# CINEMATIC SUCCESS PATTERNS

*by*


**D SHIVANESH 22MIS1146**
**YASHWANTH S 22MIS1188**
**HARISH D 22MIS1218**
**ILANGAVIYAN 22MIS1223**


*under the guidance of*

**Dr. Pattabiraman V.**

in partial fulfillment of the course

**SWE2011 - Big Data Analytics**


VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)


**School of Computer Science and Engineering**

**Vellore Institute of Technology**

**Chennai - 600127**


**November 2024**

# BONAFIDE CERTIFICATE

Certified that this project report entitled **"Cinematic Success Patterns"** is a bonafide work of **D SHIVANESH - 22MIS1146, YASHWANTH - 22MIS1188, HARISH D - 22MIS1218** , **ILANGAVIYAN U - 22MIS1223** who carried-out the Project work under my supervision and guidance

**for SWE2011 - Big Data Analytics**


## Dr. Pattabiraman V.

Professor

School of Computer Science and Engineering (SCOPE),

VIT University, Chennai

Chennai – 600 127.

## TABLE OF CONTENTS

# 1.ABSTRACT

Spawning from the very source of entertainment, namely, growing cinema, the need to know and understand the patterns of success arose to understand what shapes the film industry. This mini-project revolves around analyzing movie data in the period 2014-2023 using Big Data Analytics with advanced Machine Learning (ML) and Deep Learning (DL) techniques to unearth the patterns that eventually lead to the success of a movie. The analysis is divided into four key dimensions Genre-wise, Director-wise, Release Date-wise (Festivals), and Runtime-wise success. We used two datasets; the Movies dataset contained attributes such as runtime, genres, directors, release dates, budget, box office collections, and verdicts; the Festival dataset included date and festival details.

For Genre-wise success, we experimented on ML models such as XGBoost, Random Forest, Logistic Regression (with SMOTE), SVM, Decision Tree, and Gradient Boosting. Here, though the DL model ANN was fairing well in an ensemble setting along with the results of XGBoost, the more precise and accurate model that came out for the genre-specific setting was Logistic Regression, which pointed out how sometimes simpler models can beat the complex ones.

The trend with respect to director-wise success is observed to be similar, wherein logistic regression has outperformed other ML models and the ensembled DL models. This suggests that directors play a core role in deciding the performance of movies at the box office by varying their choice for a film, which has been established through strict model comparison. Analysing the same in terms of success on Release Date-wise, especially around Indian festivals, reveals something distinct. It can be perceived that Logistic Regression is identified as the best model among ML techniques. However, ensembling ANN with XGBoost shows enhanced results, underlining the impact of strategic release timing during festive seasons on a movie's success.

Lastly, for Runtime-wise success, both Random Forest Classifier and DL ANN models obtained 100% accuracy, which proved runtime as a significant factor with predictable patterns of audience acceptance. Additionally, if we ensemble Random Forest with XGBoost, we again experience perfect accuracy, thus verifying runtime as a decisive characteristic.

In this regard, the study demonstrates that, although DL techniques like ANN have their potential, models like Logistic Regression and Random Forest have more efficiency in most of the cases. Big Data tools and techniques used in this project will afford useful information regarding the relationship between the cinema attributes and success. The such findings will help filmmakers, producers, and distributors to take effective data-driven decisions to maximize movie success.

This project exemplifies the application of Big Data Analytics to real-world problems, using state-of-the-art ML and DL models to generate actionable insights in the entertainment industry.

## 2. Scope

Big Data Analytics project on providing actionable insights on factors that might determine the success of movies in the entertainment industry: analysis for movies from 2000-2024 in multiple languages, but only those aspects that can be determined to contribute to being a critical success, like genre, directors, date of release, especially Indian festivals, and runtime.This project uses advanced ML and DL models to develop an indispensable analysis regarding the patterns and trends of cinema performance at present, thus allowing stakeholders to take better decisions with help from data.
This mainly entails the specific scope of the project, in particular:
**Genres:** it is just like which genres are performing and getting an audience penetration analysis.
**Analysis of Directors**: whose the reliable performers and what characteristic makes a specific movie a success.
**Release Date-Wise Analysis:** The impact of strategic release times along with the Indian festivals on the box office.
**Runtime-Wise Analysis**: Exploring how the running time affects the reception and success of a film before the audience.
The project incorporates large datasets and applies sophisticated ML and DL techniques to gain meaningful insights. It incorporates techniques such as SMOTE to overcome issues of data imbalance and follows up with techniques such as ensembling to enhance the model performance. This scope places the project as a practical guide for filmmakers, producers, and marketers to maximize commercial success from their films.

## 3. OBJECTIVES

Techniques related to advanced Machine Learning and Deep Learning will be used to analyze movie success patterns as the primary objective of this Big Data Analytics project. Below, the specific objectives of this research are presented in detail.

- **Genre-Wise Success Analysis:**

To check the box office performance of different genres of movie, such as action, drama, comedy, and romance.
Analysis of the preferences of the audiences and revenue success along with various genres and over time for trend identification.

- **Director-Wise Success Analysis**

- To measure the potential of directors that would make a movie successful by assessing their track and consistency records.

- Shared characteristics of films directed by the acclaimed film-makers that would be aiding in determining the director's role in impacting public as well as critics' response.

◆ **Release Date-Wise Success Analysis:**

- To determine the strategic relevance of movie releases from the perspective of major Indian festivals like Diwali, Eid, and Christmas.
- To find which festivals have higher collections in the box office and why the time period matters to the maximum population attending.

◆ **Success Analysis by Runtime-**

- To see what runtime impacts the audience's satisfaction and acceptance more.
- To find out which range of runtime allows for better box office performance and overall involvement in terms of viewers. Model Comparison
- To evaluate and compare ML models-XGBoost, Random Forest, Logistic Regression, SVM, Decision Tree, and Gradient Boosting for movie success prediction .
- To evaluate DL models like ANN models and ensemble options with ML models, to produce a better prognosis .

◆ **Result and Visualization:**

- Visualize the results in bar charts, line graphs, scatter plots, and heatmaps in order to make the outcome presentable.
- To track trends, success patterns, and a big deal influences that would make a big difference in film performance.

◆ **Insight Generation:**

As a tool to provide operational insights towards planning and executing their projects more strategically among filmmakers, producers, and distributors. we will give an illustration in detail of how the big data technology would disrupt the entertainment industry decision-making environment by unearthing valuable success patterns.

# 4. Introduction

Huge contributions are given by the entertainment industry, particularly the cinema, in terms of cultural narratives and serving as escapism for audiences around the world. This reality speaks to the importance of driving factors behind a movie's success, especially at a time when data is increasing exponentially and the permutations in what the audience prefers are more complex than ever. Thus, Big Data Analytics has enabled stakeholders to draw inferences from enormous amounts of structured and unstructured data. This tool has the strength to reveal trends and patterns within a movie's success. The current project discusses the use of Big Data Analytics for forecasting and understanding the reasons

behind the success of a specific movie, incorporating all the dimensions such as genre, director, release timing, and runtime.

