PREDICTIVE ANALYTICS IN ENTERTAINMENT MEDIA: A COMPREHENSIVE MODEL FOR FORECASTING MEDIA SUCCESS AND APPLICATION FOR MOVIES AND SERIES

BY

ASIL AHSAN MUBIN
ID: 201-15-3555

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Narayan Ranjan Chakraborty
Associate Professor and Associate Head
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

July 2024

APPROVAL

This Project titled "**Predictive Analytics in Entertainment Media: A Comprehensive Model for Forecasting Media Success and Application for Movies and Series**", submitted by **Asil Ahsan Mubin** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **13.07.2024**.

BOARD OF EXAMINERS

Dr. Md. Ismail Jabiullah (MIJ) — **Board Chairman**
Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Mr. Abdus Sattar (AS) — **Internal Examiner 1**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Ms. Zahura Zaman (ZZ) — **Internal Examiner 2**
Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Ahmed Wasif Reza (DWR) — **External Examiner**
Professor
Department of Computer Science and Engineering
East West University

DECLARATION

I hereby declare that this project has been done by us under the supervision of **Name of the Supervisor, Designation, Department of Computer Science and Engineering, Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Narayan Ranjan Chakraborty
Associate Professor and Associate Head
Department of CSE
Daffodil International University

Submitted by:

Asil Ahsan Mubin
ID: 201-15-3555
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to almighty for His divine blessing making it possible for us to complete the final year project successfully.

I am grateful and wish my profound indebtedness to **Narayan Ranjan Chakraborty, Associate Professor and Associate Head**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine Learning*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

I would like to express my heartiest gratitude to the **Head of the Department of CSE**, for his kind help in finishing our project and also to other faculty members and the staff of the Department of CSE, Daffodil International University.

I would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, I must acknowledge with due respect the constant support and patience of my parents.

ABSTRACT

In the evolving landscape of entertainment media, accurately forecasting the success of movies and series presents a formidable challenge with significant implications for production and marketing strategies. This research project, introduces an innovative predictive model that transcends traditional approaches by integrating a myriad of factors influencing media success. Grounded in the principles of predictive analytics, the model employs advanced machine learning algorithms to analyze data collected from a meticulously designed survey, capturing audience preferences, genre trends, and the impact of social media buzz. The core of this research lies in its data-driven methodology, where both quantitative and qualitative data—ranging from budget and star power to narrative complexity and digital engagement metrics—are synthesized to predict the potential success of entertainment content. The model's capability to assimilate diverse data sets, including real-time social media sentiment analysis and traditional box office metrics, sets it apart, enabling predictions with SVM model to achieve 94% accuracy. This high degree of precision not only underscores the model's robustness but also its potential to revolutionize the way success is gauged in the entertainment industry. By providing producers, marketers, and content creators with actionable insights, the model facilitates strategic decision-making, potentially leading to more targeted and successful media productions. The implications of this study extend beyond immediate industry applications, paving the way for future innovations in predictive modeling and offering a blueprint for how data analytics can be harnessed to anticipate audience reception in the dynamic domain of entertainment media.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE
INTRODUCTION

1.1 Overview

From the very beginnings of humankind, entertainment has represented a hallmark of cultural expression and pastime. The media for entertainment have evolved dramatically since the days of ancient amphitheaters, reflecting innovation and the development of human creativity in modern streaming platforms. In this digital age, the success of entertainment media, particularly movies and series, has become a significant area of interest for both academics and industry professionals. It is a growing industry that offers more financial advancements with each new technological growth in the 21st century. This research aims to delve into the complexities of predicting the success of such media before their public release. The entertainment industry, characterized by its vibrant dynamism and rapid evolution, has always been a fertile ground for academic inquiry and practical application. What can be successful and what will not often determine funding and investment in entertainment projects. The advent of big data and advanced analytical techniques has opened new avenues for understanding the factors that contribute to the success of movies and series. This thesis proposes a predictive model based on a comprehensive survey conducted among diverse audiences, aiming to unravel the intricacies of audience preferences and market trends. This study's significance goes beyond scholarly discussions, providing insightful information for industry professionals in the strategic planning and development of entertainment projects. By exploring the statistical patterns of career lengths, activity, and productivity within the entertainment industry, as highlighted by researchers in [1], this thesis seeks to contribute to the ongoing conversation about sustainable success in show business. The empirical distributions of total productivity, in particular, reveal a 'rich-get-richer' phenomenon that underscores the competitive nature of the industry. In constructing the predictive model, this research draws upon various regression techniques and machine learning algorithms, as discussed in existing literature, to analyze the multifaceted aspects of entertainment media success. The model aims to integrate quantitative data from the survey with qualitative assessments of content, audience engagement, and market dynamics, offering a holistic approach to predicting media success. This thesis is structured to first review the historical context and theoretical foundations of entertainment media analysis. It then details the methodology adopted for the survey and model development, followed by an in-depth

discussion of the findings and their implications for both academia and the entertainment industry. The conclusion synthesizes the key insights and suggests directions for future research in this vibrant field of study.

1.2 Motivation

The motivations behind this research are rooted in a fascination with the dynamic interplay between cultural production and audience reception, particularly in the context of entertainment media. As the entertainment industry continues to grow, both in economic scale and cultural influence, understanding the determinants of media success becomes increasingly vital. This study is driven by a desire to bridge the gap between traditional predictive models, which often rely on quantitative metrics and the nuanced realities of audience engagement and content reception in the digital era.

1.3 Rational of the Study

The objectives of this research are multifaceted. Firstly, it aims to develop a comprehensive predictive model that incorporates a wide range of variables, including audience demographics, content analysis, and market trends, to forecast the success of movies and series more accurately. Secondly, the study seeks to explore the psychological underpinnings of audience engagement, examining how emotional and cognitive responses to content can influence viewing behavior and preferences. Finally, this research intends to contribute to the academic discourse on media studies by providing a nuanced understanding of the factors driving entertainment media success, offering valuable insights for academics, industry professionals, and content creators alike.

1.4 Research Questions

The research endeavors to address several pivotal questions within the domain of "Predictive Analytics in Entertainment Media: A Comprehensive Model for Forecasting Media Success and Application for Movies and Series." These inquiries act as road markers, steering the research in the direction of a thorough comprehension of the complexities involved in utilizing cutting-edge artificial intelligence systems for improved forecasts in the entertainment media industry. The overarching research questions include:

How does the incorporation of machine learning influence the sensitivity and specificity of predicting entertainment media success compared to traditional methods? This study investigates the effects of incorporating this paradigm and assesses how well it performs in comparison to conventional techniques. Through

evaluating the impact of these sophisticated methods on prediction accuracy, this research seeks to identify areas for development and obstacles related to their use in the context of successful entertainment media.

What implications do the research findings have for the broader field of entertainment business? This question explores the broader implications of the research findings for the field of entertainment media. The project aims to develop a predictive model for forecasting the success of entertainment media by integrating traditional analytics with emerging trends. It involves a comprehensive review, data collection, and analysis, culminating in a novel digital application for real-time predictions. This tool is designed for both industry professionals and enthusiasts, providing insights into media popularity dynamics. It attempts to evaluate how the study's knowledge might be applied in real-world scenarios, with an emphasis on improving investment and important decision outcomes via more accurate success prediction. It attempts to evaluate how the study's knowledge might be applied in real-world scenarios, with an emphasis on improving investment and important decision outcomes via more accurate success prediction.

1.5 Expected Output

This research-based project will include a novel predictive model for entertainment media success, incorporating a new approach that blends traditional analytics with emerging trends. It includes a thorough literature review, data collection, and analysis, leading to model creation. A significant aspect is the development of a digital application (web or Android-based) that utilizes the model for real-time predictions, aimed at industry professionals and enthusiasts. This project aims to offer a comprehensive tool for forecasting and understanding the dynamics of media popularity. The expected outcomes consist of:

Development of a Robust Entertainment Media Success Prediction System: The final outcome of this thesis presents a novel predictive model that integrates a diverse array of variables to forecast the success of entertainment media, such as movies and series, with a notable degree of accuracy.

Evaluation and Comparative Analysis of Machine Learning Models: There will be a thorough evaluation and analysis of the selected machine learning models. This study will provide light on each model's efficiency, accuracy rates, and performance indicators, enabling readers to make an informed decision about how effective it is to anticipate entertainment media.

Integration of Predictive Strategies: One important expected outcome is the integration of prediction methods into the success prediction system. The technique aims to forecast future trends in addition to media success in the present.

1.6 Report layout

The proposed report is structured with a comprehensive layout to guide readers through theresearch process, findings, and implications. Each chapter serves a specific purpose, contributing to the overall coherence and depth of the document. Chapter 1: Introduction is the introductory chapter which sets the stage for the research, providing a background on the significance of entertainment media, the application of machine learning and the need for a comparative analysis of success prediction. Chapter 2: Background is a detailed exploration of relevant literature and prior research is presented. Chapter 3: Research Methodology outlines the approach taken to conduct the study. It provides insight into the research design, data collection methods, model architecture, and the rationale behind choosing specific methodologies. Chapter 4: Experimental Result showcases the results obtained from the application of machine learning in entertainment media success prediction. Detailed analyses of the performance ofdifferent detection and segmentation approaches are provided, along with visual representations and statistical measures to validate the outcomes. Chapter 5: Impact on Society explores the broader implications of the research findings on entertainment industry. It discusses how an optimized entertainment media success prediction model can positively influence the gross income, global impact, and the overall success of entertainment media. Chapter 6: Summary, Conclusion, Recommendation, and Implications for Future Research, the final chapter serves as a synthesis of the entire report. It summarizes the key findings,reiterates the conclusions drawn from the research, and offers recommendations forpractical applications and further studies.

CHAPTER TWO
BACKGROUND

2.1 Terminologies

In any research project, such as the endeavor to predict viewer satisfaction for entertainment media, background research is paramount. This research equips researchers with a deep understanding of the subject, guiding the exploration of various facets essential for developing accurate predictive models. To understand the field of entertainment media prediction, addressing previous studies is necessary as they will provide valuable insights on what needs to be improved, what approaches should be followed, and what to ignore. These insights, derived from meticulous research, serve as a solid foundation for crafting strategies that enhance and optimize my model design and approach to successfully predict the success of entertainment media.

2.2 Related Works

Nahid Quader, Dipankar Chaki, Md. Osman Gani, and Md. Haider Ali embarked on a research project [2] to devise a machine learning-based decision support system aimed at the movie investment sector. Their study leveraged an extensive dataset extracted from popular film databases such as IMDb, Rotten Tomatoes, Box Office Mojo, and Metacritic. Utilizing a combination of Support Vector Machine (SVM), Neural Networks, and Natural Language Processing techniques, the research focused on predicting the box-office success of movies based on a comprehensive set of both pre-release and post-release features. The investigation revealed that Neural Networks achieved an impressive accuracy of 84.1% when utilizing pre-release features alone, and this accuracy further improved to 89.27% when all features were considered. Conversely, the SVM model showcased an accuracy of 83.44% for pre-released features and 88.87% for the full feature set. The study also highlighted the significant impact of factors such as budget, IMDb votes, and the number of screening theaters on a movie's box-office performance, with these elements being pivotal in predicting a film's commercial success. Narayana Darapaneni et al. conducted an extensive study [3] aimed at mitigating the financial risks inherent in the film industry by predicting movie success through machine learning. Utilizing the IMDB database, the research focused on key movie attributes such as crew, storyline, box-office revenue, and critical and audience reviews. A variety of machine learning algorithms were explored, including Random Forest, Decision Tree, KNN, NLP, XGBoost Classifier, and Deep

Neural Networks, to determine the predictive accuracy regarding the success of movies. The XGBoost Classifier (90.39%) was identified as the most effective model in this context. The study underscored the importance of user reviews, gross income, and budget as critical predictive features, offering valuable insights into the movie industry's strategic planning and decision-making processes. Anand Bhave et al. [4] delved into the complex terrain of forecasting movie success amidst the burgeoning digital and social media landscapes that are reshaping the film industry. Recognizing the challenge posed by the vast number of movies produced annually and the critical need for reliable predictive models to enhance business prospects, their research posits a novel approach. They advocate for a holistic integration of traditional movie attributes—cast, director, producer—and dynamic social factors like online community responses to more accurately gauge a movie's potential success. Their study underscores the limitations of existing predictive models that rely heavily on either classical or social elements in isolation, thus failing to capture the nuanced interplay that more accurately reflects movie success dynamics. By examining the intricate relationships among traditional movie elements and incorporating real-time social media analytics, Bhave et al. aim to refine the predictive accuracy for the movie industry's stakeholders. Currently, their accuracy rate is Linear Regression (70.57%). Rijul Dhir and Anand Raj's analysis [5] revealed that the Random Forest algorithm provided the highest prediction accuracy (61%) among various tested algorithms, a notable advancement in movie success prediction methodologies. Their exploratory analysis also highlighted that the number of voted users, critics' reviews, social media interactions, movie duration, and gross collections significantly influenced IMDb scores. Particularly, genres like Drama and Biopic were found to be more favorable among audiences, contributing positively to movie ratings. This comprehensive approach, combining traditional movie metrics with social media data, offers new insights into factors affecting movie success, underscoring the effectiveness of machine learning in film industry analytics. Jeffrey Ericson and Jesse Grodman at Stanford University conducted a study [6] to identify factors contributing to a movie's success using machine learning. They collected data from IMDB, Rotten Tomatoes, and Wikipedia, focusing on attributes known before a movie's release, such as budget, cast accolades, and genre. The final dataset comprised 6,590 movies, after excluding those with fewer than 100 votes to ensure rating reliability. The researchers applied locally weighted linear regression and SVM (Support Vector Machine) with filter feature

