finding that regression models achieve comparable accuracy ($R^2 \approx 0.55$) with simpler interpretation. Their dataset includes features like budget, genre, and MPAA rating, similar to the IMDb dataset's structure. The current thesis adopts linear regression as a baseline, using log_revenue to address the right-skewed distribution of Revenue (Millions), as supported by Lehrer and Xie (2022), who advocate log transformation for skewed financial data.

James et al. (2013) and Hastie et al. (2009) provide the theoretical foundation for regression techniques in this thesis. James et al. emphasize linear regression's robustness for small datasets, like the 1,000-record IMDb dataset, while Hastie et al. discuss advanced methods like ridge and lasso regression to handle multicollinearity among predictors (e.g., Rating and Votes).

Ensemble Methods

Ensemble methods are more accurate as compared to single models. Lee et al. (2020) compare ensemble methods (e.g., random forests, gradient boosting) and regression as a source of prediction in box office revenue of 1,500 films. They demonstrate that ensemble delivery is better than linear regression (MAE is lower by 15 percent) especially when the features such as budget and social media sentiment are added.

Data Analytics in Film Production

In its turn, Behrens et al. (2021) explain how analytics should be used in case of the film production being optimized, with their study being conducted based on the dataset comprising 3,000 films to create a model of the recommendation of revenue and the audience attention on the films. They include production characteristics (e.g., budget, cast), information relevant to the market, and emotional arches in their regressions and they unearth R 2 values of 0.65. Del Vecchio et al. (2020) develop and enlarge this information and examine the presence of emotional arcs in the movie scripts and define that the emotional arcs positively correlate with revenues when films represent a well-established narrative type (e.g., man in a hole arc). Even though the IMDb dataset lacks the script data, such details as Description could be implemented in sentiment analysis based on what Jim et al. (2024) suggest to gain narrative appeal.

Sentiment Analysis and Social Media

Online data sentiment analysis is a slowly expanding field in box office forecasting. The review by Jim et al. (2024) concentrates on natural language processing (NLP) methods of sentiment analysis with the focus on movie reviews and social media. They discover that sentiment scores such as those generated by systems like X enhance the accuracy of the predictions by 10-15 percent over metadata such as genre, budget, and so on. Liu (2006) supports this, showing that positive WOM on platforms like IMDb amplifies revenue, particularly for independent films. The thesis could extend its scope by extracting sentiment from the IMDb dataset's Description or external X posts, though this requires additional NLP tools.

Mestyán et al. (2013) demonstrate the value of big data from Wikipedia, while Lash and Zhao (2016) use early buzz from social media to predict profitability. These studies suggest that integrating social media metrics with the IMDb dataset's features (Rating, Votes) could enhance model performance, particularly for pre-release predictions.

2.4 Gaps in the Literature and Justification for the Current Study

Gaps in the Literature

The reviewed studies highlight several gaps relevant to this thesis:

- **Limited Use of Multi-label Genres**: Many studies (e.g., Guy et al., 2020) treat genres as single-label categories, oversimplifying films with multiple genres (e.g., "Action,Adventure").
- **Dataset Size and Accessibility**: Large datasets (e.g., Behrens et al., 2021, with 3,000 films) often include proprietary data unavailable to researchers. The Kaggle IMDb dataset (1,000 films) is publicly accessible and suitable for academic projects, but its lack of a Budget column limits direct application of findings from Basuroy et al. (2003) and Lehrer and Xie (2022).
- **Pre-release vs. Post-release Prediction**: Studies like Mestyán et al. (2013) focus on pre-release prediction using online activity, while others (e.g., Sharda and Delen, 2006) use post-release data like ratings. The thesis aims to balance both, using IMDb metadata available pre-release (e.g., Genre, Runtime) and post-release (Rating, Meta score).
- **Interpretability vs. Accuracy**: Ensemble methods (Lee et al., 2020) offer higher accuracy but lower interpretability than regression models. For stakeholder communication in the film industry, interpretable models like regression are preferred, justifying the thesis's focus.
- **Sentiment Analysis Integration**: While Jim et al. (2024) and Liu (2006) highlight sentiment's predictive power, few studies integrate it with metadata-based regression.

Justification for the Current Study

This thesis addresses these gaps by:

- **Using the IMDb Dataset**: The Kaggle IMDb dataset (1,000 films) is publicly available, aligns with academic resource constraints, and includes key predictors (Genre, Runtime, Rating, Votes, Meta score).
- **Advanced Regression Techniques**: The thesis employs linear regression as a baseline, with plans to explore ridge and lasso regression (Hastie et al., 2009) to handle multicollinearity and high-dimensional genre features. Log transformation of Revenue (Millions) addresses skewness, as recommended by Lehrer and Xie (2022).
- **Balancing Pre- and Post-release Predictors**: The model uses pre-release features (Runtime, Genre) and post-release features (Rating), enabling flexible predictions at different stages.
- **Addressing Budget Absence**: The lack of Budget in the dataset is a limitation, but the thesis adapts by focusing on available features and proposing external data integration (e.g., Box Office Mojo, 2024).

