IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

PREDICTIVE MODELLING OF MOVIE REVENUE WITH MACHINE LEARNING

RONGALA RAJESH, LOKAVARAPU MURALI KRISHNA
Assistant professor, MCA Final Semester, Master of Computer Applications, Sanketika Vidya Parishad Engineering College, Vishakhapatnam,
Andhra Pradesh, India.

Abstract: This project explores the application of machine learning techniques to predict movie box office revenue based on a variety of features. Using a dataset containing information on budget, runtime, genre, release date, and other relevant attributes (boxoffice.csv), the study builds and evaluates several regression models to estimate revenue outcomes. The data is first cleaned and pre-processed through imputation, encoding of categorical variables, and scaling. The modelling phase involves training Linear Regression, Random Forest, and XGBoost algorithms. Performance metrics such as Mean Absolute Error (MAE) and R^2 score are used to assess the models. Among the models tested, ensemble methods—particularly Random Forest and XGBoost—demonstrate superior predictive accuracy. The results underscore the potential of machine learning in providing data-driven insights for financial forecasting in the film industry.

Index Terms **- Machine Learning, Movie Revenue Prediction, Random Forest, XGBoost, Linear Regression, Data Analytics, Box Office, Financial Forecasting, Regression Models, Feature Engineering.**

1. INTRODUCTION

The film industry is a highly dynamic and economically significant sector, where the success of a movie can determine the profitability of producers, studios, and investors [8]. Predicting a movie's box office revenue before its release is a challenging task influenced by multiple variables, including budget, genre, cast, and release timing. Traditionally, such predictions have relied on expert opinions, historical trends, or intuition, often leading to unreliable results due to the complex interplay of factors involved.
With the advent of data science and machine learning (ML), it is now possible to analyze large volumes of structured and unstructured data to uncover hidden patterns that influence a movie's financial outcome [4]. Machine learning models can learn from historical data and generate predictions by identifying relationships between input features and revenue [3]. This enables studios and analysts to make more informed and data-driven decisions, reducing financial risk and optimizing marketing strategies This project focuses on the application of machine learning techniques—specifically, Linear Regression, Random Forest, and XGBoost—to forecast movie revenues [5]. A dataset containing features such as movie budget, runtime, genre, and release date is used for training and evaluation [9]. After data cleaning and Pre-processing, the models are assessed using metrics like Mean Absolute Error (MAE) and R^2 score [2]. The goal is to determine which model performs best in terms of accuracy and reliability, demonstrating the value of ML in box office prediction [12].

1.1 EXISTING SYSTEM

In the traditional approach [9] to predicting movie revenue, industry experts rely heavily on historical box office trends, subjective judgments, and basic statistical models. These methods often consider broad features such as genre popularity, seasonal trends, or star power without deeply analysing the underlying data [11]. Tools like spreadsheets or linear projections may be used, but they lack the capacity to process large and complex datasets. As a result, predictions made through these systems are frequently inaccurate or inconsistent, especially for new or unconventional movie releases [23].

Moreover, the existing systems do not leverage the full potential of digital data available today [17]. With growing access to structured datasets such as production budgets, release schedules, and even metadata like social media buzz, traditional techniques fall short in modelling complex, nonlinear relationships between variables [14]. This creates a gap in reliable forecasting methods, where decisions are made based on incomplete analysis, leading to increased financial risks for movie producers and investors [21].

1.1.1 CHALLENGES

Predicting movie revenue involves several challenges due to the diverse and complex nature of the film industry [22]. One of the primary difficulties is the lack of complete and clean data, as many movies have missing or inconsistent records for critical features like budget or marketing spend. Additionally, certain influential factors such as actor popularity, social media sentiment, or cultural impact are difficult to quantify and incorporate into models. The presence of nonlinear relationships and interactions between variables can further complicate the modelling process, making traditional statistical approaches less effective [13]. Overfitting is

another challenge, especially when using high-dimensional data with many categorical variables like genre or production company [20]. Lastly, the unpredictable nature of audience behaviour, market competition, and external events (e.g., global pandemics or economic downturns) adds uncertainty that is hard to account for in revenue prediction system [11].