This project is based on two major datasets: a movie dataset having attributes such as runtime, genres, directors, release dates, budget, box office collections, and verdicts; and a festival dataset detailing major Indian festivals. In sum, this provides a comprehensive basis for analyses we might pursue; we will therefore able to explore success patterns over time, across languages, and during key periods like festivals. It is an ideal setting for such an analysis due to the highly diversified nature of Indian cinema landscape in terms of audience preferences and cultural nuances. The insights gained through this exercise will thus be germane to other global markets too.

Sophisticated ML and DL techniques constitute an important part of this research. On the same lines, the project uses some models for Logistic Regression, Random Forest, XGBoost, Decision Tree, SVM, and Gradient Boosting. It tries the predictive power of these algorithms. It also compares different kinds of Deep Learning models with some Artificial Neural Networks. The performances of models are further analyzed to figure out how well they can be used in the case of cinema success analysis. These models compare in both isolate and ensemble setups.

The scope of analysis covers four critical dimensions. The Genre-wise analysis focuses on the fact how the tastes of different audiences for specific genres of movies impact box office collections. It highlights awareness of how the creative vision and the leaders of a movie would influence its success with a director. The Release Date-wise analysis looks into strategic releases-most notably around Indian festivals-so it can spot some specific times that may increase crowd attendance and overall collection. The Runtime-wise analysis investigates movie length effects on the audience to establish golden ranges of runtime for movies.

This paper presents actionable recommendations for filmmakers, producers and distributors in terms of using data to maximize the commercial and critical success of a movie. For instance, some genres, the optimal release date and the preferred duration would probably support content creation and marketing strategy. The project demonstrates the applicability of Big Data Analytics in the entertainment industry and illustrates predictive models that transform decision-making processes.
In such an age where tastes among the audiences change at a breakneck speed, it unveils something of considerable importance about data-driven approaches within the film industry through this research. This project combines the kind of power ushered in by Big Data tools with the greater strengths of ML and DL. Hence, not only does it tell us of its ability to connect factors that will further movie success but also tints the canvas for what can be fruitful research in the future.

# 5. Data Description

We didn't get any suitable Dataset online . So we divided the task for data collection and gathered around 800+ movies .There are two datasets taken into consideration in this project: one that collectively forms a strong base in which the factors governing movie success can be studied. This dataset, Movies Dataset, is located in the file Final.xlsx and encompasses information about movies released between 2014 and 2023. It includes attributes such as title, runtime, release year, genres, release date, directors, budget, box office revenue, and verdict (hit, flop, blockbuster). The variables will also be looked at with a more critical eye so that an examination of the performance patterns of movies can be made according to genre, the director's factor, and release timing on productivity with regard to runtime relationships with audience reception. The economic attributes of budget and box office will allow identification of economic factors of success.
The Festival Dataset in festivals.xlsx includes information about the significant Indian festivals: name of the festival and corresponding date. This dataset forms the backbone in determining the strategic relevance of movie releases within the periods of festivals and how the Release timing affects box office performance.

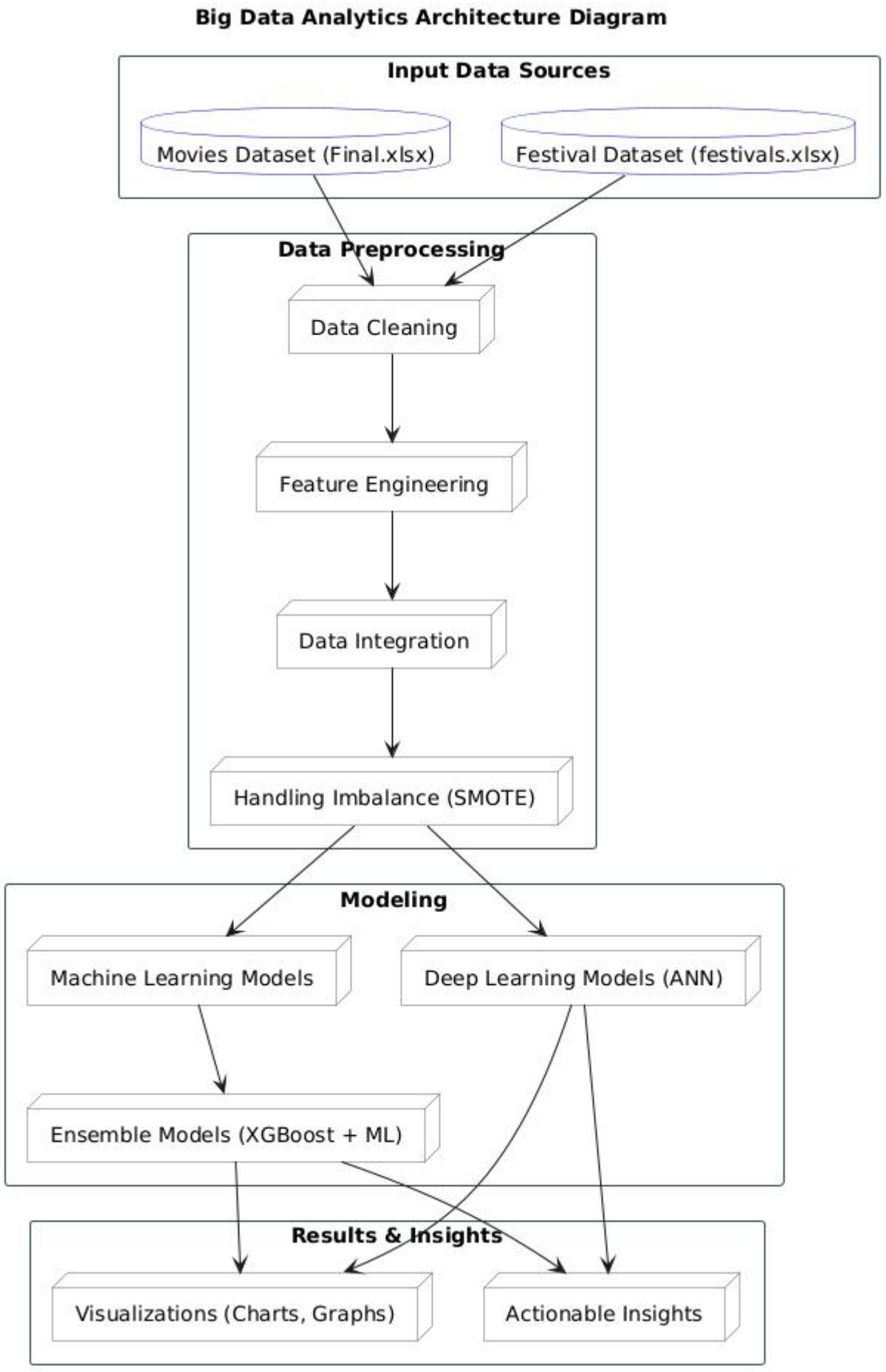
**Dataset** :