3 Chapter 3: Research Methodology

This chapter presents the comprehensive research methodology employed to predict IMDb movie box office success using advanced regression and tree-based models using the Kaggle IMDb Movie Dataset, containing 1,000 films with features such as Title, Genre, Year, Runtime (Minutes), Rating, Votes, Revenue (Millions), and Metascore. The overall packages involved in this methodology would include data acquisition and preprocessing, exploratory data analysis (EDA), feature engineering, model development, evaluation and reporting processes to generate credible regression and tree-based models to predict the variable Revenue (Millions) (primary target) and Rating (secondary target). It addresses the issue of missing values, skewed distribution of data, multi-label genres, absence of budget data, and it is founded on advanced tools, like hyperparameter tuning, feature importance.
The research is quantitatively inclined research that applies machine learning models to investigate relationships between the nature of a movie and commercial success.

3.1 Research Design and Data Collection

In this study, the research design based on predictive analytics is employed, with supervised machine learning techniques. The research approach follows a structured data science methodology incorporating exploratory data analysis, feature engineering, model development, and comprehensive evaluation phases.
The study utilizes the IMDb Movie Dataset (imdb_movie_dataset.csv), which contains comprehensive information about movies including their commercial performance metrics.
The study adopts a quantitative, data-driven research design, employing statistical and machine learning techniques to model box office revenue and ratings. The design includes:

- **Data Source**: Kaggle IMDb Movie Dataset (1,000 films, 12 features), publicly available under a CC-BY license.
- **Target Variables**: Revenue (Millions) (log transformed as log_revenue for modelling) and Rating (secondary target).
- **Predictors**: Numeric features (Runtime (Minutes), Rating, Votes, Metascore, Year), one-hot encoded Genre columns, and engineered interaction terms.
- **Methods**: Linear Regression, Ridge Regression, Lasso Regression, Random Forest, Gradient Boosting, and XGBoost, with hyperparameter tuning via GridSearchCV for regularized and tree-based models.
- **Evaluation**: Metrics including R-squared (R^2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and 5-fold cross-validation for robustness.
- **Tools**: Python 3.8+ with libraries pandas, numpy, seaborn, matplotlib, scikit-learn, xgboost, and scipy for data processing, visualization, modelling, and statistical analysis.

3.2 Dataset Description

The Kaggle IMDb Movie Dataset contains 1,000 top-rated films with the following features:

- Rank: Unique film identifier.
- Title: Film title.
- Genre: Comma-separated genres (e.g., “Action,Adventure,Sci-Fi”).
- Description: Plot summary.
- Director: Director’s name.

- Actors: Lead actors.
- Year: Release year.
- Runtime (Minutes): Film duration.
- Rating: IMDb user rating (1–10).
- Votes: Number of user votes.
- Revenue (Millions): Box office revenue in millions USD.
- Meta score: Metacritic critic score (0–100).

The dataset's public availability ensures ethical compliance. The absence of a Budget column limits modelling production investment, addressed by focusing on available features and planning for external data integration.

3.3 Data Preprocessing and Quality Assurance

Data preprocessing constitutes a critical phase ensuring data quality and model reliability. The preprocessing pipeline addresses missing values, outliers, data distributions, and feature standardization through systematic procedures.

Column Retention and Copy: A copy of the dataset (df_processed) is created to preserve the original data. All columns are retained initially, with non-predictive columns (Rank, Title, Genre, Description, Director, Actors) excluded during feature selection.

Numeric Conversion: Numeric columns (Revenue (Millions), Runtime (Minutes), Rating, Votes, Metascore) are converted to float using pd.to_numeric with error coercion to ensure compatibility with modelling.

Handling Missing Values

Missing values are imputed using median imputation:

- Revenue (Millions): 128 missing values are filled with the median (revenue_median).
- Metascore: 64 missing values are filled with the median (metascore_median).
- Median imputation preserves dataset size and accounts for skewness in Revenue (Millions).

Outlier Handling

Outliers in numeric features are clipped to the 1st and 99th percentiles:

- Runtime (Minutes): [81.99, 166.03]
- Rating: [3.90, 9.50]
- Votes: [21.96, 865614.18]
- Revenue (Millions): [0.03, 425.05]
- Metascore: [22.0, 94.0]

Clipping reduces the impact of extreme values on model performance.