selection to predict success across five different measures. They discovered that genre, MPAA rating, runtime, and studio were significant predictors, with budget and cast awards also playing a role depending on the success measure. Their SVM model achieved about 70% accuracy for rating predictions and 88% for monetary gross predictions. The study highlights the complexity of predicting movie success and suggests that while their model offers valuable insights, the diversity and unpredictability of the film industry limit the practical application of such predictive models. Interactive Storytelling (IS) technologies, integrating advancements like advanced visualization and artificial intelligence, present a significant evolution in entertainment experiences, allowing users to actively influence and shape narratives. Christoph Klimmt, Christian Roth, Ivar Vermeulen, Peter Vorderer, and Franziska Susanne Roth conducted comprehensive research [7] to explore the potential user experiences facilitated by IS technologies. Through theoretical analysis and expert interviews, their research highlighted the importance of system usability and character believability as essential for engaging user interaction. These foundational elements support users' perception of their impact on the story and foster emotional connections with characters, leading to immersive experiences. The study identified various experiences such as curiosity, suspense, and flow, illustrating the diverse entertainment possibilities offered by IS. This intricate ability to deliver personalized narrative experiences underscores the critical role of interactive storytelling in their research on the future of entertainment technologies which impacted this research’s approach as well. Christina Hofmann-Stölting, Michel Clement, Steven Wu, and Sönke Albers conducted a study [8] on sales forecasting for new entertainment media products, exploring the effectiveness of diffusion models and simple regression analyses compared to traditional management predictions. Their research, spanning the German markets for music, film, and literature, indicates that model-based forecasts generally outperform management predictions for a majority of products, except for top-sellers, where management's insights, possibly due to more focused attention and resources, yield better accuracy. The study highlights that advertising, product quality indicators like star power and critics' reviews, and the product's country of origin are consistently significant across all three industries for sales forecasting. Other factors, such as price and distribution power, showed variable significance depending on the industry. The research reveals that while complex diffusion models didn't significantly outperform simpler regression models for total sales predictions, they offer detailed insights into

sales distribution over time, which can be valuable for strategic planning. This comprehensive analysis underscores the complexity of forecasting in the entertainment sector and suggests that integrating model-based forecasts with managerial insights, especially for potentially high-selling products, could enhance prediction accuracy and strategic decision-making. This approach aligns with the broader need in media economics to refine demand forecasting for new media products, contributing valuable insights to both academic and practical realms. In their research [9], Masrury et al. explored the use of machine learning techniques to predict the success of Hollywood movies prior to their release. They compared three models: Artificial Neural Network (ANN), Naïve Bayes, and Support Vector Machine (SVM), using pre-release data such as genre, budget, and director from IMDb. The study aimed to mitigate the financial risks for industry stakeholders by enhancing the predictability of movie success. The findings revealed that ANN was the most effective model, achieving an 80% accuracy rate and high scores in precision and recall, making it a reliable tool for assessing movie success risks. Naïve Bayes showed moderate performance, while SVM was the least suitable due to lower accuracy. This research by Masrury et al. underscores the potential of ANN in forecasting Hollywood movie outcomes, offering valuable insights for producers, distributors, and exhibitors in their decision-making processes. In a study [10] focused on evaluating machine learning classification techniques for predicting movie box office success, Quader and colleagues compared seven algorithms: Support Vector Machine (SVM), Logistic Regression, Multilayer Perceptron Neural Network (MLP), Gaussian Naive Bayes, Random Forest, AdaBoost, and Stochastic Gradient Descent (SGD). The research utilized a dataset of 755 movies, incorporating both pre-released and post-released features, sourced from IMDb, Rotten Tomatoes, Box Office Mojo, and Meta Critic. The success of movies was quantified based on box office profit, with films categorized into five classes from flop to blockbuster. The models' performances were gauged through the exact match and one-away predictions, the latter accounting for minor deviations in class predictions due to marginal target class values. The Multilayer Perceptron Neural Network emerged as the most effective model, achieving a 58.53% success rate in exact match predictions. This highlights MLP's capability to handle the complex data patterns and overlapping points characteristic of movie success classification, providing valuable insights for decision-making in the movie industry. Dewan Muhammad Qaseem and his team developed a method [11] to predict movie success rates using optimal sentiment analysis of Twitter data. The

research utilized a lexicon-based approach, applying three different dictionaries—AFINN, Socialsent, and Rating Warriner—to assess the sentiment of tweets related to movie trailers. Each dictionary has a unique word count and rating scale, which were normalized to a 0-5 star rating system for comparison. The team collected Twitter data using hashtags related to movie trailers and calculated the sentiment scores for each tweet. These scores were then averaged to predict the overall movie rating. The predicted ratings were rescaled to a 0-5 star system to align with conventional movie rating scales. In the comparative study of machine learning algorithms for sentiment analysis of Twitter data to predict movie success rates, the Linear SVC algorithm emerged as the top performer with an impressive 84% precision. Alongside, the K-Nearest Neighbor algorithm also demonstrated commendable efficacy, matching the 84% precision mark set by Linear SVC. In their study [12], Michael T. Lash and Kang Zhao developed a system aimed at predicting the profitability of movies in the early stages of production. By integrating data from various sources and employing techniques such as social network analysis and text mining, the system evaluates multiple aspects including the cast involved ("who"), the movie's theme ("what"), and the release timing ("when"). This approach seeks to aid movie investors by providing early predictions of movie profitability, thus supporting more informed decision-making. The research findings demonstrated that the system outperformed benchmark methods significantly. Novel features introduced by Lash and Zhao, particularly those related to the cast's profitability records and the movie's content, contributed substantially to the prediction accuracy. The study not only presents a practical tool for the movie industry but also sheds light on the key factors influencing movie profitability. Furthermore, the research illustrates how the system can be used prescriptively to suggest a profit-maximizing cast, showcasing the potential of predictive and prescriptive data analytics in strategic business decisions within the entertainment sector. The study revealed that the Random Forest algorithm emerged as the most effective model, boasting an impressive 90% accuracy rate. This high level of precision underscores the model's reliability in evaluating the potential success of movie projects. The findings highlight the significant impact of incorporating novel features, especially those related to the cast's track record in profitability and the thematic elements of the movie, on the accuracy of predictions. The research not only offers a valuable tool for decision-making in the film industry but also emphasizes the critical factors that contribute to a movie's financial success. Moreover, it demonstrates

the system's capability to provide recommendations for assembling a cast that maximizes profitability, showcasing the practical applications of predictive and prescriptive analytics in the entertainment industry's strategic planning processes.

In their comparative analysis [13], Verma and Verma utilized a range of machine-learning techniques to predict the success of Bollywood movies prior to their release. The study focused on models including Decision Tree, Random Forest (RF), Support Vector Machine, Logistic Regression, Adaptive Tree Boosting, and Artificial Neural Networks, with predictors like the lead actor's ratings, IMDb movie rankings, music rankings, and the total number of screens allocated for the movie's release. The research findings indicated a general accuracy range between 80% to 90% for most models. However, the models based on Random Forest and Logistic Regression techniques stood out, achieving a higher accuracy rate of 92%. This highlighted the significant predictive power of these two models in assessing a movie's potential success. Among the predictors analyzed, the music rating emerged as the most influential factor, followed by the movie's IMDb rating and the number of screens planned for its release.

Ashutosh Kanitkar's research [14] focuses on using machine learning algorithms to predict the success of Bollywood movies before their release. The study is significant due to the substantial financial implications involved in the Bollywood industry. Kanitkar employed seven regression algorithms and six classification algorithms to predict box office success and categorize movies into nine profitability classes. The best-performing regression technique was found to be the Artificial Neural Network (ANN), which excelled in predicting net India box office collections. For classification, the study identified ANN, K Nearest Neighbors (KNN), and Random Forest as the top performers, with ANN achieving an accuracy of 50% and KNN and Random Forest both reaching 49.33%. These results highlight the potential of machine learning in providing valuable insights for producers and distributors, enabling more informed decision-making to ensure profitability in the highly competitive Bollywood film market.

In the project "Movie Success Prediction using Data Mining" by Saurabh Kumar, Avinay Mehta, and Joy Pal [15] data mining techniques and machine learning algorithms were applied to predict movie successes. Utilizing R software, the study focused on various factors such as budget, cast, and release timing, gathered from IMDb. The project aimed to provide insights for stakeholders in the movie industry to make informed decisions regarding potential blockbusters. The methodology involved preprocessing IMDb data, generating training and testing datasets, and building

predictive models. Key attributes like critic scores, IMDb ratings, and audience scores were analyzed, revealing a strong correlation between critic and audience scores as significant predictors of movie success. The use of linear regression in the analysis yielded a confidence level of 95%, indicating a high reliability in the model's predictions. This project highlights the potential of leveraging data analytics in the movie industry to predict film performance, offering a data-driven approach to enhance movie production and marketing strategies. In the study [16] by Kyuhan Lee, Jinsoo Park, Iljoo Kim, and Youngseok Choi, the focus was on enhancing the predictive accuracy of movie success using machine learning techniques. The research highlighted two main strategies for improving prediction models: introducing new features and employing ensemble methods. the study employed an ensemble approach, named the Cinema Ensemble Model (CEM), which combined predictions from multiple models to enhance overall accuracy. It incorporated linear regression which resulted in the most desired output success of 88.3%. This study [17] conducted a comparative analysis of data mining techniques to predict the success of Bollywood movies, utilizing a dataset compiled from sources like IMDb, Wikipedia, and Rotten Tomatoes. The dataset included various movie-related parameters, such as title, runtime, release date, genre, and director. The researchers employed several prediction algorithms, including Decision Tree, Naïve Bayes, Logistic Regression, Adaboost, and KNN. Among the tested algorithms, Naïve Bayes emerged as the most effective, offering the best result with an accuracy rate of 82.8%. This finding suggests that Naïve Bayes could be a valuable tool for production houses and investors in making informed decisions about movie releases and investments in the Bollywood film industry. In the research [18] conducted by Partha Chakraborty, Md. Zahidur Rahman, and Saifur Rahman from Comilla University, a modified form of linear regression was utilized to predict movie success. The model demonstrated a 100% success rate in their selected experiments. However, the researchers did not expand their study further to obtain more detailed or comprehensive information. This limited scope means the model's effectiveness across a broader range of movies and variables remains untested. Pojan Shahrivar and Carl Jernbäcker [19] explored the use of machine learning to predict movie success, focusing on ratings and box office revenue from pre-release metadata. They found that Support Vector Machine (SVM) was the most successful technique, achieving a 68.8% accuracy rate in their predictions. While the study demonstrated the potential of machine learning in the film industry, it also highlighted the need for larger datasets and more

comprehensive features to improve prediction accuracy. In their research [20], Manav Agarwal et al. conducted a comprehensive comparison of various models to predict movie success, focusing on box office performance and popularity. They explored Regression models, Machine Learning models, a Time Series model, and a Neural Network, with the Neural Network achieving the highest accuracy at approximately 86%. The analysis included an examination of 2020 movie releases to enhance the dataset. Key methods employed were Simple Linear Regression (SLR) and Multiple Linear Regression (MLR), with SLR using the Variance Inflation Factor (VIF) to select predictive attributes, and MLR utilizing multiple variables, including genre, to predict movie success. Both models underwent statistical testing for validation. Subramaniyaswamy V. et al. [21] explored the effectiveness of multiple linear regression and Support Vector Machine (SVM) Classification in predicting the box-office success of Hollywood movies. The research involved collecting and cleaning data from BoxOfficeMojo, Wikipedia, and YouTube for movies released in 2016. The final dataset comprised 138 films after pruning incomplete or irrelevant entries. These films were categorized based on their budgets to calculate an adjusted ROI, which served as a measure of success. To predict the best ROI factor, the researchers employed SVM Classification, categorizing movies into four groups based on their adjusted ROI. This classification aimed to distinguish between box office flops, break-even films, moderately successful films, and hugely successful films. The SVM model achieved an accuracy rate of 56.52%, which was a significant improvement over previous studies that relied solely on IMDb data. In their research [22], Márton Mestya´n, Taha Yasseri, and János Kertész explored the potential of using "big data" from Wikipedia to predict the box office success of movies. They demonstrated that the level of activity on a movie's Wikipedia page, including the number of page views, edits, and the diversity of editors, can serve as an early indicator of a movie's popularity and financial performance. The study focused on movies released in the United States in 2010, tracking 312 films for which Wikipedia data was available. The researchers observed that Wikipedia articles about movies are often created well before the movie's release, allowing them to analyze the buildup of interest over time. The further exploration of this study is showcased in my research’s data collection approach. Ankit, Lakshmi M, K. Aditya Shastry, Aditya Sandilya, and Rahul Shekhar conducted a comparative analysis [23] of various Machine Learning approaches for predicting movie success. They applied algorithms like Support Vector Machine (SVM), Gaussian Naïve Bayes

(GNB), and k-nearest Neighbors (k-NN) to a dataset from IMDb, focusing on attributes such as actors, genres, directors, and budget. Among these, the Gaussian Naïve Bayes method outperformed others, achieving a success rate prediction accuracy of 85.4%. This study underscores the potential of Machine Learning in forecasting the financial outcomes of movies, thereby aiding stakeholders in making informed decisions. Muzammil Hussain Shahid and Muhammad Arshad Islam [25], propose a novel approach to predict movie profitability using genre popularity features derived from time series analysis. They suggest that a movie's box office success is influenced by the popularity of its genre at the time of release. To test this hypothesis, they introduce new features based on genre popularity, including budget, revenue, frequency, success, and return on investment (ROI), and integrate these with the movie's release timing to train machine learning classifiers. The study reveals that the Gradient Boosting classifier, when enhanced with the proposed features, significantly outperforms existing methods, achieving an accuracy of over 92.4%. This represents a 35.7% improvement over the state-of-the-art, addressing a multi-class problem in movie success prediction. The research highlights the dynamic nature of genre popularity and its impact on movie profitability, suggesting that investors and producers can benefit from considering these factors in the early stages of movie development.