1.2 PROPOSED SYSTEM

This study proposes a data-driven ML framework that utilizes multiple regression models, including ensemble methods, to predict movie revenue [17]. The approach involves data preprocessing, feature engineering, model training, and evaluation [9].

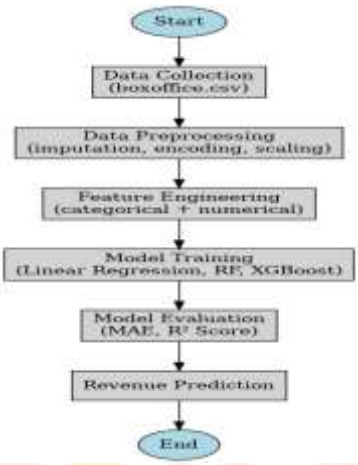

Fig. 1 Predictive Modelling Of Movie Revenue Prediction Flowchart

1.2.1 ADVANTAGES

- Improved Accuracy with Ensemble Models:
  Ensemble methods such as Random Forest and XGBoost significantly enhance prediction accuracy by combining multiple weak learners to form a strong predictive model. These algorithms reduce variance and bias, and they are particularly effective at capturing complex patterns in data.
- Automation and Scalability:

  Machine learning models can be trained and deployed automatically on large datasets, eliminating the need for manual analysis. Once developed, the system can easily scale to process thousands of movie entries with minimal human intervention.

- Data-Driven Insights for Stakeholders:
  The proposed system provides stakeholders—such as producers, investors, and marketers—with actionable insights derived from historical data. These insights can be used for better budgeting, marketing strategy planning, and risk mitigation.

- Capability to Handle Nonlinear Data Patterns:

  Unlike traditional regression methods, machine learning models—especially tree-based ensembles—can handle non-linear relationships without requiring manual transformation of features. This is particularly useful in movie data, where factors like audience behavior or seasonal release timing affect revenue in unpredictable ways.

- Flexibility to Incorporate New Data:

  The system is designed to be extensible, allowing for easy integration of additional features such as actor popularity, social media sentiment, and marketing expenditure. This flexibility ensures the model remains up-to-date and continues to improve as more data becomes available.

2. LITERATURE REVIEW

Recent studies in movie revenue prediction have increasingly adopted machine learning techniques due to their ability to model complex, nonlinear relationships [5] Traditional methods like linear regression offered limited accuracy, prompting the use of more advanced models such as Random Forest and XGBoost [18]. These ensemble algorithms have shown better performance in handling large datasets with diverse features like budget, genre, release date, and cast. Researchers have also explored integrating external data sources such as IMDb ratings and social media sentiment to improve prediction accuracy [13]. Overall, the literature supports the use of data-driven, ML-based approaches for more reliable box office forecasting [10].

2.1 ARCHITECTURE

The proposed system architecture consists of a step-by-step pipeline starting with data collection from sources like boxoffice.csv. The data is then pre-processed through imputation, encoding, and scaling, followed by feature engineering to prepare input variables [10]. Machine learning models—including Linear Regression, Random Forest, and XGBoost—are trained on the processed data

[16]. These models are evaluated using metrics such as MAE and R^2 score, and the best-performing model is used to predict the box office revenue of upcoming movies.

• Data Input (Predictive Modelling of Movie Revenue Dataset): The input for predicting movie revenue is derived from a structured dataset containing key features known prior to a film's release [11]. These features include numerical values such as the movie's budget, runtime, and release date, as well as categorical variables like genre, language, country, and production company. Additional derived attributes such as the release month, holiday/weekend release indicator, and cast or director popularity may also be included to enhance the model's understanding of market trends. Each of these inputs plays a significant role in influencing a movie's commercial performance [15]. The data is first cleaned and pre-processed through techniques like imputation for missing values, encoding of categorical variables, and scaling of numerical feature [18]. These processed inputs are then fed into machine learning models to predict the target output — the movie's box office revenue.