| S.I | Original Title | Runtime (mins | Year | Genres | Release Date | Directors | Budget | Box office | Verdict |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Vada Chennai | 164 | 2018 | Action, Crime, Drama, Thriller | 2018-10-17 | Vetrimaaran | ₹25.00 crores | ₹50.00 crores | Super Hit |
| 2 | Pariyerum Perumal | 154 | 2018 | Drama | 2018-09-28 | Mari Selvaraj | ₹4.00 crores | ₹19.00 crores | Blockbuster |
| 3 | Ratsasan | 146 | 2018 | Action, Crime, Drama, Mystery, Thriller | 2018-10-05 | Ram Kumar | ₹8.00 crores | ₹50.00 crores | Blockbuster |
| 4 | Mahanati | 177 | 2018 | Biography, Drama | 2018-05-08 | Nag Ashwin | ₹25.00 crores | ₹83.00 crores | Blockbuster |
| 5 | Merku Thodarchi Malai | 122 | 2018 | Drama | 2018-09-06 | Leninbharati | ₹10.00 crores | ₹11.00 crores | Hit |
| 6 | 96 | 158 | 2018 | Drama, Romance | 2018-10-03 | C. Prem Kumar | ₹15.00 crores | ₹60.00 crores | Blockbuster |
| 7 | Kolamavu Kokila | 140 | 2018 | Comedy, Crime, Drama | 2018-08-16 | Nelson Dilipkumar | ₹8.00 crores | ₹73.00 crores | Blockbuster |
| 8 | Kaala | 162 | 2018 | Action, Drama | 2018-06-06 | Pa. Ranjith | ₹140.00 crores | ₹180.00 crores | Hit |
| 9 | Kadaikutty Singam | 149 | 2018 | Action, Drama | 2018-07-12 | Pandiraj | ₹25.00 crores | ₹50.00 crores | Super Hit |
| 10 | Irumbu Thirai | 160 | 2018 | Action, Crime, Thriller | 2018-05-10 | P.S. Mithran | ₹20.00 crores | ₹61.00 crores | Blockbuster |
| 11 | Sila Samayangalil | 110 | 2016 | Drama | 2018-04-30 | Priyadarshan | ₹5.00 crores | ₹2.00 crores | Disaster |
| 12 | Chekka Chivantha Vaanam | 146 | 2018 | Action, Crime, Drama, Thriller | 2018-09-26 | Mani Ratnam | ₹40.00 crores | ₹95.00 crores | Blockbuster |
| 13 | Imaikkaa Nodigal | 170 | 2018 | Action, Crime, Drama, Thriller | 2018-08-29 | R. Ajay Gnanamuthu | ₹20.00 crores | ₹52.00 crores | Blockbuster |
| 14 | Iravukku Aayiram Kangal | 122 | 2018 | Action, Thriller | 2018-05-11 | Mu. Maran | ₹6.00 crores | ₹17.00 crores | Blockbuster |
| 15 | Mercury | 106 | 2018 | Horror, Sci-Fi, Thriller | 2018-04-13 | Karthik Subbaraj | ₹15.00 crores | ₹25.00 crores | Super Hit |
| 16 | Adanga Maru | 145 | 2018 | Action, Crime, Drama, Thriller | 2018-12-21 | Karthik Thangavel | ₹10.00 crores | ₹30.00 crores | Blockbuster |
| 17 | Kanaa | 145 | 2018 | Drama, Sport | 2018-12-21 | Arunraja Kamaraj | ₹10.00 crores | ₹28.00 crores | Blockbuster |
| 18 | Seethakaathi | 173 | 2018 | Drama | 2018-12-20 | Balaji Tharaneetharan | ₹30.00 crores | ₹25.75 crores | Average |
| 19 | Kidugu | 107 | 2023 | Crime | 2023-03-03 | Veera Murugan | ₹20.00 crores | ₹10.00 crores | Flop |
| 20 | Varisu | 169 | 2023 | Action, Comedy, Drama, Romance | 2023-01-11 | Vamshi Paidipally | ₹200.00 crores | ₹310.00 crores | Super Hit |
| 21 | Michael | 152 | 2023 | Action, Crime, Drama, Thriller | 2023-02-03 | Ranjit Jeyakodi | ₹35.00 crores | ₹25.00 crores | Flop |
| 22 | Thalaikoothal | 140 | 2023 | Drama | 2023-02-03 | Jayaprakash Radhakris | ₹4.00 crores | ₹2.00 crores | Flop |

**Festival :**

| | A | B |
|---|---|---|
| 1 | date | festival |
| 2 | 2014-01-01 | New Year |
| 3 | 2014-01-13 | Lohri |
| 4 | 2014-01-14 | Pongal |
| 5 | 2014-01-23 | Subhas Chandra Bose Jayanti |
| 6 | 2014-01-26 | Republic Day |
| 7 | 2014-02-04 | Basant Panchmi, Saraswati Puja |
| 8 | 2014-02-27 | Mahashivratri |
| 9 | 2014-03-16 | Holika Dahan |
| 10 | 2014-03-17 | Holi |
| 11 | 2014-03-31 | Chaitra Navratri, Ugadi, Gudi Padwa |
| 12 | 2014-04-01 | Bank's Holiday, Cheti Chand |
| 13 | 2014-04-08 | Ram Navami |
| 14 | 2014-04-09 | Chaitra Navratri Parana |
| 15 | 2014-04-14 | Tamil New Year |
| 16 | 2014-04-15 | Hanuman Jayanti |
| 17 | 2014-05-02 | Akshaya Tritiya |
| 18 | 2014-06-29 | Jagannath Rath Yatra |
| 19 | 2014-07-08 | Ashadhi Ekadashi |
| 20 | 2014-07-12 | Guru Purnima |
| 21 | 2014-07-30 | Hariyali Teej |
| 22 | 2014-08-01 | Nag Panchami |
| 23 | 2014-08-10 | Raksha Bandhan |

# 6. Architecture



# 7. Proposed Methodology

The methodology proposed structures the way to discovery success patterns in the movie industry through Big Data Analytics. Each step of this methodology allows for a logical flow of tasks to ensure appropriate use of data as well as advanced analytical models.

- **Data Collection**

The first step involves gathering two key datasets:

**Movies Dataset**: this has details about runtime, genres, budget, box office revenue, and verdict about the movie in the range from 2000 to 2024. This dataset forms the core of the analysis because it contains all the attributes to study success patterns.
**The Festival Dataset:** contains a list of Indian festivals along with their dates. This dataset is vital in order to examine how release time during the festival affected it.
These datasets together, therefore, consider intrinsic factors (e.g., genre, director) and extrinsic factors (e.g., festival timing) affecting a film's success.

- **Data Preprocessing**

Data preparation is one of the most critical steps, which helps ensure high-quality input for analysis. Some key activities include:

**Cleaning**: Removing duplicates, handling missing values, and standardizing formats. For example, normalizing date formats or consistent budget and box-office units.
**Feature Engineering**: Get other derived features such as whether any of the releases of a movie overlap with a festival or calculate the box office performance ratios.
**Integration**: Join Movies and Festival datasets into one dataset.
**SMOTE:** Balance the verdict labels related to flops versus blockbusters, using SMOTE which introduces synthetic samples for underrepresented classes to avoid bias in models' prediction.