Sagar Sadashiv et al. [26] proposed a methodology that involves data preprocessing to clean and prepare the dataset, importing classification algorithms like Decision Tree, KNN, SVM, Logistic Regression, and Random Forest, and evaluating the performance of these models to select the one with the highest accuracy. The chosen model, Random Forest, demonstrated an 85.20% accuracy rate, making it the preferred model for predicting movie success. In their research [27], K. Meenakshi, G. Maragatham, Neha Agarwal, and Ishitha Ghosh employed the K-means clustering technique as part of their data mining approach to analyze and predict movie success. The K-means clustering was utilized to categorize movies based on various attributes relevant to their performance, such as actors, directors, genres, and more. Their analysis using this method yielded a success rate of 68.8%, demonstrating the potential effectiveness of data mining techniques in forecasting the success of movies in the Bollywood film industry.

2.3 Comparative Analysis and Summary

The comparative analysis of my literature review is given below:

TABLE 2.1: Comparative Analysis with Previous Work

SL No	Author Name	Used Algorithm	Best Accuracy Algorithm
1	Nahid Quader, Dipankar Chaki, Md. Osman Gani, Md. Haider Ali [2]	SVM, Neural Networks, NLP	Neural Networks (89.27%)
2	Narayana Darapaneni et al. [3]	Random Forest, Decision Tree, KNN, NLP, XGBoost Classifier, Deep Neural Networks	XGBoost Classifier (90.39%)
3	Anand Bhave et al. [4]	Linear Regression	Linear Regression (70.57%)
4	Rijul Dhir and Anand Raj [5]	Random Forest	Random Forest (61%)
5	Jeffrey Ericson and Jesse Grodman [6]	Locally Weighted Linear Regression, SVM	SVM (88% for monetary gross)
6	Masrury et al. [9]	ANN, Naïve Bayes, SVM	ANN (80%)
7	Quader et al. [10]	SVM, Logistic Regression, MLP, Gaussian Naive Bayes, Random Forest, AdaBoost, SGD	MLP (58.53%)
8	Dewan Muhammad Qaseem et al. [11]	Lexicon-based Sentiment Analysis, Linear SVC, K-Nearest Neighbor	Linear SVC and KNN (84%)
9	Michael T. Lash and Kang Zhao [12]	Random Forest, SVM	Random Forest (90%)
10	Verma and Verma [13]	Decision Tree, Random Forest, SVM, Logistic Regression, Adaptive Tree Boosting, Artificial Neural Networks	Random Forest and Logistic Regression (92%)
11	Ashutosh Kanitkar [14]	ANN, KNN, Random Forest	ANN (50%)

12	Saurabh Kumar, Avinay Mehta, Joy Pal [15]	Linear Regression	Linear Regression (95%)
13	Kyuhan Lee et al. [16]	Linear Regression, Ensemble Approach (CEM)	CEM (88.3%)
14	Khandelwal, R., & Virwani, H. [17]	Decision Tree, Naïve Bayes, Logistic Regression, Adaboost, KNN	Naïve Bayes (82.8%)
15	Partha Chakraborty et al. [18]	Modified Linear Regression	100%
16	Pojan Shahrivar, Carl Jernbäcker [19]	Support Vector Machine (SVM)	SVM (68.8%)
17	Manav Agarwal et al. [20]	Regression models, Machine Learning models, Time Series models, Neural Network	Neural Network (86%)
18	Subramaniyaswamy V. et al. [21]	Multiple Linear Regression, SVM Classification	SVM Classification (56.52%)
19	Ankit et al. [23]	SVM, GNB, k-NN	GNB (85.4%)
20	Muzammil Hussain Shahid and Muhammad Arshad Islam [25]	Gradient Boosting	Gradient Boosting (92.4%)
21	Sagar Sadashiv et al. [26]	Decision Tree, KNN, SVM, Logistic Regression, Random Forest	Random Forest (85.20%)
22	K. Meenakshi et al. [27]	K-means Clustering	K-means Clustering (68.8%)

2.4 Gap Analysis

Entertainment media prediction is an exciting topic that has been covered by many. These studies have employed a wide range of machine-learning algorithms and data collection approaches. However, upon reviewing the studies, research, and projects the following gap analysis can be made:

High Dependence on Historical Data: Almost every model in this form of prediction relies heavily on historical data, which may not fully capture rapidly changing trends in movie consumption and audience preferences. Factors before the release of any media aren't considered most of the time.

Lack of Real-Time Data: There is a gap in incorporating real-time data, such as evolving social media sentiment or real-time box office numbers, into predictive models which will lead to prediction failure as the models aren't addressing the psychology of the viewers.

Handful Interdisciplinary Approaches: Few studies combine insights from machine learning, film studies, psychology, and marketing to create more holistic and accurate predictive models.

Limited Cross-Cultural Studies: Few studies address the predictive models' applicability across different cultural and regional markets, a significant gap given the global nature of the film industry.

Inconsistency in Feature Importance: While there are clear indications among studies, there is no consensus on which features are most predictive of movie success, indicating a need for more systematic feature analysis and selection methods.

2.5 Scope of the Problem

The extent to which the issue covered in this study, with a focus on "Predictive Analytics in Entertainment Media: A Comprehensive Model for Forecasting Media Success and Application for Movies and Series," extends to the realm of sentiment analysis, specifically within the context of behavioral patterns and personal preference among adults. The entertainment industry is always expanding, and successful products need to be predicted with accuracy and timeliness. The range includes the difficulties posed by conventional prediction techniques, which may have restrictions due to their heavy reliance on historical data, ignoring social norms, global trends, and overall prediction accuracy.

CHAPTER THREE
RESEARCH METHODOLOGY

3.1 Research Subject

The experiments for this study were conducted on Google CoLab using Matlab and Panda's library. TensorFlow, one of the greatest deep learning frameworks, was used to handle machine learning methods on Python. Each model was developed by Google and trained on a Tesla graphics processing unit (GPU) in the cloud before being made available through the Google Collaboratory framework. The Collaboratory framework provides up to 16GB of RAM (random-access memory) and about 360GB of GPU in the cloud for research use. The study makes use of frameworks and programming languages like TensorFlow and Python to create deep learning models and related techniques. The utilization of GPUs for quicker processing in computational infrastructure is essential for managing the intricacy of deep learning.

3.2 Data Collection Procedure

To conduct this research, I worked with raw data. For data collection, I used survey-based questionnaires. However, selecting the appropriate questions for this research was a tough choice. This is due to the fact that, unlike other predictive analyses, there isn't a specific scale like SAS that I can follow. Upon research and countless paper reviews, I came across Praxis [28] who created their own standardized scale from 0 to 9 for predicting movie success which I based my questions upon as my target was to obtain real-time data on audience preference. Therefore, I utilized tools like Google Forms, SurveyMonkey, and Qualtrics. However, the most difficult task was to determine the questions. I had to ensure the design aligned with my research questions and objectives which led me to revisit research and studies of my predecessors in this field to base my (26) questions on.

After selecting the questions, I created a Google form to collect data from people. I used a six-point Likert scale to collect the data that was used for this research. Basically, it is a psychometric measurement scale that is frequently used in research and surveys to evaluate participants' attitudes, opinions, and feelings. It gives participants a variety of ways to indicate whether they agree, are satisfied, or disagree with a statement. Participants are given six response options on a six-point Likert scale, each of which represents a different degree of agreement or disagreement, satisfaction or dissatisfaction, frequency, or intensity. Usually, the scale goes from extremely satisfied

to extremely dissatisfied, or from strongly agree to strongly disagree. The points are given below:

- Strongly Disagree
- Disagree
- Somewhat Disagree
- Somewhat Agree
- Agree
- Strongly Agree

After creating the form, I spread it using social media and survey-based platforms. Most of the responses it got were from students, especially university students. Collecting data was also a hard process as I was lacking manpower. So, I implied physical surveys on restaurants, local malls, and parks (popular places where my target audience roams) to conduct the survey. Fortunately, I managed to get 2835 responses. The questions that were used in the research are given below:

TABLE 3.1: Questionaries Used and Their Related Factors

Question	**Factor**	**Reference**
Do you think having a celebrity personality (actor, director, writer) makes you want to see a movie or show?	Historical Data	[8]
How important are ratings and reviews (IMDB, Rotten Tomatoes) of trailers, teasers, and sneak peeks influencing your decision to watch or interact with entertainment media?	Critical Reception	[2]
Do you prefer entertainment media that offers diverse representation in terms of gender, race, ethnicity, and LGBTQ+ characters?	Sentiment Metrics	[7]
Do you watch movies, or series based on the hype and buzz they have generated on social media platforms (X, Reddit, Facebook, Instagram)?	Online Engagement	[4]
Do you think a complex plot is necessary for a movie or show to be successful?	Interactive Storytelling	[15]
When selecting a show or movie to watch, do prefer familiarity with the source material or follow-up (book or comic adaptation, Sequel/Prequel, Franchise)?	Box Office Metrics	[3]

Do you pick or drop a movie or show from your watchlist after getting initial screening reviews?	Critical Reception	[2]
How important is the availability of entertainment media on streaming platforms (Netflix, Amazon, Disney Plus) for your viewing?	Distribution and Accessibility	[27]
Do you tend to watch entertainment media that aligns with your cultural or social interests?	Sentiment Metrics	[12]
Are you more inclined to watch media that address social, political, or historical events/issues?	Sentiment Metrics	[25]
Have you ever decided to watch movies or series based on a recommendation from a social media influencer or celebrity endorsement?	Trend	[11]
Do you tend to watch or read entertainment media that explores unconventional or non-mainstream topics and narratives?	Interactive Storytelling	[25]
Do you look into the production values (e.g., cinematography, set design, costumes) before making your decision to watch movies or TV series?	Budget & Financials	[2]
Do you prefer entertainment media that offers a thought-provoking or intellectual experience?	Interactive Storytelling	[25]
Are you more likely to watch or read entertainment media based on a true story or real-life events?	Historical Data	[5]
How important are character development and depth in your decision to watch or read entertainment media?	Interactive Storytelling	[3]
Do you feel interested in a movie or show based on the artwork or music featured in it?	Sentiment Metrics	[19]
Do you feel interested in a movie or show with an extremely high budget?	Budget & Financials	[2]
Are you likely to watch animated content formatting more than regular video formatting?	Interactive Storytelling	[25]
Movies and shows with fantasy/supernatural elements are a must-watch for you?	Sentiment Metrics	[22]

Do you feel compelled to watch movies and series that are R-rated (Explicit scenes, Blood-gore, Nudity)?	Sentiment Metrics	[17]
How likely are you to watch a movie or show that is creating/marketing some sort of trend (E.G Barbenheimer)?	Trend	[4]

The Metrics/Factors for the Questions

Box Office Metrics:

- **Opening Weekend Revenue:** Often serves as a strong indicator of a movie's overall financial success.
- **Total Gross Revenue:** The total worldwide box office earnings.
- **Revenue Over Time:** How earnings change week-by-week can indicate a movie's longevity and audience retention.

Budget and Financials:

- **Production Budget:** The cost to produce the movie, which when compared to earnings, can indicate profitability.
- **Marketing Spend:** The amount spent on promoting the movie, can impact its initial box office performance.
- **Return on Investment (ROI):** A ratio of net profit to cost, indicating overall financial success.

Online Engagement and Sentiment Metrics:

- **Social Media Mentions:** The volume of mentions on platforms like Twitter, Instagram, and Facebook can reflect audience interest.
- **Sentiment Analysis:** Evaluating the sentiment (positive, negative, neutral) of social media posts, reviews, and comments to gauge public opinion.
- **Engagement Rate:** Likes, shares, comments, and other interactions relative to the audience size on social media platforms.

Critical Reception:

- **Critics' Reviews:** Aggregated scores from critics, such as those on Rotten Tomatoes or Metacritic.
- **Awards and Nominations:** Recognition from film festivals and award bodies can contribute to a movie's prestige and audience interest.
- **Viewer Ratings:** Audience ratings on platforms like IMDb and Rotten Tomatoes.

Historical Data and Trends:

- **Genre Performance:** Historical performance of movies within the same genre.
- **Seasonal Trends:** Movie success trends related to release timing (e.g., summer blockbusters, holiday releases).
- **Director and Cast Track Record:** Past success of the director, lead actors, and supporting cast in similar roles or genres.

Distribution and Accessibility:

- **Number of Screens:** The number of theaters showing the movie can affect its box office earnings.
- **International Release:** Availability in key international markets and performance in those markets.
- **Streaming and Digital Sales:** Performance on digital platforms, especially for movies released directly online.

Interactive Storytelling

- **Story Complexity:** How the story is expressed to the viewers.
- **Casting, Development & Relatability:** Availability of understandable and relatable entertainment

3.3 Statistical Analysis

Most of the data that was used in this research was collected through google form. I selected those questions through meticulous calculation and theories based off of previous researches in the field. I collected a total of 2835 data. All of the data were scale based where the scale value was from 1 to 6. The collected data followed on mentioned metrics like box office information, budget, online engagement, crowd sentiment, critical reception, historical data, storytelling and plot. These features help significantly in identifying the pattern which were used to predict success of entertainment media prior to their release. As for demographic data, I collected age and gender of the participants. The visual representation of the demographic data is given below:

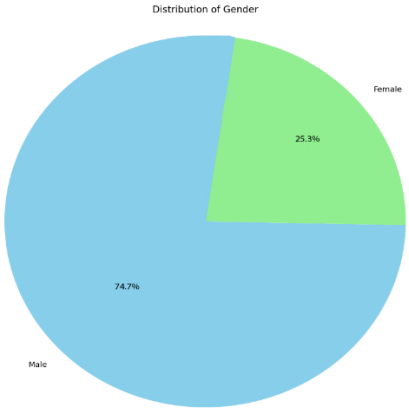

Figure 3.1: Gender Distribution

In the pie chart from figure 3.1, we can see that the majority of the individuals who participated in this survey are male which is 74.7% and the rest is female with 25.3%.