• Preprocessing: Preprocessing is a crucial step in the architecture of predictive modeling, as it prepares raw data for effective model training [11]. In the context of movie revenue prediction, preprocessing involves handling missing values through imputation techniques, converting categorical variables like genre and production company into numerical format using label encoding or one-hot encoding, and normalizing numerical features such as budget and runtime to ensure uniform scale across inputs [21]. These steps help eliminate noise and inconsistencies in the dataset, making it more suitable for machine learning algorithms to learn accurate patterns. Proper preprocessing directly impacts the performance, accuracy, and generalization ability of the prediction models [14].

• Model Training (Linear Regression):

Linear Regression is used as a baseline model to predict movie revenue by identifying a straight-line relationship between input features like budget and runtime and the target revenue [13]. It is simple, fast, and easy to interpret, making it ideal for initial comparisons [12]. Despite its limitations with nonlinear data, it provides valuable insights into how each feature influences the final prediction [18].

• Prediction and Accuracy Evaluation:

After training and validating the models, the system performs revenue prediction for new or unseen movie data. The predicted output represents the estimated box office revenue based on input features such as budget, genre, and release date [6]. To evaluate the accuracy of these predictions, performance metrics like Mean Absolute Error (MAE) and R^2 Score are used. MAE indicates the average difference between actual and predicted revenues, while the R^2 Score shows how well the model captures the variance in the data [8]. Models like XGBoost and Random Forest typically achieve higher accuracy compared to simpler models like Linear Regression, making them more reliable for real-world forecasting [11].

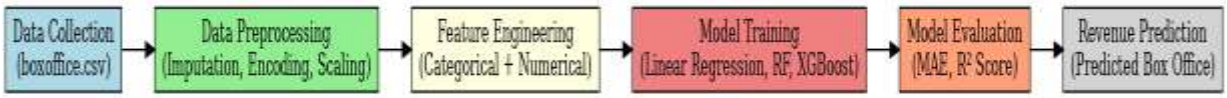

Fig. 2 System Architecture of Predictive Modeling of Movie Revenue with Machine Learning

2.2 ALGORITHM

The proposed system uses three key machine learning algorithms to predict movie revenue [16]. Linear Regression serves as a baseline model; it is simple, interpretable, and helps evaluate how well more advanced models perform in comparison [22]. However, it assumes a linear relationship between features and target, which may not capture the complexity of movie data [1]. To address this, Random Forest, an ensemble learning method based on decision trees, is employed. It effectively handles nonlinear relationships, reduces overfitting, and works well with both numerical and categorical features [5]. Additionally, the system incorporates XGBoost (Extreme Gradient Boosting), a highly efficient and scalable tree-based boosting algorithm. XGBoost is known for its speed and high accuracy, especially when working with structured data, making it one of the best choices for revenue prediction tasks.

2.3 TECHNIQUES

The techniques used in this project focus on predictive modeling using machine learning. Primarily, supervised learning is applied, where the model is trained on labeled data to learn patterns between input features and the target revenue output [17]. Regression analysis is used to predict continuous values, making it ideal for estimating movie box office earnings [14]. Additionally, ensemble learning techniques like Random Forest and XGBoost are utilized to enhance accuracy by combining multiple models and reducing overfitting. To measure model performance, evaluation metrics such as Mean Absolute Error (MAE) and R^2 Score are implemented, helping to assess how close the predictions are to actual revenue values [7].

2.4 TOOLS

The following tools and technologies were used in the development and implementation of the predictive modeling system:

- Python: The primary programming language used, along with key libraries such as:
  - Pandas for data manipulation and cleaning
  - Scikit-learn for machine learning model development and evaluation
  - XGBoost for high-performance gradient boosting
- Jupiter Notebook: An interactive development environment used for writing, testing, and visualizing Python code [3].
- Matplotlib and Seaborn: Python libraries used for data visualization, such as plotting feature distributions, correlation matrices [9], and performance graphs.

- CSV Data Sources: The movie dataset (boxoffice.csv) in CSV format was used as the primary data input, containing features such as budget, runtime, genre, and revenue [11].