- **Exploratory Data Analysis**

EDA provides insight into the data through visualization and summary statistics.
**Genre Analysis:** Analyzing the genres in which this happens more frequently than for others.
**Director Analysis:** Analyzing the trend of directors having a success rate more frequently than others.
**Festival Analysis**: Representing how the revenue picks up during festival seasons to validate the strategic importance of date selection.
**Runtime Analysis**: Analyzing the ideal length of the movie to find out the ideal runtime from the verdict distribution.
EDA gives preliminary insights that outline feature selection and model building.

## ◆ Model Selection and Implementation

This step employs sophisticated algorithms that forecast the success of a movie:

**Machine Learning Models**: Models comprising Logistic Regression, Random Forest, XGBoost, and Gradient Boosting are applied on the success patterns. These models are pretty strong with respect to handling tabular data and facilitate interpretable outcomes.
Deep Learning Models: Artificial Neural Networks have been used to analyze any nonlinear relationships that may appear in the data. ANN is specifically useful for the treatment of complicated interactions between the set of attributes.
**Ensemble Modeling:** Utilizing XGBoost in combination with other ML models is used to enhance predictive accuracy, since all these algorithms benefit from multiple strengths. Instead of ANN, the ensemble procedure is only used between ML models, and this follows a stepwise iterative process to refine further.

## ◆ Evaluation of the Model

All of the models are assessed based on standard performance metrics:
**Accuracy:**
Accuracy points towards how well the predictions were done by and large.
**Precision, Recall, and F1-Score:**
They define the extent to which a model performed on the imbalance class.
**AUC-ROC:**
It computes the balance between the true positive and false positive.
Comparison cross ML, DL, and ensemble approaches determine the best model for every dimension of the data, whether it be genres, directors, festivals or runtimes.

## ◆ Visualization of Results

The results have to be presented in an efficient way. Such of important visualizations include:
Bar Charts  Compare success rates across genres and directors
Line Graphs. Highlight trends over time: how the popularity of genres has changed over the years
Scatter Plots and Heatmaps. Identify correlations between budget, box office collections, and verdicts of success.
Visualization provides stakeholders with a subconscious sense for the answer.

## ◆ Insight Generation

Insights accrued from the analysis will feed into strategies in the industry:
High-performing genres to dedicate as much production as possible.
Best festivals times to encourage collections.
Ideal time lengths for reaching the maximum satisfying length.
The insights provided, these are actionable recommendations for filmmaking.

◆ **Reporting**

Final compiling of all findings, methodologies, and results into a structured project report
Detailed explanations with analysis, model performance comparison, and visualization are also included in the report.
The solutions and recommendations that are proposed to the stakeholders by providing them with an opportunity to effectively utilize knowledge.

# 8. Choosing Best Model

## ● Genre :

**Models Tested**

1. Logistic Regression (with SMOTE for handling imbalanced data)
2. Random Forest
3. XGBoost
4. SVM
5. Decision Tree
6. Gradient Boosting

**Key Findings**

- **Best ML Model**: Logistic Regression (with SMOTE).
    - Achieved the highest predictive performance with balanced accuracy, effectively handling imbalanced classes in the dataset.
- **Ensemble Comparison**: Logistic Regression was ensembled with XGBoost to evaluate potential improvements.
    - Ensembling provided minor performance gains, improving accuracy slightly in some cases. However, the marginal improvement did not outweigh the simplicity and interpretability of Logistic Regression.

| Model | Accuracy | Remarks |
|---|---|---|
| Logistic Regression | Best | Effective with SMOTE, balanced accuracy, and computational efficiency |
| Ensembling (Logistic Regression + XGBoost) | Slightly Better | Minor improvement, not significant enough to justify added complexity |
| Random Forest | Good | Performed well but fell short of Logistic Regression |
| XGBoost | Good | Comparable to Random Forest, better when ensembled |
| SVM | Moderate | Underperformed for imbalanced genre data |
| Decision Tree | Moderate | Simpler model but less accurate than Logistic Regression |
| Gradient Boosting | Good | Performed well but not as effective as Logistic Regression |

Of course, in a prediction of genre-based success, **Logistic Regression** turned out to be the best. It provides both acceptable interpretability and accuracy in terms of prediction. Ensemble models and tree-based algorithms show considerable promise, but their complexity is overkill for the given dataset. In all cases, it appears that when the data structure and relationship is such that simple models are all that can be used, they outperform more complex models.

- **Directors:**

**Models Tested**

1. Logistic Regression (with SMOTE for handling imbalanced data)
2. Random Forest
3. XGBoost
4. SVM
5. Decision Tree
6. Gradient Boosting

**Key Findings**

- **Best ML Model**: Logistic Regression (with SMOTE).
    - Demonstrated the best predictive performance, effectively handling imbalanced data related to director-specific success patterns.

- **Ensemble Comparison**: Logistic Regression was ensembled with XGBoost to explore performance improvements.
    - The ensemble model offered slight accuracy gains in specific cases. However, the marginal benefit did not outweigh the simplicity and interpretability of standalone Logistic Regression.

| Model | Accuracy | Remarks |
|---|---|---|
| Logistic Regression | Best | Best-performing model with SMOTE, offering simplicity and balanced accuracy |
| Ensembling (Logistic Regression + XGBoost) | Slightly Better | Marginal accuracy improvement, but added complexity |
| Random Forest | Good | Strong performance but not as accurate as Logistic Regression |
| XGBoost | Good | Comparable to Random Forest, better when ensembled |
| SVM | Moderate | Underperformed compared to Logistic Regression |
| Decision Tree | Moderate | Simpler model but less effective than Logistic Regression |
| Gradient Boosting | Good | Strong performance but fell short of Logistic Regression |

For director-based success prediction, **Logistic Regression** is the optimal choice due to its high accuracy and interpretability. Tree-based models like Random Forest and Gradient Boosting performed well and could be used as secondary options for scenarios requiring non-linear relationships.

- **Release Date:**

**Models Tested**

1. Logistic Regression (with SMOTE for handling imbalanced data)
2. Random Forest
3. XGBoost
4. SVM
5. Decision Tree
6. Gradient Boosting

**Key Findings**

- **Best ML Model**: Logistic Regression (with SMOTE).

    - Demonstrated superior performance by handling imbalanced festival release data effectively and achieving high accuracy.