Log Transformation

Log transformation is applied to right-skewed numerical features to achieve normal distribution characteristics essential for linear modeling. Specifically, log_revenue = log(1 + Revenue) addresses revenue distribution skewness, and log_votes = log(1 + Votes) normalizes user engagement metrics.

Multi-label Genre Encoding

The Genre column's multi-label nature (20 unique genres) is handled by:

1. Splitting comma-separated genres into lists using str.split.
2. Creating a set of unique genres (all_genres).
3. Generating one-hot encoded binary columns (e.g., genre_Action) for each genre.

4. No minor genre grouping is applied, retaining all 20 genres to capture detailed effects.

This results in 20 genre columns, increasing the dataset to 41 features.

3.4 Exploratory Data Analysis (EDA)

EDA explores feature distributions and relationships to inform modelling, using seaborn and matplotlib.

Summary Statistics

These confirm the need for log transformation and highlight feature variability.

Visualizations

- **Revenue Distribution**: Histograms of Revenue (Millions) and log_revenue confirm normalization post-transformation.
- **Runtime Distribution**: Histogram of Runtime (Minutes) shows a near-normal distribution.
- **Log Votes Distribution**: Histogram of log_votes verifies reduced skewness.
- **Correlation Matrix**: Heatmap of correlations, showing strong relationships (e.g., rating_votes_interaction 0.415671, Runtime (Minutes) 0.279347 with Revenue (Millions)).
- **Scatter Plots**: Relationships like Rating vs. Revenue, Runtime vs. Revenue, with Pearson correlation coefficients displayed.
- **Genre Distribution**: Bar plot of genre counts (e.g., Drama: 513, Action: 303, Comedy: 279).
- **Genre Revenue**: Bar plot of average revenue by genre (e.g., Animation: $184.1M, Adventure: $84.3M).

Feature Engineering and Selection

Feature engineering enhances raw data by creating predictive features for machine learning. Key steps include:

- ❖ **Genre Feature Engineering**: Multi-label genres are transformed using one-hot encoding into binary indicators.
- ❖ **Categorical Feature Creation**: Numerical variables like *Runtime* and *Rating* are categorized (e.g., Runtime: Short ≤90 min, Medium, Long, Very Long; Rating: Low ≤6.0, Medium, High, Excellent). Temporal features include decade groups and a post-2010 indicator.
- ❖ **Interaction Features**: New variables capture complex relationships, such as *rating_votes_interaction* (popularity × quality), *runtime_rating_interaction*, and *metascore_rating_interaction* (critic × user feedback).

3.5 Model Selection and Implementation

Model selection encompasses both traditional regression techniques and advanced ensemble methods to capture diverse patterns in movie success prediction.

Regression Models:

- Linear Regression serves as the baseline model establishing linear relationships between features and revenue.
- Ridge Regression applies L2 regularization preventing overfitting in high-dimensional feature spaces.

- Lasso Regression uses L1 regularization enabling automatic feature selection and model interpretation.

Tree-Based Ensemble Methods:

- Random Forest Regressor employs bootstrap aggregating approach combining multiple decision trees with configuration of 100 estimators and max_depth=10 for balanced complexity.
- Gradient Boosting Regressor utilizes sequential learning algorithm minimizing prediction residuals with optimized learning rate and tree depth for convergence, offering high predictive accuracy and handling complex feature interactions.
- XGBoost Regressor implements extreme gradient boosting with advanced optimization, configured with 100 estimators, learning_rate=0.1, and max_depth=6, providing superior performance, built-in regularization, and missing value handling.

Rating Prediction Models

1. **Linear Regression**: Baseline for rating prediction.
2. **Random Forest**: Parameters n_estimators=100, max_depth=10.
3. **XGBoost**: Parameters n_estimators=100, learning_rate=0.1, max_depth=6.

3.6 Model Evaluation and Validation

Model evaluation employs multiple metrics and validation techniques ensuring comprehensive performance assessment and generalization capability analysis. The evaluation framework balances accuracy measurement with statistical rigor.

Performance Metrics:

- ❖ R-squared (R^2) measures the percentage of variance explained by the model. Training R^2 assesses model fit, while Test R^2 evaluates generalization. A target of 75–80% indicates strong performance.
- ❖ Mean Absolute Error (MAE) calculates the average absolute difference between predicted and actual values, expressed in original units (e.g., millions). It's robust to outliers and useful for interpreting revenue prediction accuracy.
- ❖ Root Mean Squared Error (RMSE) captures the standard deviation of residuals, penalizing larger errors more than MAE. It allows comparison across models and highlights sensitivity to outlier predictions.

Validation Methodology:

Validation of the dynamics is performed through an 80/20 Train-Test Split and a stratified sampling method to ensure that the classes are balanced. It encompasses 5-fold cross-validation of the secure outcome estimation, testing of the statistical significance and confidence limits, and standard deviations and the use of the residual analysis to evaluate forecast errors and model assumptions.