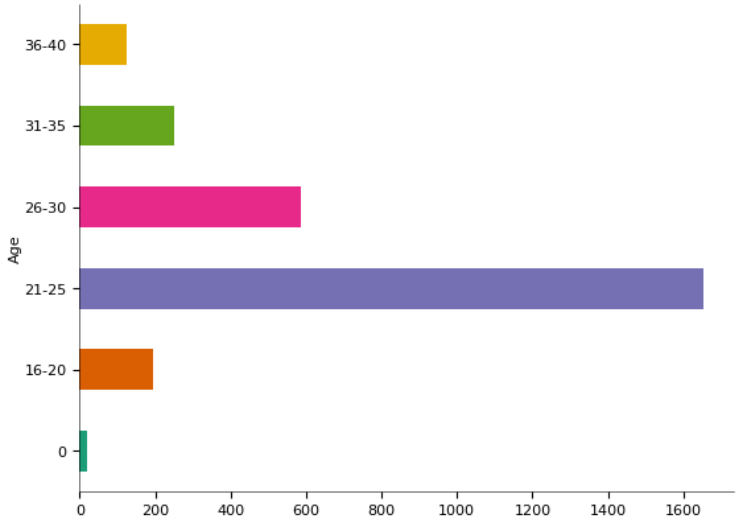

Figure 3.2: Age Distribution

In the figure 3.2, we can see that, most of the participants are from age group 21-25 followed by the age group 26-30 If we go beyond age 30, the number significantly decreases.

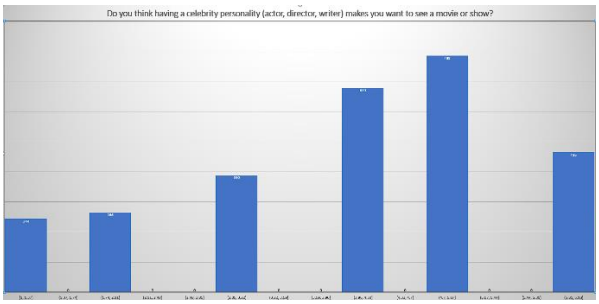

Figure 3.3: Celebrity Preference

In the figure 3.3 it shows that majority (788 out of 2835) of people agrees that the inclusion of celebrities significantly improves the chances of them being interested to

watch a movie or show.

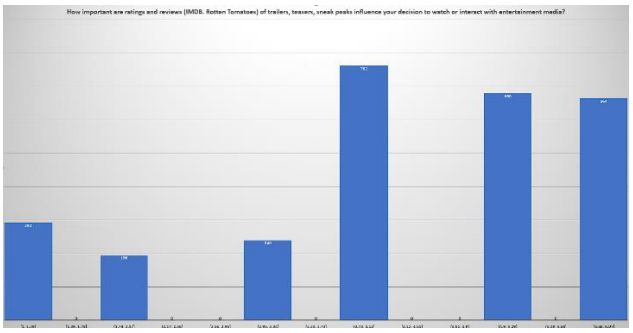

Figure 3.4: Ratings & Reviews

In the figure 3.4, we can see that majority people (762) somewhat agrees selecting entertainment media based on their initial reviews while 664 people out of 2835 strongly agrees on online reviews before selecting to invest themselves on a movie or show.

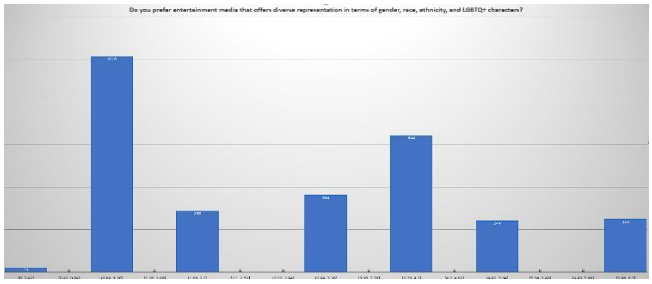

Figure 3.5: Diversity

In the figure 3.5 , it can be seen that most people (1016) don't pick an entertainment media based on the representation of diversity. While some agrees (644) that diversity can persuade them to invest on a show and there are others (252) heavily rely on diverse cast and plot for their media preference.

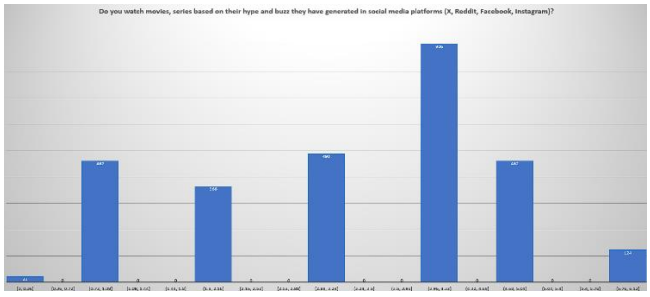

Figure 3.6: Social Media Buzz

In the figure 3.6 it is showcased that people would likely to pick up on a new entertainment media, if it is attracting a great deal of attraction in social media platform. At the same time, there are a significant portion of the audience that ignores social media buzz completely.

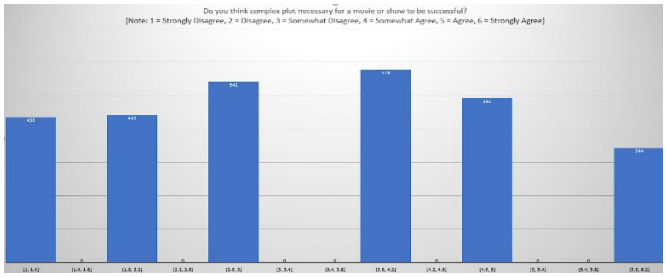

Figure 3.7: Intriguing Plot

Figure 3.7 shows the audience is equally divided when it comes to complex plots.

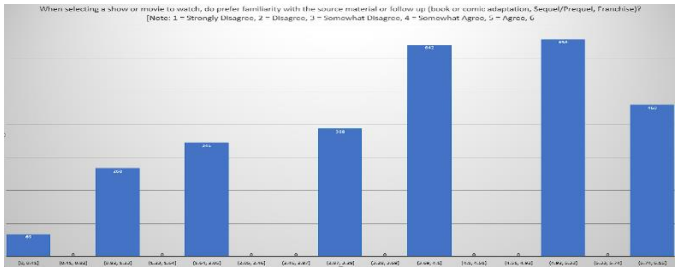

Figure 3.8: Source Material Familiarity

In the figure 3.8, we can see that most audiences tend to interact with entertainment media which have previous source materials that they are familiar with.

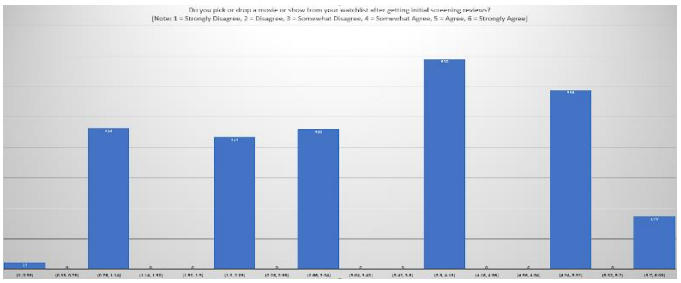

Figure 3.9: Screening Reviews

In the figure 3.9 it is showcased that most audiences (690) somewhat agree that they rely on initial screening reviews for movies and shows to invest themselves into them.

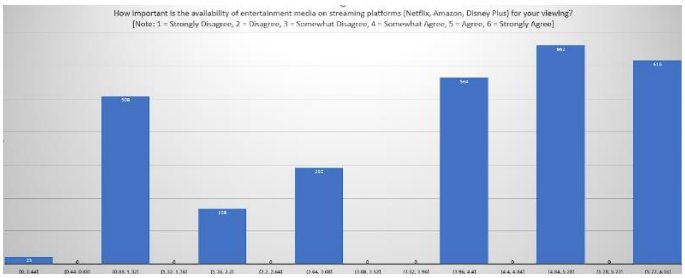

Figure 3.10: Age Distribution

In the figure 3.10 we see that online platform releases are most preferred by users.

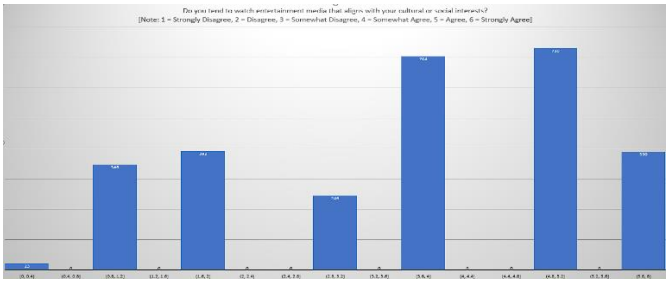

Figure 3.11: Cultural Influence

In the figure 3.11, entertainment media that explores cultural and social influences are highly likely to be accepted by audiences as 1400+ users agree that they are likely to engage with media that offers such experience.

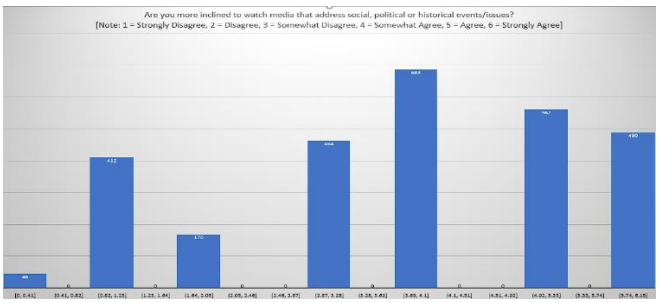

Figure 3.12: Social, Political & Historical Issues

In the figure 3.12, we see that a high number of audiences prefers to see entertainment media that addresses social, political or historical issues.

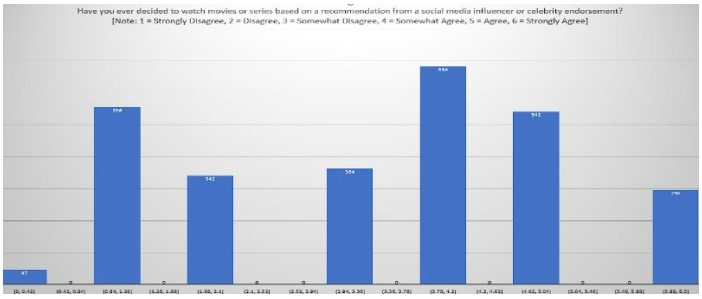

Figure 3.13: Influence of Social Media Personalities

In the figure 3.13, it is showcased that majority of people somewhat agree that they might start watching something if it's recommended by social media influencers.

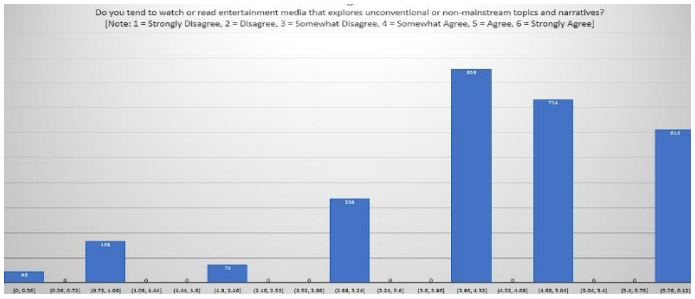

Figure 3.14: Unique Content

In the figure 3.14, we see that almost 90% of the users would like to interact with a movie or show if it is capable of introducing unique and complex stories that doesn't entertain standard storytelling practice.

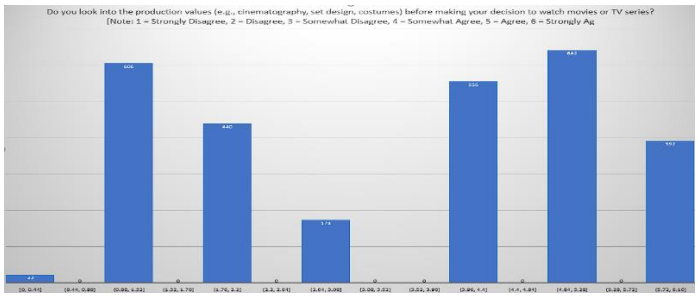

Figure 3.15: Production Cost

In the figure 3.15 we see a diverse opinion on production value. A high majority of users 1500+ prefer shows interest in the production value while similar albeit a bit lower amount shows no interest at all.

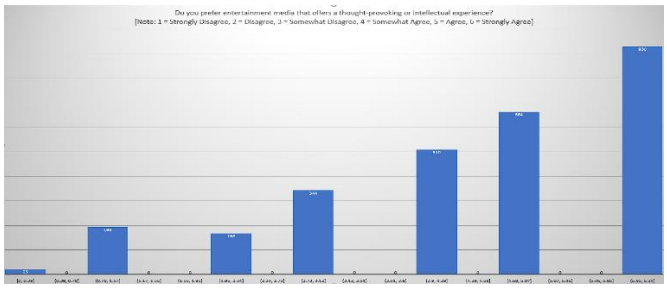

Figure 3.16: Thought Provoking Experience

In the figure 3.16, we see a rapid growth of users being interested in entertainment media that provides an intellectual experience more than anything. Only a handful of users declined to mark this as a requirement for success in their books.

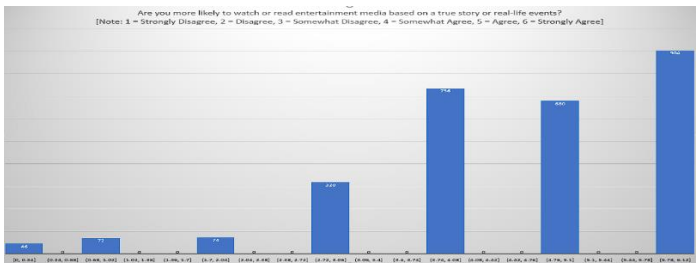

Figure 3.17: Real Life Stories

In the figure 3.17, it can be noted that majority of audience prefers to see stories that portray real life events.