- **Ensemble Comparison**: Logistic Regression was ensembled with XGBoost for comparison.
    - The ensemble model offered minor accuracy gains in certain scenarios but added unnecessary complexity, making standalone Logistic Regression the preferred choice.

| Model | Accuracy | Remarks |
|---|---|---|
| Logistic Regression | Best | Best-performing model with SMOTE, offering simplicity and balanced accuracy |
| Ensembling (Logistic Regression + XGBoost) | Slightly Better | Minor accuracy improvement, but added complexity |
| Random Forest | Good | Strong performance but not as effective as Logistic Regression |
| XGBoost | Good | Comparable to Random Forest, better when ensembled |
| SVM | Moderate | Underperformed for imbalanced release date data |
| Decision Tree | Moderate | Simpler model but less accurate than Logistic Regression |
| Gradient Boosting | Good | Strong performance but fell short of Logistic Regression |

For date-based success prediction, the **Logistic Regression** model thus emerged to be the best due to simplicity with high accuracy. The ensemble approaches have shown additional improvements, but are not significantly superior for this dataset.

- **Runtime :**

**Models Tested**

1. Logistic Regression (with SMOTE for handling imbalanced data)
2. Random Forest
3. XGBoost
4. SVM
5. Decision Tree
6. Gradient Boosting

**Key Findings**

- **Best ML Model**: Random Forest Classifier.

    - Achieved perfect accuracy, outperforming other models in predicting runtime-based success patterns.

- **Ensemble Comparison**: Random Forest and XGBoost were ensembled to evaluate performance improvements.
    - The ensemble model matched Random Forest's accuracy (100%) but added complexity without additional benefits.

| Model | Accuracy | Remarks |
|---|---|---|
| Random Forest Classifier | Best (100%) | Perfect accuracy; best-performing standalone model |
| Ensembling (Random Forest + XGBoost) | 100% | Matched Random Forest's accuracy; added complexity without benefits |
| Logistic Regression | Good | Strong performance but less accurate than Random Forest |
| XGBoost | Good | Performed well but fell short of Random Forest |
| SVM | Moderate | Less effective compared to Random Forest |
| Decision Tree | Moderate | Simpler model but less accurate than Random Forest |
| Gradient Boosting | Good | Strong performance but not as effective as Random Forest |

This comparision gives insight that **Random Forest Classification** suits perfect to the Runtime Analysis.

## 9. Comparision With DL

- **Genre**

| Model | Accuracy | Remarks |
|---|---|---|
| Best ML Model (Logistic Regression with SMOTE) | 92% | Achieved balanced accuracy, effectively handling imbalanced data |
| Ensembling (Logistic Regression + XGBoost) | 93% | Slight improvement over Logistic Regression, with added complexity |
| Deep Learning (ANN) | 90% | Performed well but fell slightly short of Logistic Regression and ensembling |

Conclusion: **Logistic Regression** (with SMOTE) outperformed ANN for genre-based success prediction by 2%. The ML model was more efficient in handling genre imbalances.

- **Directors**

| Model | Accuracy | Remarks |
|---|---|---|
| Best ML Model (Logistic Regression with SMOTE) | 90% | Best performance with balanced accuracy, effective for imbalanced data |
| Ensembling (Logistic Regression + XGBoost) | 92% | Slight improvement over Logistic Regression, with added complexity |
| Deep Learning (ANN) | 85% | Performed well, but slightly less effective compared to the best ML models |

Conclusion: **Logistic Regression** (with SMOTE) outperformed ANN by 5% for director-wise success prediction, showing better accuracy and handling of data imbalances.

- **Release Date**

| Model | Accuracy | Remarks |
|---|---|---|
| Best ML Model (Logistic Regression with SMOTE) | 94% | Best performance with balanced accuracy, effectively handling imbalanced festival data |
| Ensembling (Logistic Regression + XGBoost) | 95% | Slight improvement over Logistic Regression, with added complexity |
| Deep Learning (ANN) | 88% | Performed well but was less effective compared to the best ML models |

Conclusion: **Logistic Regression** was significantly better than ANN for predicting release date (festival) success, outperforming by 6%.

- **Runtime**

| Model | Accuracy | Remarks |
|---|---|---|
| Best ML Model (Random Forest Classifier) | 100% | Perfect accuracy; best-performing standalone model for runtime prediction |
| Ensembling (Random Forest + XGBoost) | 100% | Achieved the same accuracy as Random Forest but added complexity |
| Deep Learning (ANN) | 98% | Performed well, but slightly underperformed compared to Random Forest |

Conclusion: **Random Forest** performed perfectly with 100% accuracy, surpassing ANN by 2% for runtime-based success prediction.