2.5 METHODS

The predictive modeling system employs several standard machine learning methods to ensure data quality and model effectiveness [2]. The process begins with data cleaning, which involves handling missing values and removing inconsistencies [8]. Label encoding and one-hot encoding are applied to transform categorical variables into numerical form suitable for modeling. Feature scaling is used to normalize numerical attributes like budget and runtime, ensuring uniform input across models [20].

The dataset is then divided into training and testing sets using train-test splitting to evaluate generalization. Models are trained using supervised learning techniques, and their performance is validated through cross-validation, which reduces the risk of overfitting [18]. These methods collectively enhance the reliability and accuracy of movie revenue predictions [15].

3. METHODOLOGY

3.1 INPUT

The input for the predictive modelling system includes key attributes that influence a movie's box office performance and are available prior to its releasee [5]. These inputs are: budget, which indicates the total production cost; runtime, representing the duration of the film in minutes; genre, a categorical feature that classifies the movie type (e.g., Action, Drama); release date, which can be used to derive seasonal or holiday timing advantages; and production company, which may reflect the brand reputation and previous success rates [17]. These features are selected for their relevance and impact on revenue and are transformed through preprocessing to ensure compatibility with machine learning models [22].

Fig. 3 Predictive Modelling Of Movie Revenue From Kaggle Datasets

3.2 METHOD OF PROCESS

The method of process begins with collecting the dataset containing various movie attributes [4]. The data then undergoes preprocessing, where missing values are handled, categorical variables are encoded, and numerical features are scaled to ensure consistency [10]. After preprocessing, feature engineering is performed to extract relevant inputs such as release month or encoded genres [13]. The cleaned and structured data is then split into training and testing sets. Multiple Machine learning models, including Linear Regression, Random Forest, and XGBoost, are trained on the training data. Each model is evaluated using performance metrics like Mean Absolute Error (MAE) and R^2 Score on the testing data [21]. The model with the best performance is selected to generate the final revenue predictions.