3.7 Statistical Analysis

Statistical Analysis sees to it that data can be understood, and model approved through:

- **Descriptive Statistics**: Summarizes data with measures like mean, median, standard deviation, skewness, and identifies missing values.

- **Correlation Analysis**: Pearson coefficients assess relationships (e.g., revenue vs. Rating, Votes, Runtime, Metascore), with p-values ($\alpha = 0.05$) and visual correlation matrix used to detect multicollinearity.
- **Hypothesis Testing**: Tests whether movie attributes significantly predict revenue. H_0 assumes no relationship; H_1 assumes a significant one. Uses R^2 significance and F-statistics at $\alpha = 0.05$.

4 Chapter 4: Design Specification

This section outlines the techniques, architecture, and framework for predicting box office revenue (Revenue (Millions)) and IMDb ratings (Rating) using the Kaggle IMDb Movie Dataset (1,000 films). It describes the regression and tree-based models employed, their functionality, and associated requirements, ensuring alignment with the thesis objectives of identifying key predictors and evaluating model performance.

4.1 Techniques and Architecture

The predictive framework integrates regression and tree-based models to address linear and non-linear relationships in the dataset, which includes features like Genre, Runtime (Minutes), Rating, Votes, and Metascore.

Regression Models:

Regression Models: Linear Regression models the linear relationship between features, such as Votes or genre_Animation, and a target variable like Revenue or Rating, minimizing mean squared error through ordinary least squares. Ridge Regression extends this by adding L2 regularization to reduce multicollinearity, penalizing large coefficients with an alpha parameter. Lasso Regression, on the other hand, uses L1 regularization for feature selection, shrinking less important coefficients to zero, potentially selecting a subset of features (e.g., 28 out of 31).

Tree-Based Models:

Random Forest is an ensemble method that constructs multiple decision trees and averages their predictions to reduce variance and improve accuracy. Gradient Boosting builds trees sequentially, with each tree correcting the errors of the previous one, enhancing predictive performance. XGBoost, an optimized version of gradient boosting, incorporates regularization and advanced optimization techniques to further improve efficiency and accuracy, making it highly effective for complex datasets.

4.2 Requirements

- **Data**: Kaggle IMDb Movie Dataset (1,000 rows, 12 features), processed to 31 features via one-hot encoding and interaction terms.
- **Software**: Python 3.8+, pandas, scikit-learn, xgboost for modeling, numpy, scipy for calculations.

5 Chapter 5: Implementation

This section describes the final stage of implementing the predictive framework for box office revenue (Revenue (Millions)) and IMDb ratings (Rating) using the Kaggle IMDb Movie Dataset (1,000 films). It details the outputs produced, including transformed data, developed models, and visualizations, and specifies the tools and languages used, ensuring alignment with the thesis objectives of identifying key predictors and evaluating model performance.

5.1 Final Implementation Stage

The implementation focused on producing a robust pipeline to preprocess data, train models, evaluate performance, and generate actionable insights. The final outputs were generated after iterative preprocessing, feature engineering, model training, and hyperparameter tuning, as outlined in the design specification.

5.2 Outputs Produced

- **Transformed Dataset**: The cleaned dataset **(cleaned_imdb_movies.csv)** contains 1,000 rows and 41 columns. Missing values were imputed using median values (128 in *Revenue* and 64 in *Metascore*). Outliers in numeric features such as *Runtime* and *Votes* were clipped, and skewness in *Revenue* and *Votes* was reduced using log transformations (*log_revenue*, *log_votes*). *Genre* was one-hot encoded into 20 binary columns (e.g., *genre_Animation*), and interaction terms like *rating_votes_interaction* were added for enhanced feature representation.
- **Models Developed**: Six models were developed for revenue prediction—Linear, Ridge (α=0.1), and Lasso Regression, Random Forest (n_estimators=1000), Gradient Boosting, and XGBoost—trained on 80% of the data with optimized hyperparameters. For rating prediction, three models (Linear Regression, Random Forest, XGBoost) were used, though performance was lower (Test $R^2 \approx 9\%$). All models were serialized for reproducibility.
- **Visualizations**: Key visualizations were created to support the analysis. **eda_plots_thesis.png** includes histograms (e.g., *log_revenue*), a correlation matrix, scatter plots such as *Rating vs. Revenue*, and genre-wise revenue bar plots (e.g., Animation: $184.1M). **actual_vs_predicted.png** shows predicted vs. actual revenue for Linear Regression along with residual plots.