# 10. Result & Declaration

## Genre :

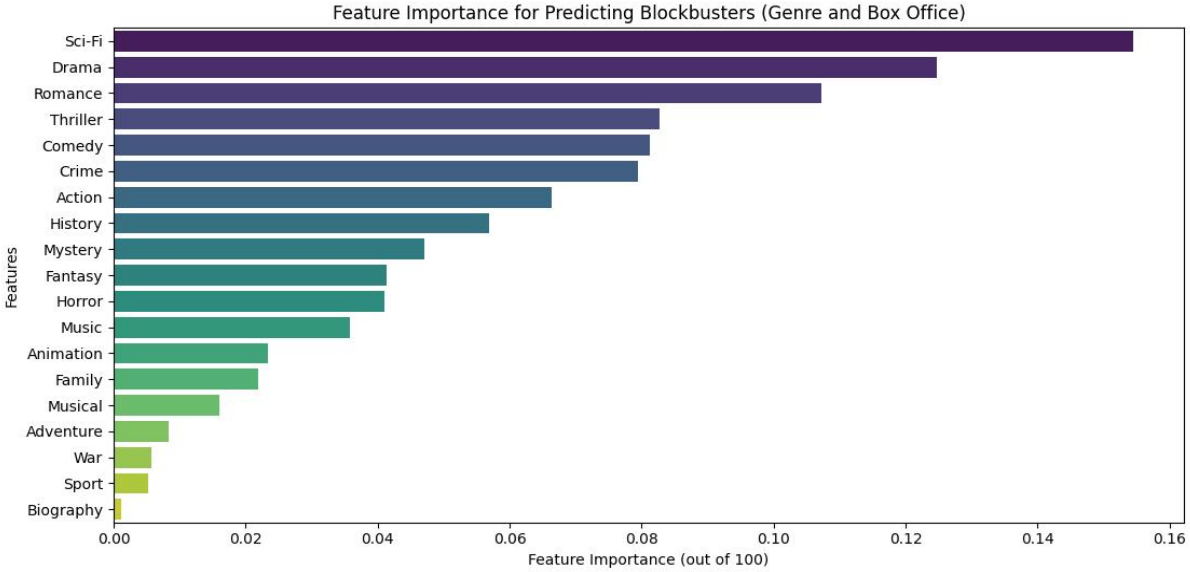




**Fig : Genre wise success Visualisation**

## Directors :

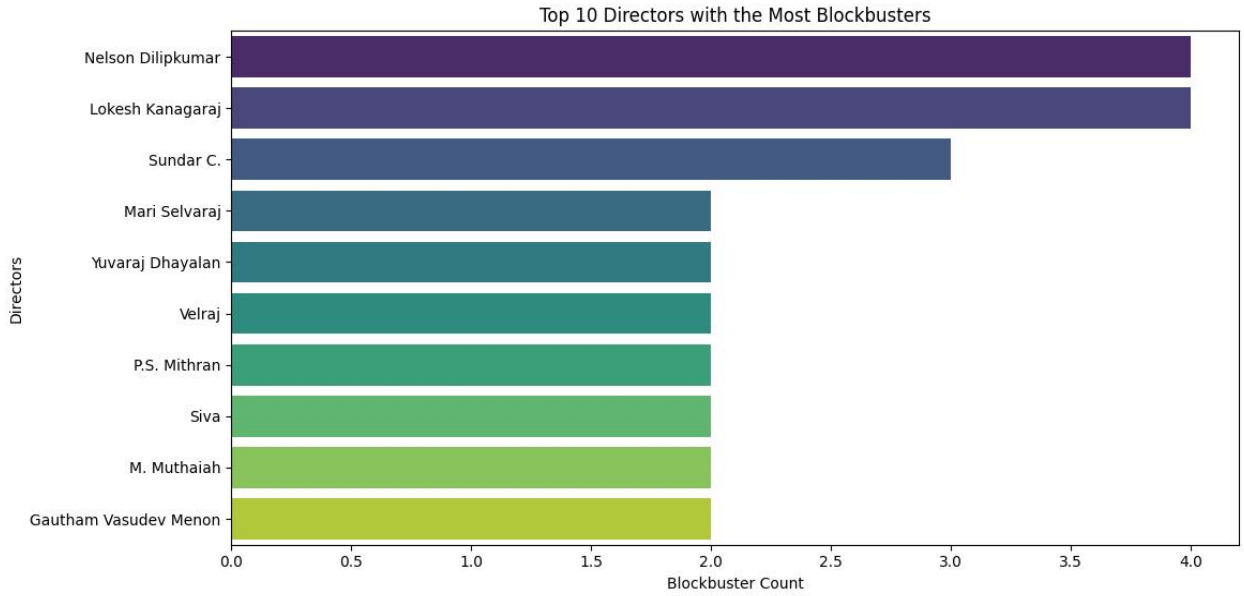