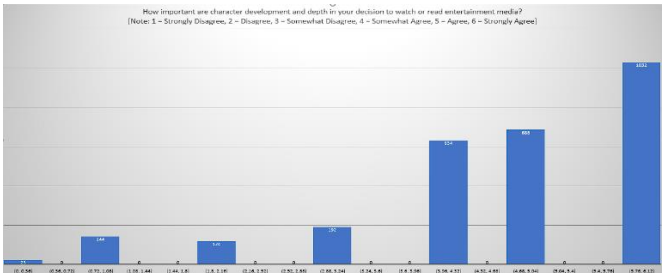

Figure 3.18: Character Development

In the figure 3.18, audiences voted on character development as a key aspect for an entertainment media to success with more than 2000 votes on the positive.

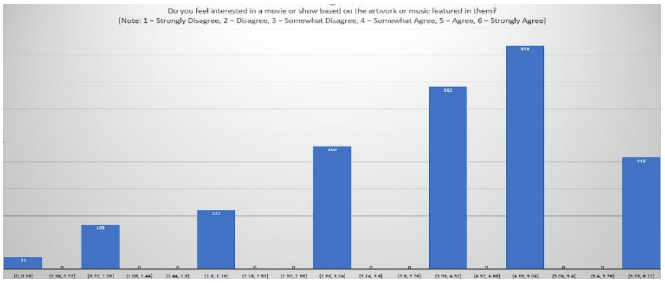

Figure 3.19: Artwork & Music

In the figure 3.19, we see that a majority of the users agree that dedicated artwork and music attracts them to support a new entertainment media such as movies or series.

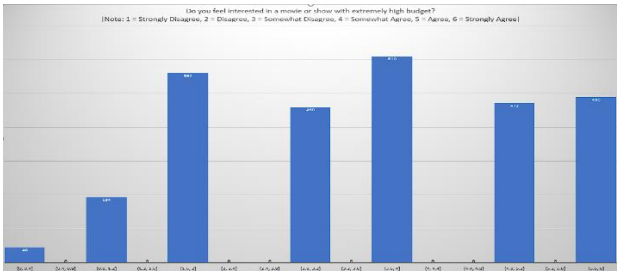

Figure 3.20: Budget

In the figure 3.20, it is showcased that people somewhat look into high budget entertainment projects a bit more than low budget ones.

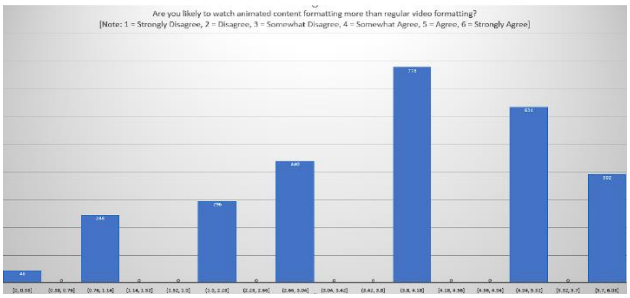

Figure 3.21: Animation

In the figure 3.21 depicts that animated shows and movies have a high possibility of getting chosen among adult audiences.

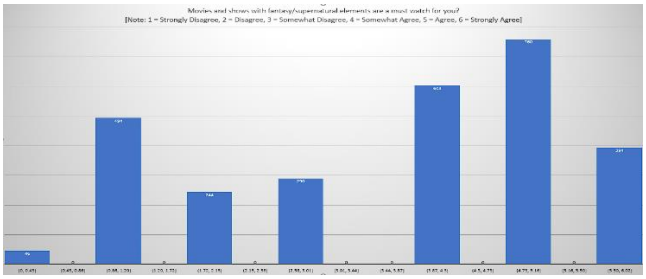

Figure 3.22: Fantasy/Supernatural

In the figure 3.22, we see that majority of the audience (1364) would like to watch supernatural and fantasy elements but there is also a significant amount (494) that declines entertainment media containing such traits.

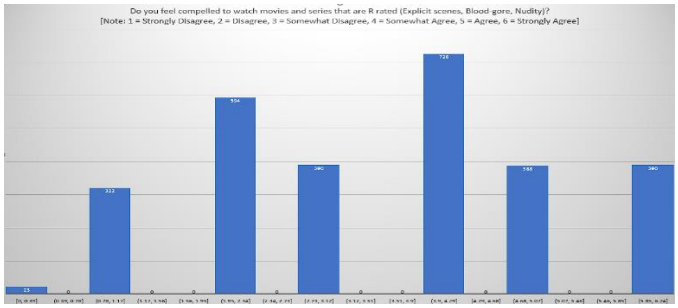

Figure 3.23: Rated R

In the figure 3.23, we find that majority of the audience (726) would appreciate R-Rated scenes in their entertainment media but a good number (594) also denies that it is necessary for an entertainment media to be successful.

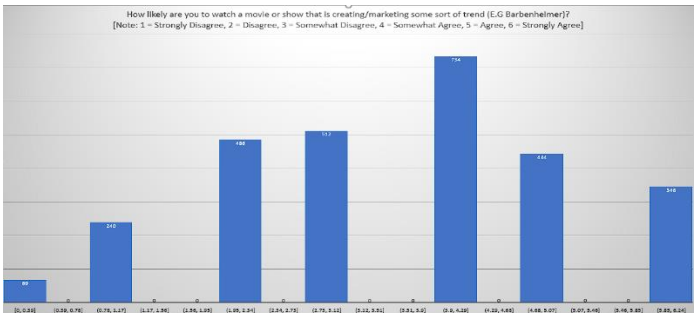

Figure 3.24: Unique Marketing Trend

In the figure 3.24, we see that majority of the audiences (734) somewhat agrees that an unique marketing approach will likely intrigue them to invest on an upcoming entertainment media project.

3.4 Proposed Methodology

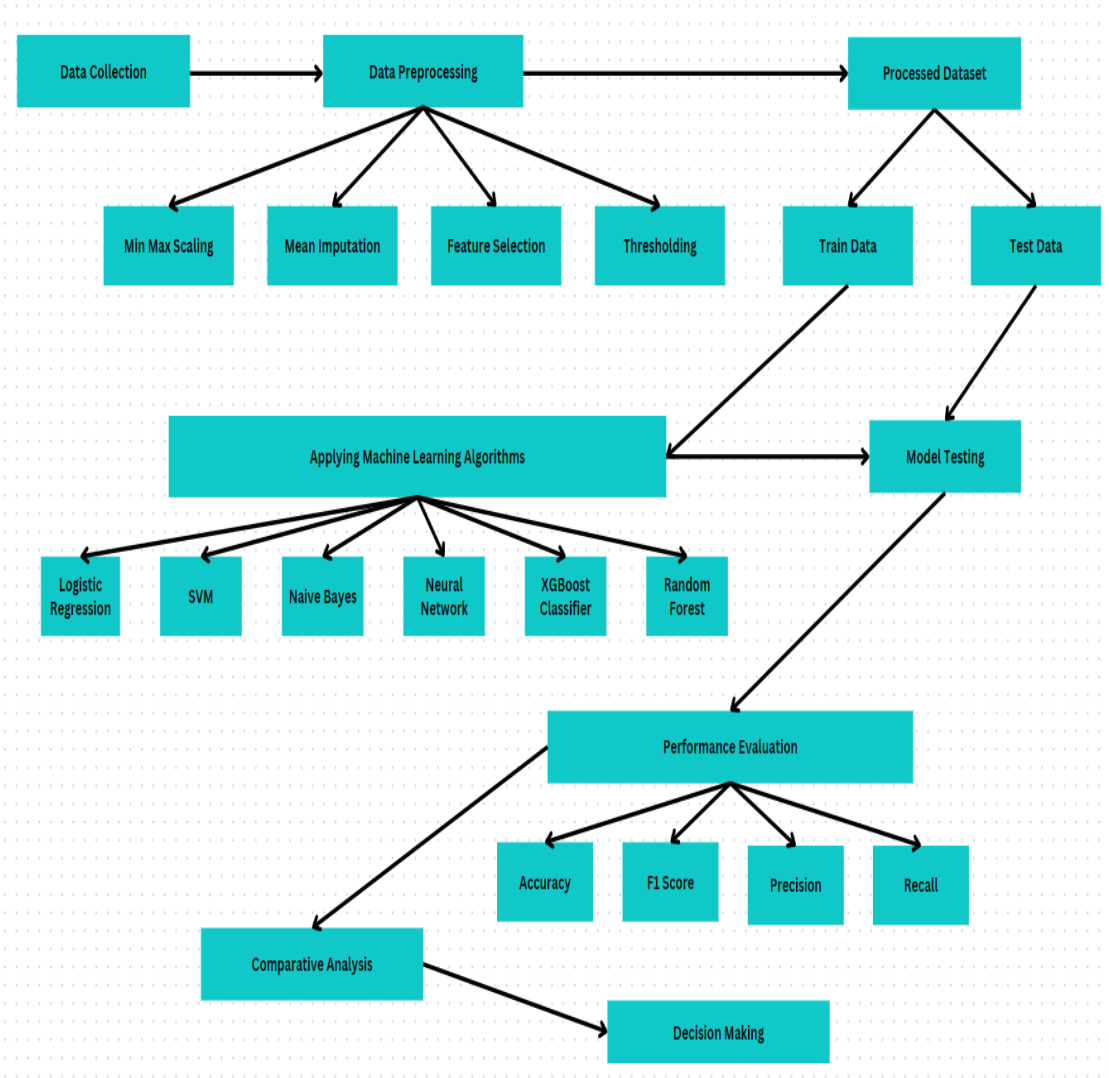

Figure 3.25: Proposed Methodology

The methodology for this entertainment media success prediction project outlines the foundational elements necessary to construct a predictive model capable of accurately forecasting media performance. This comprehensive analysis encompasses various aspects including project objectives, data and feature requirements, algorithm and model specifications, technical needs, user interface considerations, compliance, ethics, and provisions for scalability and maintenance. Key objectives include the development of a machine learning model to predict movie success, identification of critical success predictors, and model optimization for diverse datasets. Data requirements focus on sourcing reliable data, specifying data types, and determining the necessary volume for effective model training and validation. Feature requirements emphasize the importance of both pre-release and post-release attributes in influencing movie success. Algorithm and model requirements entail the selection of suitable machine learning

algorithms, such as Random Forest, SVM or Naïve Bayes with the establishment of performance metrics like accuracy and precision for model evaluation. Technical requirements specify the necessary software, tools, and computational resources, while user interface considerations aim to ensure accessibility and comprehensibility for non-technical users. Compliance and ethics address data privacy and model transparency, ensuring ethical data usage and a clear understanding of predictive mechanisms. Lastly, scalability and maintenance considerations include plans for regular model updates and infrastructure scalability to accommodate growing data volumes and user demands. By meticulously addressing these requirements, the project aspires to deliver a robust, user-friendly predictive model that effectively serves the film industry's needs, aiding stakeholders in making informed decisions based on predictive insights.

Data Pre-Processing

The data which I have collected required to go through some processes before I can introduce it to my model as it is with any raw dataset. My data was collected through online forms, physical Q/A and in person surveys. This resulted in some missing data as there were certain mishaps and miscommunications, reluctance to answer certain questions and human error. To solve this issue, I was required to perform missing value handling. Due to various data sources, I had to perform handcrafted labeling to minimalize any human error. To improve the performance of the model, I also needed to use Min-Max scaler so that I can get a much more optimal output.

Min-Max Scaling

This technique rescales features to a fixed range, typically between 0 and 1. It works by subtracting the minimum value of the feature and then dividing by the range (i.e., the difference between the maximum and minimum values). The formula for Min-Max scaling is:

$$X\text{scaled} = X - X\text{min} / X\text{max} - X\text{min}$$

where X is the original feature value, Xmin is the minimum value of the feature, and Xmax is the maximum value of the feature.

As my dataset consisted answers ranging from strongly disagree to strongly agree, so in total, I have a range from 0 to 6. This drastically changed the outlook of the dataset but also allowed to increase the performance of the model.