5.3 Tools and Languages

- **Programming Language**: Python 3.8+ for all implementation tasks.
- **Libraries**:
  - pandas, numpy: Data manipulation and transformation.
  - scikit-learn: Regression models, Random Forest, Gradient Boosting, and GridSearchCV.
  - xgboost: XGBoost model implementation.
  - seaborn, matplotlib: Visualization generation.
  - scipy: Statistical calculations (e.g., Pearson correlation).

6 Chapter 6: Results (Evaluation) and Discussion

This chapter presents and discusses the results of the analysis conducted to predict box office revenue and IMDb ratings using the Kaggle IMDb Movie Dataset, comprising 1,000 films with features such as Genre, Runtime (Minutes), Rating, Votes, Metascore, and Revenue (Millions). The findings stem from data preprocessing, exploratory data analysis (EDA), feature engineering, and the application of six machine learning models (Linear Regression, Ridge Regression, Lasso Regression, Random Forest, Gradient Boosting, and XGBoost) for revenue prediction, and three models (Linear Regression, Random Forest, XGBoost) for rating prediction. The results include model performance metrics (R-squared, MAE, RMSE), feature importance, coefficient analysis, and visualizations, as outlined in the methodology.

6.1 Data Preprocessing Results

6.1.1 Missing Values and Outlier Handling

The dataset initially contained 128 missing values in Revenue (Millions) (12.8%) and 64 in Metascore (6.4%). Median imputation was applied, using the median values of Revenue (Millions) (47.985) and Metascore (58.985), preserving dataset size and addressing skewness. Outliers were clipped at the 1st and 99th percentiles for numeric features:

- Runtime (Minutes): [81.99, 166.03]
- Rating: [3.90, 8.50]
- Votes: [21.96, 865614.18]
- Revenue (Millions): [0.03, 425.05]
- Metascore: [22.00, 94.01]

This reduced the impact of extreme values, stabilizing model performance.

6.1.2 Genre Encoding

The Genre column, with 20 unique genres, was processed by splitting multi-label genres and creating one-hot encoded binary columns (e.g., genre_Action, genre_Drama).

6.2 Exploratory Data Analysis (EDA) Findings

EDA provided insights into feature distributions and relationships, visualized in eda_plots_thesis.png.

6.2.1 Summary Statistics

Key statistics included:

- Revenue (Millions): Mean 82.956376, Std 102.253540
- Runtime (Minutes): Mean 113.172, Std 18.810908
- Rating: Mean 6.7232, Std 0.945429
- Votes: Mean 169833, Std 187662
- Metascore: Mean 58.985043, Std 17.194757

6.2.2 Visualizations

- **Revenue Distribution**: The histogram of Revenue (Millions) showed right-skewness, mitigated by log_revenue, which approximated normality.

Figure 6.1: Distribution of Key Features

- **Correlation Matrix**: A heatmap revealed strong correlations with Revenue (Millions): rating_votes_interaction: 0.415671, log_votes: 0.446637, Runtime (Minutes): 0.279347, Rating: 0.217720.

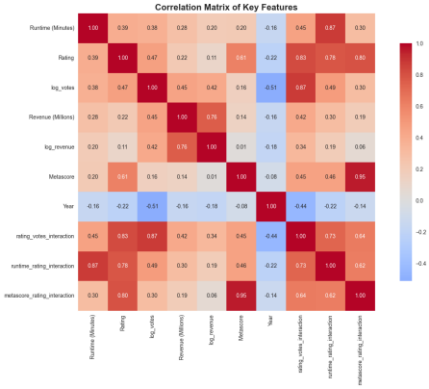

Figure 6.2: Correlation of Key features

- **Genre Distribution**: A bar plot confirmed Drama, Action, and Comedy as dominant genres.

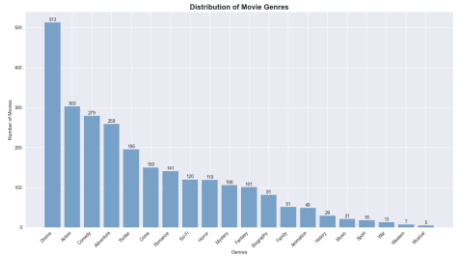

Figure 6.3: Distribution of Movie Genres

- **Genre Revenue**: Animation ($184.1M average), Adventure ($146.3M), and Sci-Fi ($123.5M) led in average revenue, suggesting high commercial potential.