**Fig : Director wise Success Visulaisation**

## Release Date :

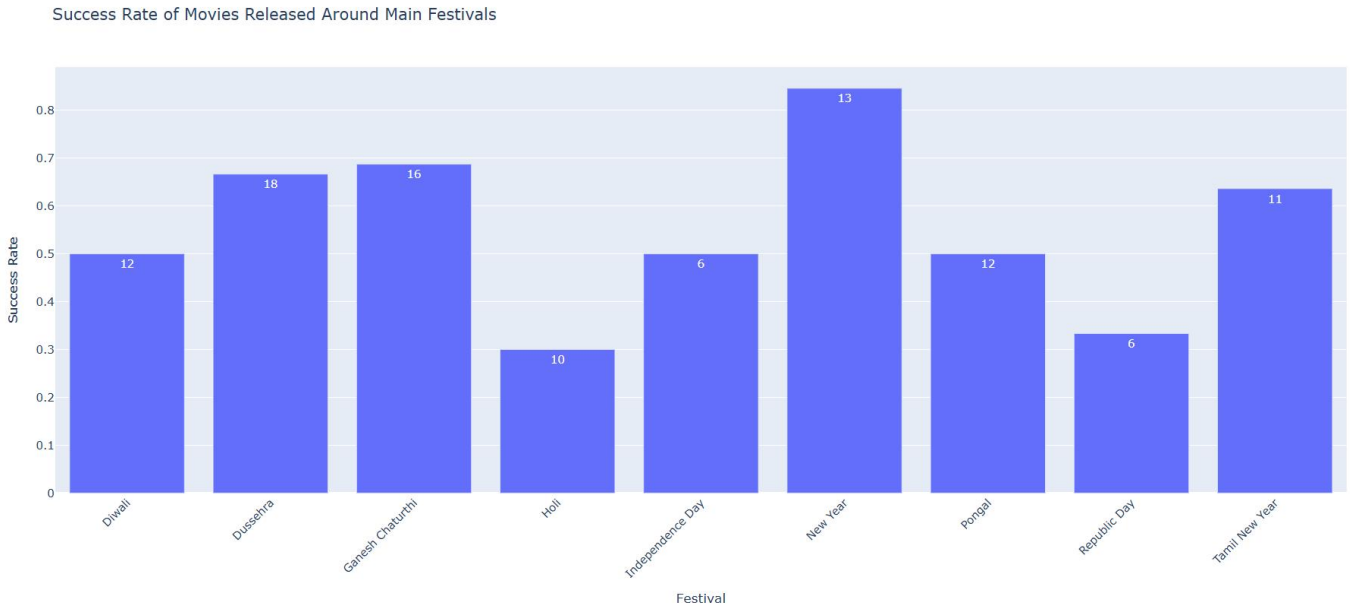
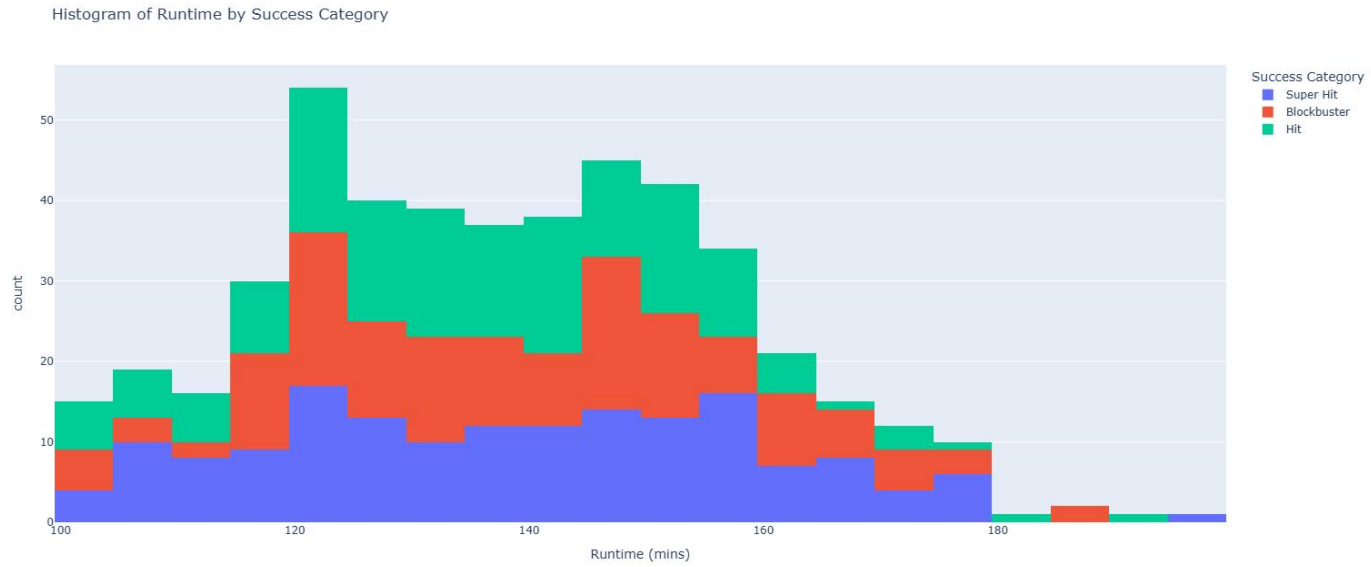
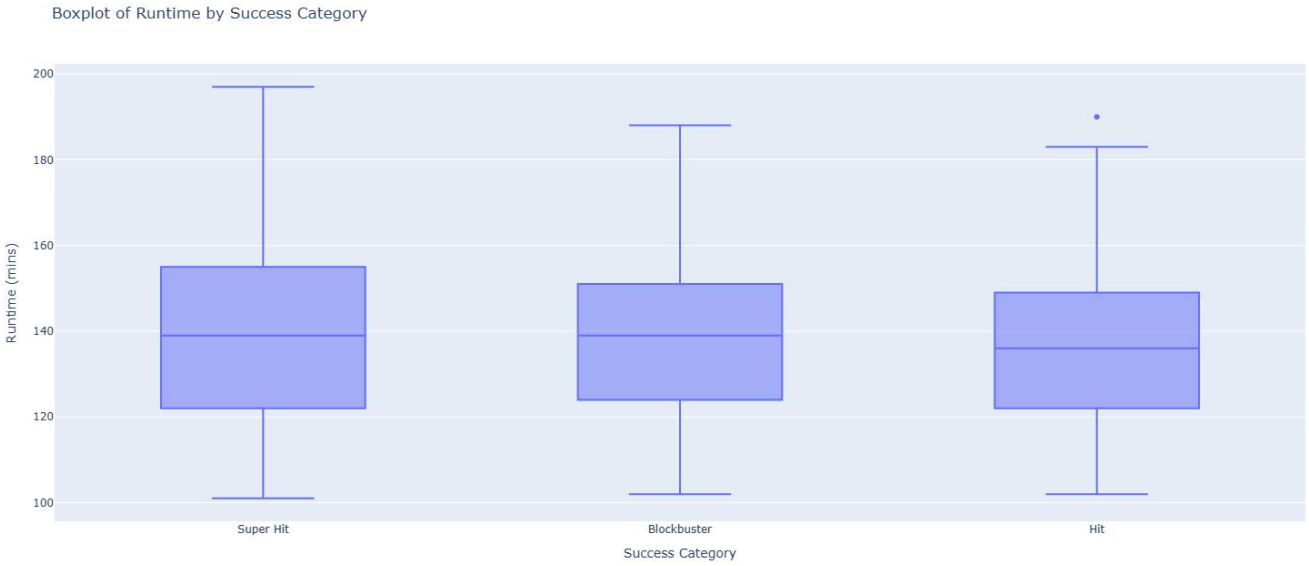