	Gender	Age	celebrity	ratings	diversity	propaganda	plot	prequel	initialreview	stream	...	production_value	intelectual	real
0	Male	21-25	3.0	1.0	3.0	4.0	1.0	5.0	1.0	1.0	...	5.0	5.0	
1	Male	26-30	6.0	3.0	1.0	4.0	4.0	2.0	4.0	1.0	...	2.0	6.0	
2	Male	21-25	5.0	6.0	2.0	5.0	4.0	5.0	5.0	6.0	...	1.0	3.0	
3	Male	21-25	1.0	6.0	4.0	5.0	4.0	5.0	4.0	1.0	...	4.0	4.0	
4	Male	21-25	2.0	5.0	1.0	5.0	4.0	5.0	2.0	6.0	...	1.0	4.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
2827	Male	21-25	3.0	1.0	3.0	4.0	1.0	5.0	1.0	1.0	...	5.0	5.0	
2828	Female	21-25	6.0	4.0	1.0	4.0	3.0	2.0	4.0	4.0	...	5.0	4.0	
2829	Female	21-25	5.0	6.0	1.0	4.0	6.0	4.0	4.0	6.0	...	6.0	3.0	
2830	Male	21-25	2.0	5.0	4.0	1.0	2.0	4.0	3.0	4.0	...	1.0	4.0	
2831	Female	21-25	5.0	6.0	1.0	4.0	6.0	4.0	4.0	6.0	...	6.0	3.0	

Figure 3.26: Before Min-Max Scaling

	celebrity	ratings	diversity	propaganda	plot	prequel	initialreview	stream	cultural	address_history_problem
0	0.500000	0.166667	0.500000	0.666667	0.166667	0.833333	0.166667	0.166667	0.666667	0.500000
1	1.000000	0.500000	0.166667	0.666667	0.666667	0.333333	0.666667	0.166667	0.833333	0.166667
2	0.833333	1.000000	0.333333	0.833333	0.666667	0.833333	0.833333	1.000000	0.833333	0.666667
3	0.166667	1.000000	0.666667	0.833333	0.666667	0.833333	0.666667	0.166667	0.666667	0.666667
4	0.333333	0.833333	0.166667	0.833333	0.666667	0.833333	0.333333	1.000000	0.666667	0.833333
...	...	...	...	...	...	...	...	...	...	...
2827	0.500000	0.166667	0.500000	0.666667	0.166667	0.833333	0.166667	0.166667	0.666667	0.500000
2828	1.000000	0.666667	0.166667	0.666667	0.500000	0.333333	0.666667	0.666667	0.500000	0.500000
2829	0.833333	1.000000	0.166667	0.666667	1.000000	0.666667	0.666667	1.000000	0.833333	0.666667
2830	0.333333	0.833333	0.666667	0.166667	0.333333	0.666667	0.500000	0.666667	0.666667	0.833333
2831	0.833333	1.000000	0.166667	0.666667	1.000000	0.666667	0.666667	1.000000	0.833333	0.666667

Figure 3.27: After Min-Max Scaling

Mean Imputation

Mean Imputation is a commonly used technique in data preprocessing to address missing values within a dataset. As I my data set showcased null values due to human error and responders ignoring questions, it was the ideal option to optimize my dataset. When encountering missing data points, mean imputation involves replacing these missing values with the mean (average) value of the respective feature. This process is straightforward and effective, as it fills in the missing values with a value that represents the central tendency of the feature. The steps involved in Mean Imputation include identifying missing values, calculating the mean of non-missing values for each feature, and replacing the missing values with the calculated mean. Mean Imputation is valued for its simplicity and ease of implementation, making it a popular choice in handling missing data. However, it's important to acknowledge potential limitations, such as the introduction of bias if missing data is not missing at random, or if there are a significant number of missing values. Researchers should consider these factors when deciding on the most appropriate approach for handling missing data in their analyses.

Feature Selection

Next step was to identify unnecessary data and deal with them to ensure optimal

performance of the model. For this, I choose Feature Selection, A straightforward approach that involves dropping columns, or features, from the dataset that are deemed irrelevant or redundant for the analysis at hand. By selectively removing unnecessary columns from the dataset, the dimensionality of the data is reduced, leading to simpler and more efficient models. Through my research and from the analysis of my predecessors, I removed column 'Age' & 'Gender' from the dataset.

Thresholding

Lastly, I had to perform thresholding or binarization on my dataset so that I could create my target column. To set the threshold value I proceeded to use the mean method and placed a condition. If the average threshold value fulfilled the condition, then it will be labeled with 1 which is 'Success' otherwise, it will be labeled 0 which is 'Failure'.

Selected Algorithms

In a research-based work where machine learning or data mining techniques are used, selecting the best algorithm is a vital part because the accuracy of the anticipated outcome is heavily dependent on the use of these chosen algorithms. For this research, I used a total of 6 algorithms. They are:

- Logistic Regression
- Support Vector Machine
- Naïve Bayes
- Neural Network
- XGBoost Classifier
- Random Forest Classifier

Logistic Regression

Logistic Regression a classification problem despite its name. it is a statical technique that is used on binary classification. This algorithm predicts the odds of internal dependency of categorical values. The working process of this algorithm involves replicating the relation between one or more variables which are independent. Also, it predicts the chances of a particular outcome to happen. This algorithm uses the logistic function which is also known as the sigmoid function to contain the predicted output between 0 and 1. This helps in mapping the values that are continuous to the probability score. Moreover, logistic regression uses a linear equation to generate the probability of each binary result. It uses this equation on the input features and then applies a non-linear transformation through this function. The logistic regression equation that is used

for a binary classification problem is given below:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Here,

- P(y=1|x) is the probability, given the input features x, that the instance to class 1.
- β_0, β_1, $\beta_2 \ldots \beta_n$ are the weights allocated to each feature x_1, $x_2 \ldots x_n$
- e is the base of the natural logarithm

$\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n$ is the linear combination that indicates the weighted sum of input features. 1/1+ e-z which is the logistic function converts this sum to a value between 0 and 1. This basically indicated to the probability of positive class. In this equation, β0, β1, β2… βn acts as coefficients. We can learn them from thee training data using optimization algorithms. The overall purpose of using these optimization algorithms is to figure out the coefficients that works best in corresponding the data and also reduces the margin of error in class prediction.

Support Vector Machine (SVM)

Support Vector Machine is a widely used supervised learning algorithm that is used for classification tasks. It calculates the best decision threshold or hyperplane that efficiently and successfully separates different classes within the parameter space of the dataset. The reason for using this algorithm in this research is that it always looks for ways to maximize the margin or the distance between the hyperplane and nearest classes which ultimately improves the model's capability to generate and work with new data. In the case of linear data, support vector machine tends to find a linear hyperplane that separates the predictable classes. But if there is nonlinear data, the algorithm will use kernel functions to convert the data into a higher dimensional space which will open the possibility to discover the nonlinear boundaries. The kernel that is used here helps this algorithm to map the input data into high dimension space without calculating the transformation in reality. For the linearly separable data, the hyperplane equation in an SVM is as follows:

$$\mathbf{w}^{\mathsf{T}}\mathbf{x} - b = 0,$$

Here,

- The weight vector perpendicular to the hyperplane is denoted by w.

- The input features are represented by x.
- Bias is represented by b.

As support vector machine is highly effective in high dimensional spaces, uses different kernel functions for versatility and works best for both linear and non-linear data, this algorithm works well in classification tasks which is a perfect match for this particular research work.

Naïve Bayes

Naive Bayes often provides competitive performance and serves as a baseline model in many applications. The algorithm assumes that features are conditionally independent given the class label, simplifying the estimation of the joint probability distribution. Naive Bayes models are trained by estimating class priors and conditional probabilities from the training data. During prediction, the algorithm calculates the posterior probability of each class given the input features and selects the class with the highest probability. Mathematically, the posterior probability in Naive Bayes can be expressed using Bayes' theorem:

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

Here,

- $P(C_k|x)$ is the posterior probability of class C_k given the input features x.
- $P(C_k)$ is the prior probability of class C_k.
- $P(x_i|C_k)$ is the conditional probability of feature x_i given class C_k.
- $\prod^{n}_{i=1} P(x_i|C_k)$ represents the product of conditional probabilities for all features given class C_k.
- $P(x)$ is the marginal probability of the input features.

In practice, Naive Bayes calculates the posterior probability $P(C_k|x)$ for each class C_k and selects the class with the highest probability as the predicted class for the input features x. This simple yet effective approach makes Naive Bayes particularly well-suited for classification tasks.

Neural Network

A Neural Network is a computational model composed of interconnected nodes organized into layers. Information flows through the network, undergoing nonlinear transformations at each layer, ultimately producing an output. Training a Neural Network involves optimizing its parameters (weights and biases) to minimize a defined

loss function, typically using backpropagation and gradient descent. The architecture of a Neural Network can vary, with common structures including feedforward networks, convolutional networks (CNNs), and recurrent networks (RNNs). Mathematically, the forward pass of a Neural Network can be represented as:

$$z^{(l)} = W^{(l)} a^{(l-1)} + b^{(l)}$$

$$a^{(l)} = \sigma(z^{(l)})$$

Here,

- $z^{(l)}$ is the pre-activation output of layer ll.
- $a^{(l)}$ is the activation output of layer ll.
- $W^{(l)}$ and $b^{(l)}$ are the weight matrix and bias vector of layer ll.
- σ is the activation function.

XGBoost Classifier

XGBoost, short for eXtreme Gradient Boosting, is a gradient boosting algorithm known for its speed, scalability, and state-of-the-art performance in various machine learning tasks. XGBoost sequentially builds a strong predictive model by adding decision trees to an ensemble, with each tree learning from the mistakes of its predecessors. The algorithm optimizes a differentiable loss function using gradient descent, efficiently handling complex nonlinear relationships in the data. Key components of XGBoost include regularization techniques to prevent overfitting and tree pruning to control model complexity. Mathematically, the prediction of an XGBoost model can be expressed as:

$$\hat{f}_{(0)}(x) = \arg\min_{\theta} \sum_{i=1}^{N} L(y_i, \theta).$$

Here,

- f is the final prediction,
- N is the number of trees in the ensemble,
- $L(y_i, \theta_i)$ represents the loss function L applied to the true label y_i and the parameters θ_i of the ith tree in the ensemble.

In practice, XGBoost iteratively adds trees to the ensemble, each minimizing the loss function with respect to the previous ensemble's predictions. This iterative process continues until a specified number of trees is reached or until the improvement in the loss function falls below a threshold. The final prediction f is then obtained by summing the contributions of all trees in the ensemble.

Random Forest

Random Forest works best for both regression and classification operations. It basically generates multiple decision trees in the time of training and provides output on the basis of mean prediction of individual trees. It efficiently operates on huge volume of data. During prediction, each tree's output is aggregated through averaging (for regression) or voting (for classification) to produce the final prediction. This ensemble approach helps to reduce overfitting and increase the model's robustness. The algorithm's performance is further enhanced by introducing randomness in tree construction. Mathematically, the prediction of a Random Forest model can be represented as the average or majority vote of individual tree predictions:

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x')$$

Here,

- f is the final prediction.
- B is the number of trees in the forest.
- $f_b(x')$ represents the prediction of the bth tree for the input data x'.

3.5 Implementation Requirements

There are several prerequisites that must be met in order for the research project "Predictive Analytics in Entertainment Media: A Comprehensive Model for Forecasting Media Success and Application for Movies and Series" to be implemented successfully.

Hardware Requirements:

High-performance computing resources: For this research to perform optimally an access to a computing infrastructure capable of handling the intricate tasks required in image analysis and deep learning, with enough memory and processing capacity.

Graphics Processing Unit (GPU): In order to drastically cut down on the amount of time needed for model construction, a GPU is necessary for speeding up the training of machine learning models.

Software Requirements:

Python programming language: Python offers a flexible and popular framework for executing machine learning algorithms and carrying out data analysis.

Machine learning libraries: Integration of machine learning libraries like Scikit-learn, TensorFlow for pre-processing and augmenting data.

Data Collection:

Diverse and representative dataset: Gathering of an extensive dataset with all the information required to guarantee the generalization and resilience of the models created.

Annotated dataset: The availability of well-labeled annotated data with explicit labels to forecast the performance of entertainment media for the models' training and validation.

Model Training and Evaluation:

Machine learning models: Implementation of machine learning architectures, leveraging pre-trained models such as SVM, Random Forest, Logistic Regression to ensure successful prediction outputs.

Metrics for evaluation: The process of defining and applying suitable assessment criteria, like accuracy, precision, recall, and F1 score, to evaluate the models' effectiveness in predicting the success of entertainment media.

Ethical Considerations:

Audience privacy and data security: Adherence to ethical guidelines andregulations to ensure the privacy and security of patient data used in the research.

Informed consent: If applicable, obtaining informed consent from individualswhose answers are included in the dataset.

Documentation and Reproducibility:

Code documentation: Detailed description of the implementation code to support future cooperation, reproducibility, and transparency.

Version control: By implementing version control tools (like Git) to monitor codebase changes will allow us to uphold an organized development process.
By fulfilling these implementation requirements, the study may be carried out methodically and thoroughly, guaranteeing the accuracy of the findings and advancing the field of machine learning-based entertainment media prediction, including movies and television shows.

CHAPTER FOUR
EXPERIMENTAL RESULT

4.1 Experimental Setup

After completing the data preprocessing part, the most important step of this research approaches which is setting up the experiment. In this part, I used the machine learning algorithms on the preprocessed dataset. To perform this, First I created a simulation-based setup. In this case, I used Google Collab. After that, I called the necessary library functions and tools to complete the setup. After setting up the necessary library functions and tools, I loaded the dataset and move towards the result and analysis of the dataset and model performance. I used the data that I gathered from online and offline sources. I applied a total of 6 machine learning algorithms for analysis purpose. As they are very well knowing machine learning algorithms, almost all of them performed very well on this dataset.

4.2 Experiment Performance Evaluation Techniques

When it comes to evaluate the performance of a model, the concepts that are considered fundamental are Confusion Matrix, Accuracy, Recall, Precision and F1 score. The explanation of each term is as follows.

4.2.1 Confusion Matrix

Confusion matrix is one of the main and fundamental tools for evaluating a model's performance. It is represented in a table that gives us meaningful insight of a model's errors, performance and weaknesses. A confusion matrix has for parts. They are True Positive which is number of correctly predicted positive values, True Negative which is the number of correctly predicted negative values, False Positive which is the number of incorrectly predicted positive values and lastly False Negative which is the incorrectly predicted negative values. The basic structure of a confusion matrix is given below:

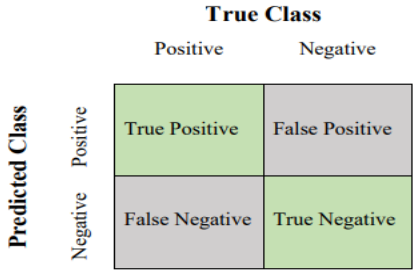

Figure 4.1: Confusion Matrix

4.2.2 Accuracy

Accuracy of the classification model constitutes a single metric. The expected percentage of the model is its accuracy. The proportion of correctly categorized images to total samples is known as accuracy. Accuracy equation:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

4.2.3 Precision

The precision indicates the number of positively predicted cases with accuracy. It is useful to be precise when FP is more important than FN. Mathematically, it's this equation:

$$Precision = \frac{TP}{TP + FP}$$

4.2.4 Recall

The number of correctly categorized positive predicted events is measured by recall or sensitivity. Recall is a useful statistic in cases where FN outperforms FP. A metric for classification accuracy that accounts for both recall and precision is the F1-score. Since recall and precision are harmonic means, a thorough comprehension of these two metrics can be obtained from the F1 score. When Precision and Recall are equal, it performs best. Recall equation:

$$Recall = \frac{TP}{TP + FN}$$

4.2.5 F-1 Score

A recall-precision measure of classification accuracy is the F1 score. The F1-score provides a comprehensive view since it is the harmonic mean of Precision and Recall. When precision and recall are equal, they are ideal. The equation is shown below:

$$F1\ Score = \frac{2 * Precision * Recall}{Recall + Precision}$$

4.3 Experimental Results & Analysis

In this section, I will discuss the performance of each algorithm. I evaluated the performance of every algorithm by using Confusion Matrix, ROC Curve, Accuracy, Precision, Recall, F1 score and Cross Validation score. The brief explanation of each evaluation method for every algorithm is mentioned in this section. After that, I selected the best algorithm based on these evaluation methods.