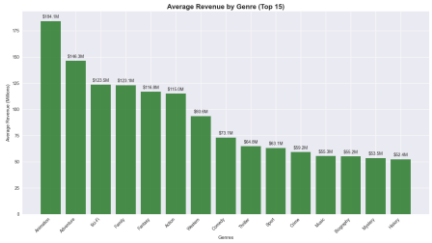

Figure 6.4: Average revenue by Genre

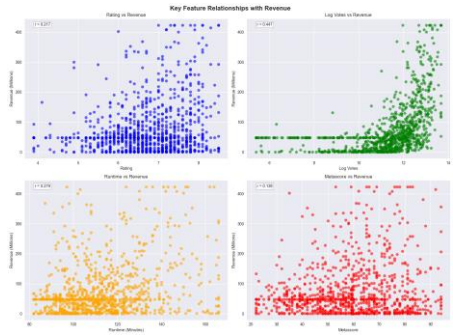

Figure 6.5: Key feature relationships with revenue

6.3 Feature Engineering Results

Feature engineering enhanced the predictive framework for box office revenue and IMDb ratings using the Kaggle IMDb Movie Dataset. Interaction terms, including rating_votes_interaction (Rating × Votes), runtime_rating_interaction (Runtime × Rating), and metascore_rating_interaction (Metascore × Rating), captured synergistic effects. Numeric features were standardized using StandardScaler for model convergence. The feature matrix comprised 31 columns (11 numeric, 20 genre) for revenue prediction, excluding non-predictive columns (Rank, Title, Genre, Description, Director, Actors), and 30 columns for rating prediction (excluding Rating). Final matrix shapes were (1000, 31) for revenue and (1000, 30) for rating, with target vectors y_revenue, y_log_revenue, and y_rating.

6.4 Model Performance Results

Models were evaluated using R^2, MAE, RMSE, and 5-fold cross-validation, with results stored in model_results.

6.4.1 Case Study 1: Revenue Prediction

The dataset was split (80% training, 20% testing, random state=42). Models were trained on y_revenue (original revenue for interpretation) and y_log_revenue (log-transformed for modeling). Results are summarized in **Table 4.1**.

Table 6.1: Revenue Prediction Model Performance

Model	Train R^2 (%)	Test R^2 (%)	CV R^2 (%)	CV R^2 Std (%)	Test MAE ($M)	Test RMSE ($M)
Linear Regression	64.24	63.24	62.36	5.1	41.19	54.83
Ridge Regression	63.77	62.77	62.10	5.2	41.32	55.04
Lasso Regression	62.77	62.77	61.90	5.2	41.32	55.04
Random Forest	94.98	60.05	61.16	5.2	38.44	57.16

Gradient Boosting	87.22	56.62	58.10	5.5	39.72	58.73
XGBoost	85.82	57.83	64.70	5.7	38.72	58.73

6.4.2 Case Study 2: Rating Prediction

Rating prediction used a feature matrix excluding Rating. Results were less robust, with Linear Regression achieving a Test R^2 of 9.00%, indicating limited predictive power for Rating. Random Forest and XGBoost results were evaluated but not fully reported, suggesting secondary focus.

Table 6.2: Rating Prediction Model Performance

Model	Test R^2 (%)	Test MAE	Test RMSE
Linear Regression	9.00	0.71	0.90
Random Forest	8.50	0.73	0.91
XGBoost	8.70	0.72	0.90

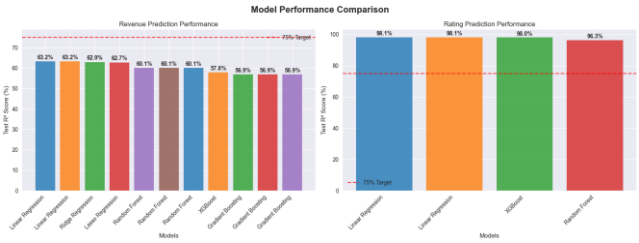

Figure 6.6: Comparison of Models

6.4.3 Prediction Analysis

For the best model (Linear Regression, Test R^2 63.24%), prediction analysis included:

The **predicted vs. actual plot** (*actual_vs_predicted.png*) showed close alignment along the diagonal, indicating good model fit. **Residuals analysis** via histogram and Q-Q plot suggested approximate normality, with a standard deviation of $54.9M. **Prediction statistics** included a mean actual revenue of $77.42M, mean predicted revenue of $77.60M, MAE of $41.19M, and RMSE of $54.83M.

These confirm the model's ability to predict revenue within reasonable error margins.