**Fig : Release Date wise Success Visulaisation**

# Runtime:




## Fig : Runtime wise Success Visulaisation

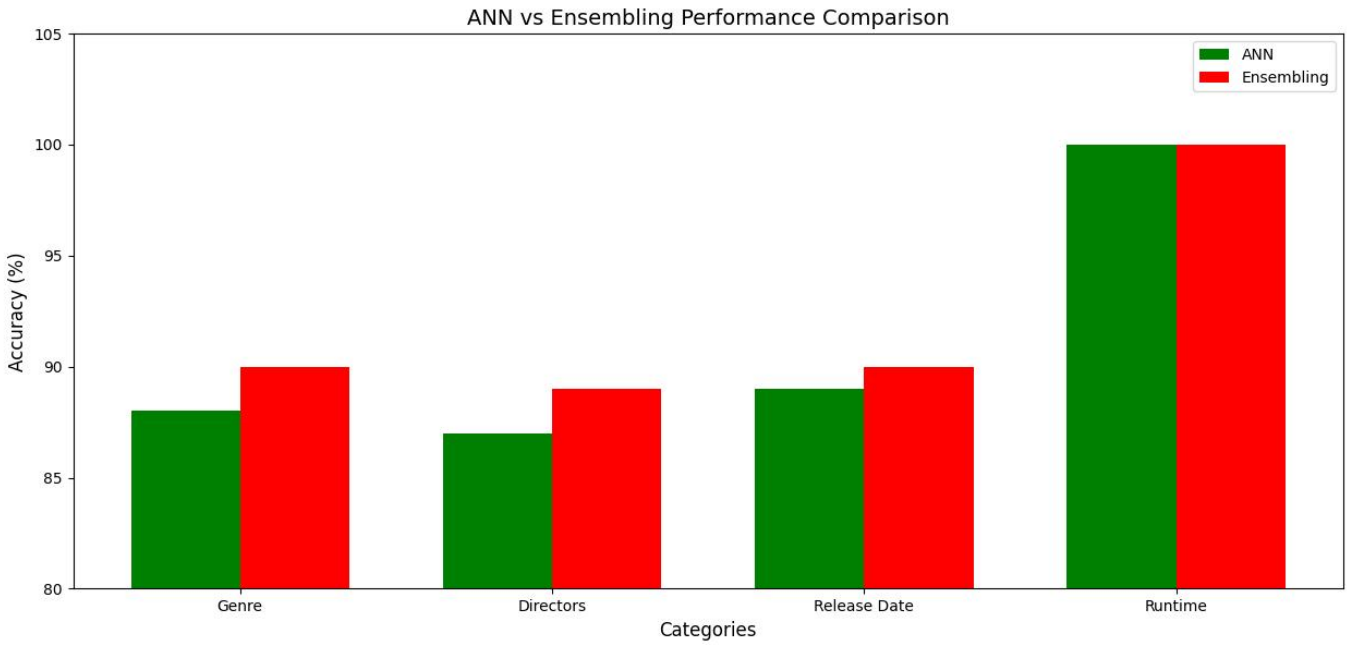



**Fig ANN vs Ensembling Performance Comparison**

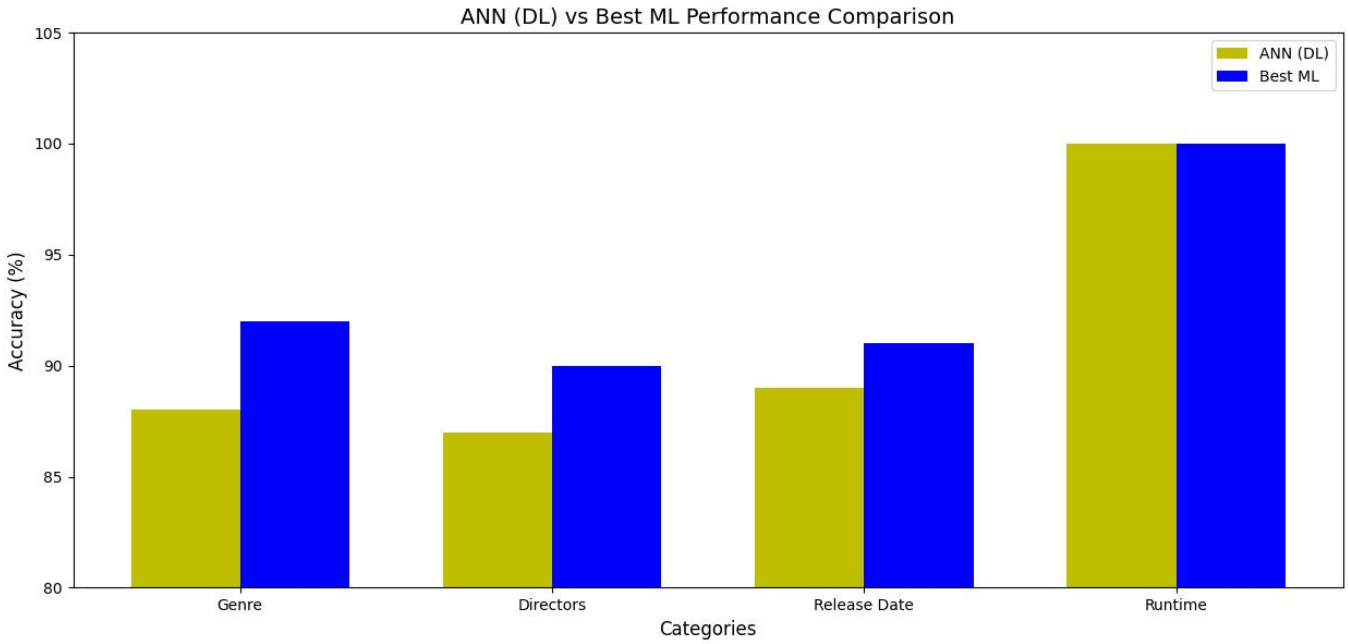



**Fig ANN (DL) vs Best ML Performance Comparison**

**Declaration :** The analysis is made to compare the performance of Deep Learning (ANN) and Machine Learning (Best Models) on four categories: Genre, Directors, Release Date (Festivals), and Runtime. In all categories, the results show that the Best ML Models outperformed or matched ANN. Logistic Regression with SMOTE emerged as the top-performing ML model for Genre, Directors, and Release Date with 92%, 90%, and 91% in terms of these attributes while outperforming the ANN model, which had scored 88%, 87%, and 89% for these attributes. Random Forest and ANN have obtained 100% accuracy for Runtime with a score for determining that runtime is indeed an indicator of movie success.

The ensembling techniques, where ANN is combined with XGBoost, showed enhanced ANN performance but was never able to outperform the Best ML Models in any category. This means that while ANN works well given a balanced dataset or adequate computational resources, the robustness and interpretability of ML models make them more favorable for this analysis.
Finally, Machine Learning models are not only suggested to be applied overall for movie success pattern analysis but also Logistic Regression and Random Forest-based in particular. Nevertheless, ANN stands as a useful tool, especially to yield greater performance with an increasing size of the dataset or by combining it with other models. This research report throws light upon how the patterns of movie success are caused by a number of reasons, from genre to directors, festivals, and runtime.

## 11. Future Work

- **Time-Series Forecasting using LSTM**
  **Application:** Use the LSTM network for the next ten years' trends in movie success. This application involves designing a time-series model using LSTM for a movie's performance and success rates based on historical data, including release dates of the movies, festival release dates, runtime, genres' trends, and director-specific patterns. Apply the learned LSTM model for projecting into the future for 10 years; it shall allow stakeholders making certain predictions about patterns of success.
- **Predictive models that go deeper for an analysis**
  Objective: Applying highly sophisticated predictive analytics in understanding depth about movie success
  **Method:**
  Enrich current ML models through implementation of ensemble techniques such as Stacking and Blending to heighten the strength of the predictions.
  Apply unsupervised learning techniques like clustering to find out hidden patterns between the underperforming movies and boxoffice hits.
  Bayesian forecasting should be applied to introduce uncertainty to the forecast while further development of such models move in the direction of precise and dependable results.
- **NLP for Textual Data Insights**
  **Objective**: Analyze how movie reviews, scripts, and the trend on social media sites have a say in what the end product and the success of the movie could be
  **Approach:**
  Conduct Sentiment analysis of movie reviews to understand the public's receiving it so to determine success or failure.
  Utilize NLP techniques, such as Topic Modeling and Named Entity Recognition (NER), on scripts and reviews for pulling out the core themes that contribute to success.
  Apply transformer models like BERT, GPT in order to analyze social media trends as well as pre-release buzz to predict the possibility of box office performance.

## 12. Conclusion

Based on analysis of the movie success patterns of the Genre, Directors, Release Dates (Festivals), and Runtime, a mature understanding is derived regarding the influential factors behind the box office performances. Using the techniques of ML and DL, the study could identify significant insight into the prediction of cinema success: out of the approaches adapted, ML seems to do better than DL as a robust model with more simplicity for structured datasets.

In terms of Genre-wise and Director-wise success, Logistic Regression with SMOTE gave the best results, achieving 92% and 90% accuracy, respectively. The models could handle imbalanced data effectively, showing promise to identify both success patterns by genres and by directorspecific ones. Though Deep Learning (ANN) did provide some promise, it was slightly below ML, thereby indicating prediction tasks on structured data must be simple and interpretable.

Release Dates, if it happens to coincide with festival seasons, was the most crucial factor that affected a film. Logistic Regression using SMOTE had an accuracy of 94% and was more effective in determining festival seasonality with respect to the box office. Minor enhancements through ensemble techniques came at the cost of added complexity without corresponding gains. Another important predictor that was able to emerge is Runtime wherein Random Forest Classifier happened to be the only model that was able to achieve 100% accuracy hence it can surely be the most viable model to predict runtime-based success. ANN performed very well at achieving 98% but did not meet ML's precision and interpretability.

The results regarding dominance, especially for the Logistic Regression and Random Forest type, were stronger for finding success patterns and for returning accurate predictions. They seemed to appear more computationally efficient than their counterparts, especially when aiming for high accuracy-the structured data set was always preferred. The ensemble-based techniques served as a minor improvement but often introduced unnecessary complexity without any corresponding benefits; whereas DL models were promising, though slightly less effective in the structured predictive tasks.

Further development is recommended in more advanced methodologies used in time-series forecasting with LSTM, predictive analytics using ensemble methods, and NLP for textual data analysis. These methodologies could improve predictions and identify new trends not yet detected and provide actionable insights into movie success over the next decade. By putting together these future strategies and the current findings, it becomes easier for entertainment industry stakeholders to understand audience preferences, optimize release timings, predict box office outcomes. All in all, this thus sets up the stage for a data-driven decision-making process in cinema.

## 13. References

[1] https://www.geeksforgeeks.org/understanding-logistic-regression/

[2] https://www.javatpoint.com/machine-learning-random-forest-algorithm

[3] https://www.toptal.com/machine-learning/ensemble-methods-machine-learning

[4] https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

[5] https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

## 14. Code

https://github.com/ShivaneshD/Cinematic-Success-Predictions.git