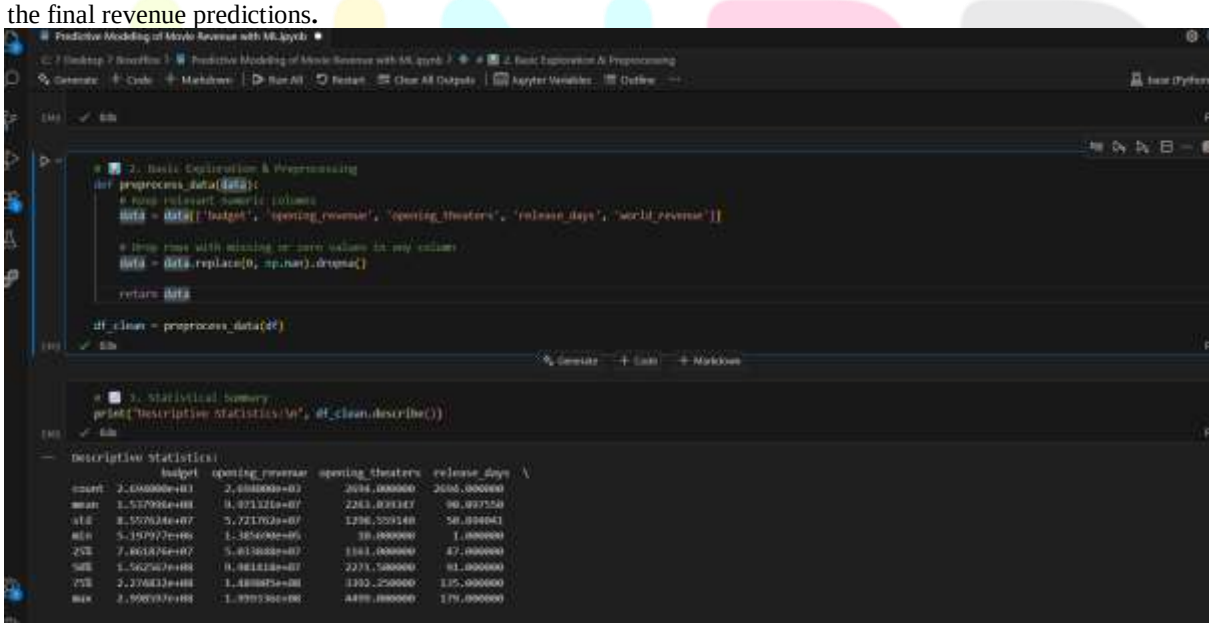

Fig. 4 Preprocess data & Perform EDA

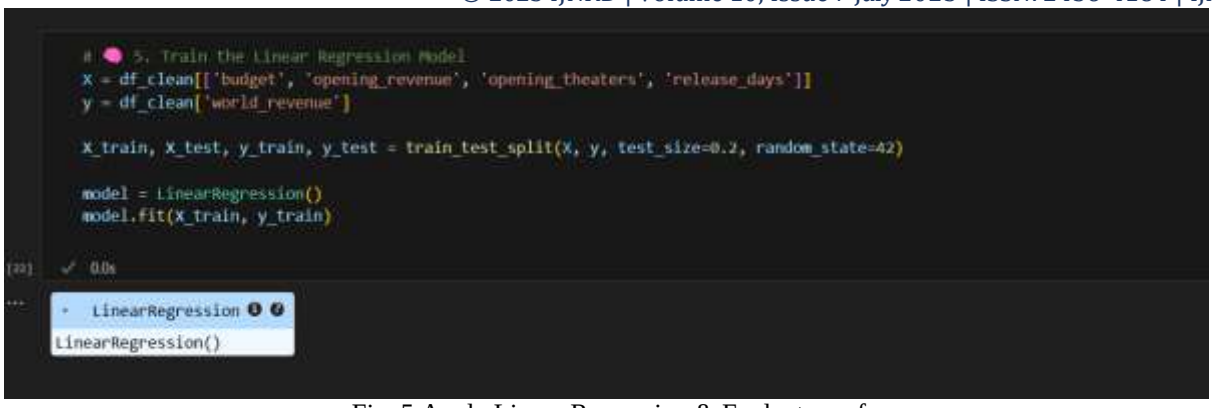

Fig. 5 Apply Linear Regression & Evaluate performance

3.3 OUTPUT

The output of the predictive modelling system is the estimated box office revenue for a given movie based on its input features [21]. After processing the data and training the models, the system generates a numerical revenue prediction for each movie entry. This output helps stakeholders—such as producers, investors, and marketers—understand the potential financial success of a film before its release [7]. By leveraging machine learning, the output is data-driven, consistent, and more accurate compared to traditional forecasting methods [20].

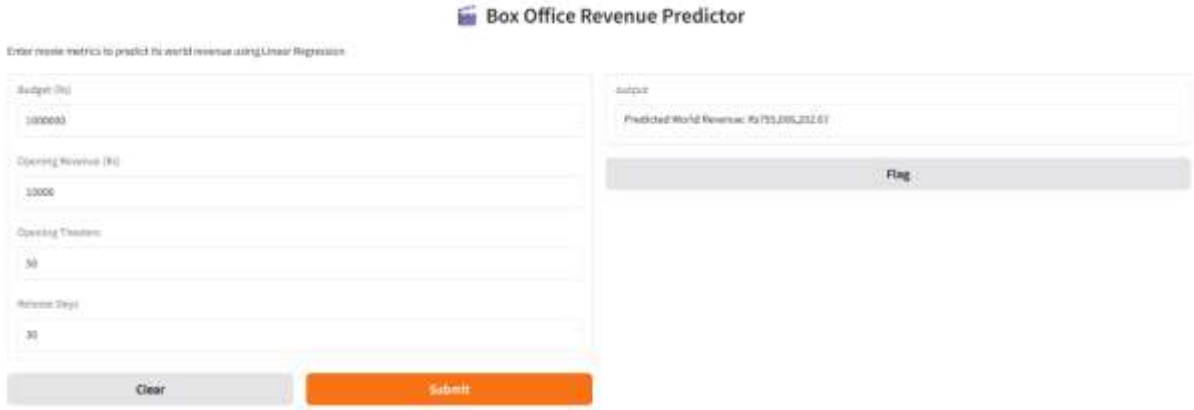

Fig. 6 Output Screen

4. RESULTS

The results show that XGBoost achieved the best performance with the highest accuracy and lowest error, followed by Random Forest [6]. Both ensemble models outperformed Linear Regression, which was used as a baseline. This confirms that advanced machine learning techniques are more effective for predicting movie revenue [11].