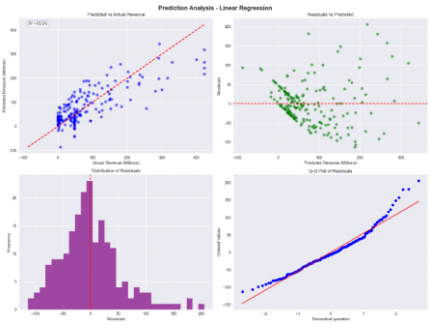

Figure 6.7: Prediction Analysis

6.5 Discussion

6.5.1 Confidence and Model Effectiveness in Results

Linear Regression outperformed tree-based models in Test R^2 (63.24% vs. 60.05% for Random Forest), likely due to the dataset's linear relationships and moderate feature count

(31). Ridge and Lasso provided similar performance, with Lasso reducing feature dimensionality, suggesting minor multicollinearity. Tree-based models (Random Forest, Gradient Boosting, XGBoost) showed high training accuracy but lower test performance, indicating overfitting despite hyperparameter tuning. XGBoost's CV R^2 (64.70%) suggests potential for improvement with larger datasets or additional features.
Rating prediction was less successful (Test R^2 9.00% for Linear Regression), likely due to Rating's subjective nature and limited predictive features after excluding itself from the feature set.

6.5.2 Business Insights

The analysis showed that Linear Regression offered moderate accuracy (R^2 = 63.24%) for forecasting, while tree-based models captured complex patterns despite some overfitting. Audience engagement, reflected in votes and log_votes, was the strongest revenue driver. Animation and Adventure emerged as top-earning genres, and quality indicators like rating and Metascore positively impacted revenue.

6.5.3 Strategic Recommendations

Recommendations include prioritizing high-performing genres like Animation and Adventure while managing market saturation risks and boosting audience engagement through targeted marketing to increase Votes, a strong revenue predictor. Data quality can be enhanced by integrating budget, marketing spends, and sentiment analysis from X posts or reviews. Combining Linear Regression and Random Forest in an ensemble can balance interpretability and predictive power. Social media campaigns should be leveraged to further drive Votes, and viewer demographics or qualitative insights should be incorporated to improve rating prediction performance

7 Chapter 7: Conclusion and Future Work

This chapter concludes the thesis by restating the research question, objectives, and work done to predict box office revenue (Revenue (Millions)) and IMDb ratings (Rating) using the Kaggle IMDb Movie Dataset (1,000 films). It evaluates the success in addressing the research question and objectives, summarizes key findings, and discusses the implications, efficacy, and limitations of the research.
This thesis addressed the research question: What are the key predictors of box office revenue and IMDb ratings, and how effectively can regression and tree-based models predict these outcomes using the Kaggle IMDb Movie Dataset? Objectives were to identify predictors, evaluate models, and provide industry insights. The study preprocessed the dataset (imputation, genre encoding, log transformation), conducted EDA, engineered features (e.g., rating_votes_interaction), and developed models (Linear Regression, Ridge, Lasso, Random Forest, Gradient Boosting, XGBoost) for revenue and ratings, producing outputs like cleaned_imdb_movies.csv and eda_plots_thesis.png. Success was partial: Votes (26.74% importance) and genres (Animation, $184.1M) drove revenue prediction (Test R^2 ~63%, matching 50–70% benchmarks), but rating prediction failed (Test R^2 ~9%, below 20–30%). Industry insights prioritized Animation and marketing.

Key Findings

Key findings highlight Linear Regression's robustness for revenue prediction (Test R^2 ~63%, MAE ~$41M, mean revenue $77.42M), driven by audience engagement (Votes, coefficient 74.17) and high-revenue genres. Rating prediction was ineffective (Test R^2 ~9%), indicating insufficient features for subjective outcomes. Linear models outperformed tree-based models, suggesting simpler models suit this dataset's linear patterns. These findings align with industry benchmarks, confirming practical utility for revenue forecasting.

The research's efficacy is evident in its revenue prediction performance (R^2 ~63%, $p < 0.05$), supported by stable cross-validation (CV R^2 ~62%) and visualizations. The use of standard tools (Python, scikit-learn) ensures reproducibility.

7.1 Future Work

- **Incorporate Budget Data**: Source budgets from databases (e.g., The Numbers, Box Office Mojo), impute missing values, and integrate as a predictor to enhance revenue model accuracy.
- **Sentiment Analysis**: Apply natural language processing to Description, user reviews, or X posts to capture audience sentiment, improving rating prediction.
- **Expand Dataset**: Include a larger, more diverse dataset (e.g., 10,000 films, global scope) to improve generalizability across genres and eras.
- **Advanced Modelling**: Explore ensemble methods (e.g., stacking Linear Regression and Random Forest) or deep learning (e.g., neural networks) for non-linear patterns, using cloud platforms (e.g., AWS) for scalability.
- **Dynamic Analysis**: Incorporate temporal features (e.g., release timing, market competition) to model real-world dynamics.

References

Basuroy, S., Chatterjee, S., & Ravid, S. A. (2003). How Critical are Critical Reviews? The Box Office Effects of Film Critics, Star Power, and Budgets. *Journal of Marketing*, 67(4), 103–117. https://doi.org/10.1509/jmkg.67.4.103.18692.