4.3.1 Logistic Regression

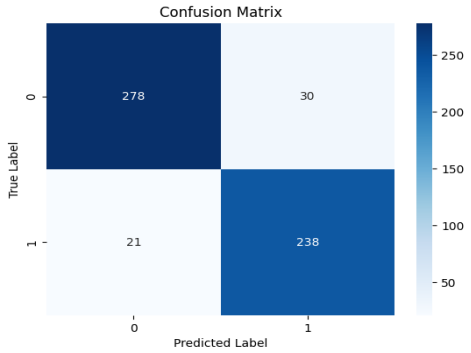

Figure 4.2: Confusion Matrix (Logistic Regression)

In figure 4.2, we see the confusion matrix for the logistic regression model. This model correctly predicted 278 positive classes and 238 negative classes. The number of incorrectly classified classes is low with 30 false positive and 21 false negative values.

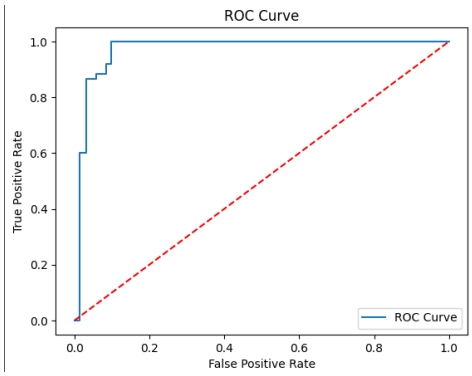

Figure 4.3: ROC Curve (Logistic Regression)

In figure 4.3, we see the ROC curve, here AUC is higher which suggest a good performance by the model. The AUC score is 0.91 which is close to 1.

The model obtained an accuracy of 91.005291% with a precision score of 0.91, recall 0.91 and F1 score of 0.91. This suggests that the model's performance is relatively better than others.

4.3.2 Support Vector Machine (SVM)

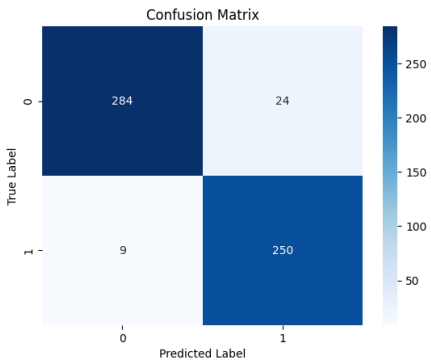

Figure 4.4: Confusion Matrix (SVM)

SVM put on a magnificent performance with this dataset as we can see in figure 4.4, the confusion matrix for the logistic regression model showcases that this model correctly predicted 284 true positive values and 250 true negative values. The number of incorrectly classified classes are extremely low with 24 false positive and 9 false negative values. If we compare this performance with other models in this project, we can see that this model clearly outperformed others when it came to confusion matrix values.

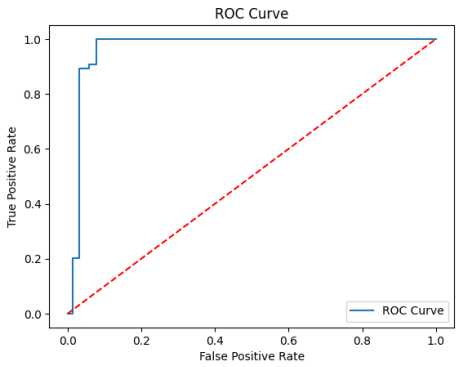

Figure 4.5: ROC Curve (SVM)

As shown in figure 4.5, the ROC curve of the SVM model is also impressive with an overall AUC score of 0.94. This is very close to 1 making the performance of this model stand out more.

The model obtained an accuracy of 94.17989% with a precision score of 0.94, recall 0.94 and F1 score of 0.94. This suggests that the model's performance is the best performing model in these settings.

4.3.3 Naïve Bayes

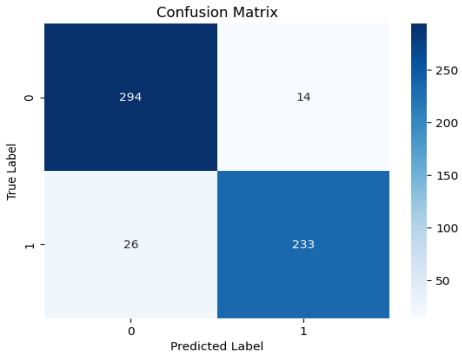

Figure 4.6: Confusion Matrix (Naïve Bayes)

In figure 4.6, we see the confusion matrix for the naïve bayes model. This model correctly predicted 294 positive classes which is the highest positive classes prediction among all the models used in these settings. It also showcased 233 true negative classes. The number of incorrectly classified classes is relatively low with 14 false positive and

26 false negative values. When compared to other models, this model outclassed all the other models except for SVM, making it the second highest performing model.

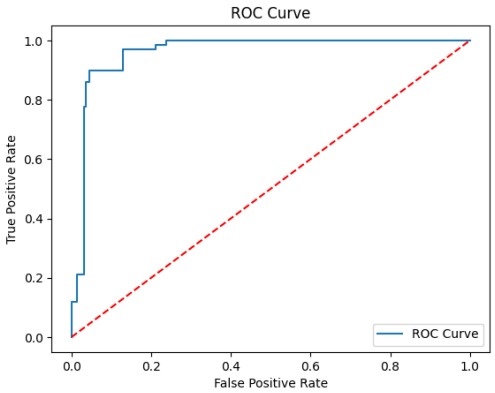

Figure 4.7: ROC Curve (Naïve Bayes)

In figure 4.7, we see the ROC curve of the Naïve Bayes model, here AUC remains higher than most other models which suggest a good performance. The AUC score is 0.93 which is very close to 1.

The model obtained an overall accuracy of 92.45326% with a precision score of 0.93, recall 0.93 and F1 score of 0.93. This suggests that the model's performance is significantly better than others.

4.3.4 Neural Network

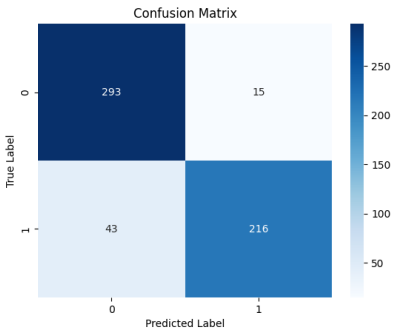

Figure 4.8: Confusion Matrix (Neural Network)

In figure 4.8, we see the confusion matrix for the Neural Network model. This model correctly predicted 293 positive classes and 216 negative classes. The number of incorrectly classified classes isn't as low as the others with 15 false positive and 43 false negative values. When compared to other models, this model didn't performed significantly well, but provided a competitive result.

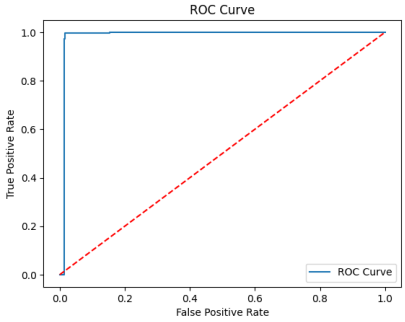

Figure 4.9: ROC Curve (Neural Network)

In figure 4.9, we see the ROC curve of the Neural Network model, here AUC higher than other models which suggests a good performance by the model. The AUC score is 0.95 making it the best score in this project.

The model obtained an accuracy of 89.77072310% with a precision score of 0.90, recall 0.90 and F1 score of 0.90. This suggests that the model's performance is relatively better but not the best.

4.3.5 XGBoost Classifier

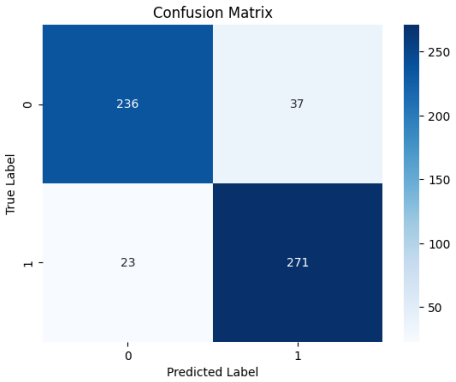

Figure 4.10: Confusion Matrix (XGBoost)

In figure 4.10, we see the confusion matrix for the XGBoost Classifier model. It predicted 236 positive classes and 271 negative classes. On the other side, there is 37 false positive classes and 23 false negative classes. When compared to other models, this model's performance is relatively poor.

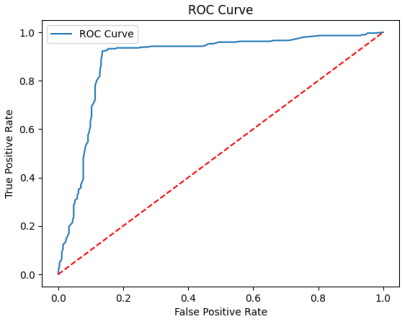

Figure 4.11: ROC Curve (XGBoost)

The ROC curve in figure 4.11, showcases that XGBoost's performance was underwhelmed as the AUC was lower than others. The AUC score of this model is 0.88 which isn't much close to 1.

The model obtained an accuracy of 89.4179% with a precision score of 0.89, recall 0.89 and F1 score of 0.89. This suggests that the model's performance is subpar to the other models and isn't suited for this dataset.

4.3.6 Random Forest

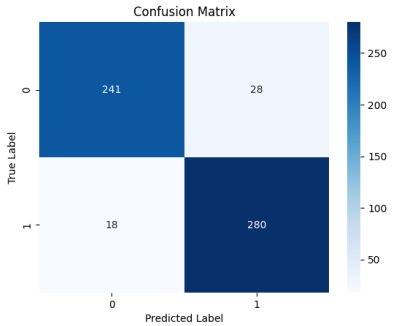

Figure 4.12: Confusion Matrix (Random Forest)

In figure 4.12, we see the confusion matrix for the Random Forest model. This model correctly predicted 241 positive classes and 280 negative classes which is highest among all the other models. The number of incorrectly classified classes is relatively low with 28 false positive and 18 false negative values. When compared to other models, this model performed quite well. ensuring a good true positive and true negative value with low false positive and false negative value.

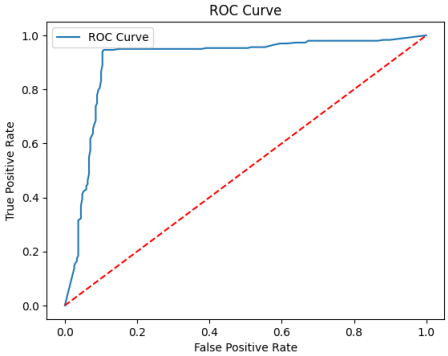

Figure 4.13: ROC Curve (Random Forest)

In figure 4.13, we see the ROC curve of the Random Forest model, here AUC isn't as high as the others suggesting a decrease in performance by the model.

The model obtained an accuracy of 91.8871% with a precision score of 0.92, recall 0.92 and F1 score of 0.92. This suggests that the model's performance is relatively better than others.