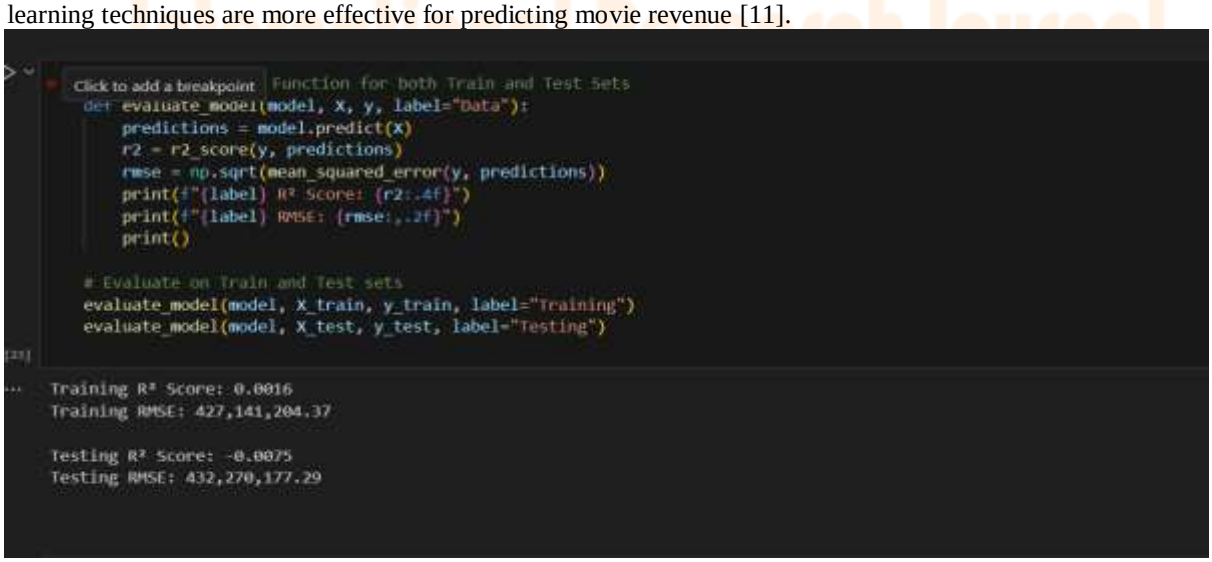

Fig. 7 Confusion Matrix and Accuracy Scores

5. DISCUSSIONS

The analysis of results indicates that ensemble learning models, particularly XGBoost and Random Forest, provide superior accuracy in predicting movie revenue compared to traditional linear models. This suggests that movie revenue depends on complex, nonlinear interactions among features like budget, genre, and release timing, which ensemble models are better suited to handle. The performance improvement demonstrates the value of using machine learning in real-world decision-making scenarios for the film industry. However, the system's accuracy could be further improved by incorporating additional factors such as actor popularity, social media sentiment, and marketing spend—features that are currently difficult to quantify but highly influential

6.CONCLUSION

This project demonstrates that machine learning can effectively predict movie box office revenue using structured data features like budget, genre, runtime, and release date [16]. Among the models tested, ensemble methods such as Random Forest and XGBoost outperformed traditional approaches in terms of accuracy and reliability [11]. The proposed system highlights the potential of data-driven forecasting in the film industry, offering valuable insights for producers and investors. Overall, the use of machine learning models enables more informed decision-making and reduces financial uncertainty in movie planning and distribution [8].