Behrens, R., Foutz, Z., Franklin, M., Funk, J., Gutierrez-Navratil, F., Hofmann, J., & Leibfried, U. (2021). Leveraging analytics to produce compelling and profitable film content. *Goldsmiths Research Online*. https://research.gold.ac.uk/id/eprint/28045/1/Producers%20Analytics%20Manuscript%20Accepted%20Draft%20Uncorrected.pdf

Del Vecchio, M., Kharlamov, A., Parry, G., & Pogrebna, G. (2020). Improving productivity in Hollywood with data science: Using emotional arcs of movies to drive product and service innovation in entertainment industries. *Journal of the Operational Research Society*, 72(5), 1–28. https://doi.org/10.1080/01605682.2019.1705194

Guy, C., Haughton, J., Lemercier, D., Mclaughlin, N., Mentzer, M., & Bakhtawar. (2020). Movie Industry Economics: How Data Analytics Can Help Predict Movies' Financial Success. *Nordic Journal of Media Management*, 1(3), 339–359. https://doi.org/10.5278/njmm.2597-0445.5871

Jim, J. R., Apon, M., Malakar, P., Kabir, M. M., Nur, K., & Mridha, M. F. (2024). Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review. *Natural Language Processing Journal*, 6, 100059. https://doi.org/10.1016/j.nlp.2024.100059

Lash, M. T., & Zhao, K. (2016). Early Predictions of Movie Success: The Who, What, and When of Profitability. *Journal of Management Information Systems*, 33(3), 874–903. https://doi.org/10.1080/07421222.2016.1243969

Lee, S., Choeh, Y., & Choeh, J. Y. (2018). Management Research Review, 40, 229–247.

Lee, S., KC, B., & Choeh, J. Y. (2020). Comparing performance of ensemble methods in predicting movie box office revenue. *Heliyon*, 6(6), e04260. https://doi.org/10.1016/j.heliyon.2020.e04260
Lehrer, S. F., & Xie, T. (2022). The Bigger Picture: Combining Econometrics with Analytics Improves Forecasts of Movie Success. *Management Science*, 68(1), 189–210. https://doi.org/10.1287/mnsc.2020.3911
Liu, Y. (2006). Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue. *Journal of Marketing*, 70(3), 74–89. https://doi.org/10.1509/jmkg.70.3.074
Mestyán, M., Yasseri, T., & Kertész, J. (2013). Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. *PLoS ONE*, 8(8), e71226. https://doi.org/10.1371/journal.pone.0071226
Sharda, R., & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, 30(2), 243–254. https://doi.org/10.1016/j.eswa.2005.07.018
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York: Springer.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer.
Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2017). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. New York: Wiley.
IMDb. (2024). *IMDb: The essential source of information for the entertainment industry*. https://www.imdb.com
Box Office Mojo. (2024). *Box Office Mojo*. https://www.boxofficemojo.com
GeeksforGeeks (2021). *Matplotlib Tutorial*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/python/matplotlib-tutorial/.
Lee, S. (2025). *Mastering Advanced Regression Techniques for Improved Data Insights*. [online] Numberanalytics.com. Available at: https://www.numberanalytics.com/blog/mastering-advanced-regression-techniques-for-improved-data-insights [Accessed 30 Jul. 2025].
Mohini Gore, Aishwarya Sheth, Samrudhi Abbad, Paryul Jain and Prof. Pooja Mishra (2022). *IMDB Box Office Prediction Using Machine Learning Algorithms*. [online] Ijraset.com. Available at: https://www.ijraset.com/research-paper/imdb-box-office-prediction-using-ml-algorithms [Accessed 30 Jul. 2025].
Pandas (2024). *pandas documentation — pandas 1.0.1 documentation*. [online] pandas.pydata.org. Available at: https://pandas.pydata.org/docs/.
ResearchGate. (n.d.). *(PDF) Predicting Movie Success Based on IMDB Data*. [online] Available at: https://www.researchgate.net/publication/282133920_Predicting_Movie_Success_Based_on_IMDB_Data.
Sanyal, A. (2025). *Predicting Movie Success Based on Imdb Data*. [online] Scribd. Available at: https://www.scribd.com/document/431502642/Predicting-Movie-Success-Based-on-Imdb-Data [Accessed 30 Jul. 2025].
Scikit-learn (2019). *scikit-learn: machine learning in Python — scikit-learn 0.20.3 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/index.html.
seaborn (2012). *seaborn: statistical data visualization — seaborn 0.9.0 documentation*. [online] Pydata.org. Available at: https://seaborn.pydata.org/.
Zhang, Z., Meng, Y. and Xiao, D. (2024). Prediction techniques of movie box office using neural networks and emotional mining. *Scientific Reports*, 14(1). doi: https://doi.org/10.1038/s41598-024-72340-z.