4.4 Model Performance Evaluation

The following table showcases the performance evaluation methods I have utilized for this research:

TABLE 4.1: Model Performance Evaluation

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	91.01	0.91	0.91	0.91	0.91
SVM	94.18	0.94	0.94	0.94	0.94
Naïve Bayes	92.45	0.93	0.93	0.93	0.93
Neural Network	89.77	0.90	0.90	0.90	0.94
XGBoost	89.42	0.89	0.89	0.89	0.88
Random Forest	91.89	0.92	0.92	0.92	0.91

To get a more robust idea of each model's performance we can observe the following figure:

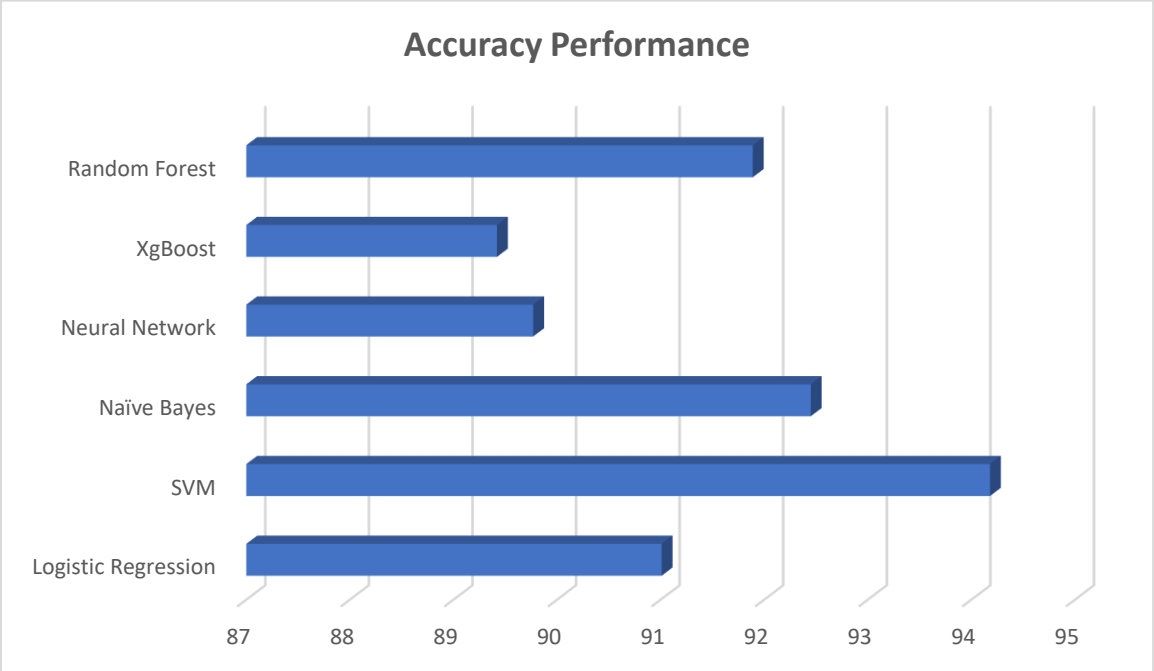

Figure 4.14: Accuracy Performance

4.5 Discussion

This section assesses the efficacy of six distinct machine learning models: Logistic Regression, Support Vector Machine (SVM), Naïve Bayes, Neural Network, XGBoost Classifier, and Random Forest. The evaluation measures employed are Accuracy,

Precision, Recall, F1 Score, and Area Under the Curve (AUC). The Logistic Regression model produced a precision rate of 91.01%. and exhibits consistent and equitable performance across all criteria, achieving a Precision, Recall, and F1 Score of 0.91 for each. The AUC, or Area Under the Curve, is 0.91, this suggests that the model effectively differentiates between the positive and negative classifications. As it is a linear model, Logistic Regression is effective when there is a close to linear relationship between the input features and the target variable. This case is true for the dataset I have assembled for this research. But the performance of this model was inferior to more intricate models due to its potential inability to successfully capture nonlinear patterns in the data. The Support Vector Machine (SVM) model demonstrated superior performance compared to other models, with an accuracy rate of 94.18%. The model achieved good scores on Precision, Recall, and F1 Score, with each metric measuring 0.94. This indicates that the model's predictions are both accurate and reliable. The SVM's AUC is 0.94, which is in close proximity to the optimal score of 1. SVM algorithm operates by identifying the most favorable hyperplane that maximizes the separation between different classes. It exhibits outstanding performance in environments with a large number of dimensions and is highly effective when the decision boundary is intricate. My current dataset fulfills this requirement which is why we can see the most optimal result from SVM. The second best performer in this research is the Naïve Bayes model. It demonstrated an excellent performance, achieving an accuracy rate of 92.45%. This happened due to the fact that Naïve Bayes operates under the assumption that features are conditionally independent of each other, given the class label. However, this assumption is seldom valid when dealing with real-world data. Nevertheless, it exhibits remarkable performance as a result of its straightforwardness and resilience, particularly when dealing with extensive datasets. However, it may not successfully capture intricate correlations between features as other models do, which is why it did not perform as well as the SVM. The Neural Network model attained an accuracy of 89.77%, which is comparatively lower than certain other models, second lowest in this case. Neural networks provide a strong capability to capture intricate patterns in data as a result of their capacity to represent non-linear connections. Nevertheless, training them efficiently necessitates a substantial amount of data and processing resources. The decreased accuracy may be attributed to overfitting a phenomenon in which the model has memorized the training data excessively yet fails to perform well on fresh, unseen data. The lowest performance

in this research is from XGBoost Classifier. Compared to others the results are relatively disappointing, achieving an accuracy of 89.42%. The Precision, Recall, and F1 Score of the model are all 0.89, suggesting steady but not outstanding performance. The AUC score is 0.88, indicating that the model's performance in class differentiation is comparatively less than that of the other models. XGBoost is a technique for ensemble learning that constructs numerous decision trees and merges their results to enhance performance. Nevertheless, the performance of the model may be compromised if the hyperparameters are not properly optimized. The inferior performance in this instance implies that the model may not have been the most suitable for the data, maybe due to inadequate capture of the data distribution. Lastly, The Random Forest model attained a precision rate of 91.89%. The Precision, Recall, and F1 Scores are all 0.92, demonstrating a high level of accuracy and consistency. The high performance can be due to the fact that Random Forest is an ensemble technique that constructs numerous decision trees and combines their predictions to enhance reliability and mitigate overfitting. The outstanding performance of the system can be attributed to its capacity to effectively process huge datasets with higher dimensionality and accurately capture intricate connections between characteristics. Nevertheless, its efficacy may be inferior than that of SVM or Naïve Bayes due to its occasional insensitivity towards smaller, more nuanced patterns in the data.

CHAPTER FIVE
IMPACT ON SOCIETY & ENVIRONMENT

5.1 Impact on Society

The implications of the research on "Predictive Analytics in Entertainment Media: A Comprehensive Model for Forecasting Media Success and Application for Movies and Series" extend far beyond the realm of academia, promising significant impacts on society and business. Entertainment is a key part of a society's growth. With the help of this model, producers and directors can ensure success of their project. It has the potential to revolutionize the landscape of how entertainment media investment works, how the industry accepts new projects. In a general term, this research would allow professionals in the entertainment industry to make informed decisions with greater confidence. This way more projects targeted to provide a message to the people can be created and be ensured of their success.

5.2 Ethical Aspects

The research on "Predictive Analytics in Entertainment Media: A Comprehensive Model for Forecasting Media Success and Application for Movies and Series " is supported by a dedication to using ethical principles in all aspects of its work. When using data collected from a large number of participants, it is important to protect their privacy and confidentiality, which means following legal and ethical requirements. When required, informed consent—a fundamental component of moral research procedures—is duly acquired. Transparent documentation of the research process facilitates responsible knowledge sharing, examination, and reproducibility—an additional way in which the ethical dimension is expressed.

5.3 Sustainability Plan

The sustainability plan for the research on "Predictive Analytics in Entertainment Media: A Comprehensive Model for Forecasting Media Success and Application for Movies and Series" is designed to minimize the chances of failure for new entertainment media such as movies, advertisements, series and such. The focus will be on improving the computational efficiency by using energy-efficient methods and utilizing cloud computing technologies to minimize the energy consumption during model training and deployment. In addition, the research team is dedicated to use environmentally responsible approaches in data management and storage.

CHAPTER 6
CONCLUSION AND FUTURE WORK

6.1 Conclusions

Early and accurate prediction of upcoming entertainment media not only allow the investors and creators an edge but will also provide a much more curated option for the audience. In my research, I have showcased that Machine Learning algorithms such as SVM is highly capable of predicting the success of entertainment medias such as movies and series based on user preference. My model was able to ensure 94% accuracy throughout the study but it requires refinement and more critical data management to reach perfection.

6.2 Implication for Further Study

The findings of this study shares valuable insights on how Machine Learning can help predicting entertainment media's success. Future researches can focus on the aspect of including genre specifications in the mix to further refine upon my work. As majority of the user's data on my research is from a certain age group it is limited to predict what these people might like in a movie or show. This is another aspect that can be looked upon when basing another research based on my studies.

6.3 Limitations

No research project is without its handful of limitation and my work falls under the same jurisdiction. My research didn't include the aspect of different ethnicities and how people belonging to those said ethnicity would react to content that is related to them in a conflicted manner. As mentioned beforehand, majority of the data is collected from a certain age group, diversity in this field will result in much more accurate predictions. There is also room to include more complex models, different parameters to dive deeper in the field of entertainment media success prediction.

REFERENCES

[1] Williams, O.E., Lacasa, L. & Latora, V., 2019. Quantifying and predicting success in show business. Nature Communications, 10, Article number: 2256.

[2] Quader, N., Chaki, D., Gani, M.O. & Ali, M.H., 2017. A Machine Learning Approach to Predict Movie Box-Office Success. IEEE. DOI: 978-1-5386-1150-0

[3] Darapaneni, N., Entoori, S., Vybhav, S.V., Bellarmine, C., Kumar, A., Mondal, K., & Paduri, A.R. Movie Success Prediction Using ML

[4] Bhave, A., Kulkarni, H., Biramane, V., & Kosamkar, P., 2015. Role of Different Factors in Predicting Movie Success. In 2015 International Conference on Pervasive Computing (ICPC), Dept. of Computer Engg., MIT, Pune.

[5] Dhir, R., & Raj, A., 2018. Movie Success Prediction using Machine Learning Algorithms and their Comparison. In 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Department of Computer Science, IIT Patna, Bihar, India

[6] Ericson, J. & Grodman, J., "A Predictor for Movie Success." CS229, Stanford University

[7] Klimmt, C., Roth, C., Vermeulen, I., Vorderer, P., & Roth, F.S., 2012. Forecasting the Experience of Future Entertainment Technology: "Interactive Storytelling" and Media Enjoyment. Games and Culture, 7(3), pp. 187-208. DOI: 10.1177/1555412012451123.

[8] Hofmann-Stölting, C., Clement, M., Wu, S., & Albers, S., 2017. Sales Forecasting of New Entertainment Media Products. Journal of Media Economics, 30(3), pp. 143-171. DOI: 10.1080/08997764.2018.1452746

[9] Masrury, R.A., Saputra, M.A.A., Alamsyah, A., & Primantari, M.A.S., 2019. A Comparative Study of Hollywood Movie Successfulness Prediction Model. In Proceedings of the 7th International Conference on Information and Communication, School of Economics and Business, Telkom University, Indonesia

[10] Quader, N., Gani, M.O., & Chaki, D., 2017. Performance Evaluation of Seven Machine Learning Classification Techniques for Movie Box Office Success Prediction. In Proceedings of the 3rd International Conference on Electrical Information and Communication Technology (EICT), 7-9 December 2017, Khulna, Bangladesh

[11] Qaseem, D.M., Ali, N., Akram, W., Ullah, A., & Polat, K., 2022. Movie Success-Rate Prediction System through Optimal Sentiment Analysis. Journal of the Institute of Electronics and Computer, 4, pp. 15-33. DOI: 10.33969/JIEC.2022.41002

[12] 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), pp. 874-903. DOI: 10.1080/07421222.2016.1243969

[13] Verma, H. & Verma, G., 2019. Prediction Model for Bollywood Movie Success: A Comparative Analysis of Performance of Supervised Machine Learning Algorithms

[14] Kanitkar, A., 2018. Bollywood Movie Success Prediction Using Machine Learning Algorithms. In IEEE Third International Conference on Circuits, Control, Communication and Computing. Pune, India

[15] Kumar, S., Mehta, A., & Pal, J. (2019). Movie Success Prediction Using Data Mining for Data Mining and Business Intelligence (ITA5007) of Master of Computer Application. School of Information Technology and Engineering

[16] Lee, K., Park, J., Kim, I., & Choi, Y. (2016). Predicting movie success with machine learning techniques: ways to improve accuracy. *Information Systems Frontiers*

[17] Khandelwal, R., & Virwani, H. (2019). Comparative analysis for prediction of success of Bollywood movies. In *Proceedings of the International Conference on Sustainable Computing in Science, Technology & Management (SUSCOM-2019)*

[18] Chakraborty, P., Rahman, M. Z., & Rahman, S. (2019). Movie success prediction using historical and current data mining. *International Journal of Computer Applications*, *178*(47)

[19] Shahrivar, P., & Jernbäcker, C. (2017). Predicting movie success using machine learning techniques

[20] Agarwal, M., Venugopal, S., Kashyap, R., & Bharathi, R. (2021). A comprehensive study on various statistical techniques for prediction of movie success

[21] Subramaniyaswamy, V., Vaibhav, V. M., Prasad, V. R., & Logesh, R. (2017). Predicting movie box office success using multiple regression and SVM. In Proceedings of the International Conference on Intelligent Sustainable Systems

[22] Mestya´n, M., Yasseri, T., & Kerte´sz, J. (2013). Early prediction of movie box office success based on Wikipedia activity big data. *PLoS ONE*, 8(8), e71226

[23] Ankit, Lakshmi M., K. Aditya Shastry, Aditya Sandilya, & Rahul Shekhar. (Year). A comparative analysis of machine learning approaches for movie success prediction. In *Proceedings of the Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)*. IEEE

[24] Sivakumar, P., Rajeswaren, V. P., Abishankar, K., Ekanayake, E.M.U.W.J.B., & Mehendran, Y. (2020). Movie success and rating prediction using data mining algorithms. *Journal of Information Systems & Information Technology (JISIT)*, 5(2), 72-80.

[25] Shahid, M. H., & Islam, M. A. (2023). Investigation of time series-based genre popularity features for box office success prediction. *PeerJ Computer Science*, 1603

[26] Sadashiv, S., Seli, S., & Sreram, S. (2021). Movie Success Prediction Using Machine Learning. *International Research Journal of Modernization in Engineering Technology and Science*, 3(7). Impact Factor- 5.354

[27] Meenakshi, K., et al. (2018). A Data Mining Technique for Analyzing and Predicting the Success of Movie. *Journal of Physics: Conference Series*, 1000, 012010.

[28] Praxis. (n.d.). Movie Analytics 2.0: Predicting the next 100 crore club movie.

Predictive Analytics in Entertainment Media A Comprehensive Model for Forecasting Media Success AND Application for Movies and Series.pdf

ORIGINALITY REPORT

23%	18%	10%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	9%
2	Submitted to Daffodil International University Student Paper	1%
3	www.researchgate.net Internet Source	1%
4	"Forthcoming Networks and Sustainability in the AIoT Era", Springer Science and Business Media LLC, 2024 Publication	<1%
5	Mohammed Awad, Salam Fraihat. "Recursive Feature Elimination with Cross-Validation with Decision Tree: Feature Selection Method for Machine Learning-Based Intrusion Detection Systems", Journal of Sensor and Actuator Networks, 2023 Publication	<1%