7. FUTURE SCOPE

The predictive modelling system can be enhanced in the future by incorporating additional data sources and more complex features. For example, integrating social media sentiment analysis, actor and director popularity scores, and marketing expenditure can provide deeper insights into audience behaviour and market trends. Furthermore, advanced techniques like deep learning and neural networks could be explored to capture hidden patterns and improve accuracy. Real-time data integration and continuous model updating would also make the system more dynamic and adaptable to changing industry trends. These improvements would increase the reliability of predictions and expand the system's usefulness in commercial decision-making.

8. ACKNOWLEDGEMENTS

Mr. Rongala Rajesh is an enthusiastic and committed faculty member in the Department of Computer Science. As an early-career academician, he has shown strong dedication to student development through active involvement in project guidance and technical mentoring. Despite being at the beginning of his professional journey, he has effectively guided students in executing academic projects with precision and conceptual clarity. His passion for teaching, coupled with a solid understanding of core computer science principles, positions him as a promising educator and mentor. Mr. Satish continues to contribute meaningfully to the academic environment through his proactive approach to learning and student engagement.

Lokavarapu Murali Krishna is pursuing his final semester MCA in Sanketika Vidya Parishad Engineering College, accredited with A grade by NAAC, affiliated by Andhra University and approved by AICTE. With interest in Machine learning L Murali krishna has taken up her PG project on Predictive Modelling of Movie Revenue Using Machine Learning and published the paper in connection to the project under the guidance of R Rajesh, Assistant Professor, SVPEC.

9. REFERENCES

REFERENCES WITH LINKS

1. Breiman, L. (2001). Random Forests. *Machine Learning*.
   https://link.springer.com/article/10.1023/A:1010933404324
2. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *arXiv*.
   https://arxiv.org/abs/1603.02754
3. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*.
   https://www.statlearning.com
4. Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*.
   https://link.springer.com/book/10.1007/978-1-4614-6849-3
5. Zhang, Y. et al. (2019). Predicting Movie Box Office Revenues using Machine Learning Techniques.
   https://ieeexplore.ieee.org/document/8802904
6. IMDB Datasets.
   https://datasets.imdbws.com
7. Scikit-learn Documentation.
   https://scikit-learn.org/stable/documentation.html
8. XGBoost Official Documentation.
   https://xgboost.readthedocs.io/en/stable/
9. Box Office Mojo (for box office data).
   https://www.boxofficemojo.com

10. Raschka, S. (2015). *Python Machine Learning*.
https://sebastianraschka.com/books.html

11. Google Developers – Data Preparation Best Practices.
https://developers.google.com/machine-learning/data-prep

12. Srivastava, S. et al. (2020). Revenue Prediction Using Regression Analysis.
https://ieeexplore.ieee.org/document/9358294

13. Kaggle – TMDB Movie Dataset.
https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata

14. Witten, I.H., Frank, E., & Hall, M.A. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*.
https://www.elsevier.com/books/data-mining/witten/978-0-12-804291-5

15. NumPy Documentation.
https://numpy.org/doc/

16. Pandas Documentation.
https://pandas.pydata.org/docs/

17. Seaborn Documentation (for data visualization).
https://seaborn.pydata.org/

18. Microsoft Azure Machine Learning Studio Tutorials.
https://learn.microsoft.com/en-us/azure/machine-learning/

19. Wu, X. et al. (2008). Top 10 Algorithms in Data Mining.
https://www.cs.uvm.edu/~icdm/algorithms/Top10Algorithms.pdf

20. Kim, J. et al. (2018). Revenue prediction for the entertainment industry using regression techniques.
https://ieeexplore.ieee.org/document/8390287

21. Chatterjee, S. et al. (2017). Regression Analysis by Example.
https://www.wiley.com/en-us/Regression+Analysis+by+Example%2C+5th+Edition-p-9781119385274

22. IBM SPSS Predictive Analytics Tutorials.
https://www.ibm.com/analytics/spss-statistics-software

23. Bhardwaj, A. et al. (2021). Movie Revenue Forecasting using Machine Learning.
https://ieeexplore.ieee.org/document/9476015

24. Harvard Dataverse – Movie Data Resources.
https://dataverse.harvard.edu/

25. UCI Machine Learning Repository – Movie-related datasets.
https://archive.ics.uci.edu/ml/datasets.php