MOVIE SUCCESS PREDICTION: PROJECT REPORT
Presented by: Kunal Tyagi

---

1. Introduction

The movie industry involves high-stakes investment, where inaccurate forecasting of a film's success can lead to major financial losses. This capstone project aims to leverage machine learning to predict whether a movie will be a *Flop*, *Average*, or *Hit* before its release, focusing especially on accurately identifying potential *Hit* movies.

---

2. Problem Statement

Film production companies often face difficulties in determining a movie's potential success at the early stages. The objective is to develop a predictive model that classifies movies into success categories (Hit, Average, Flop) using relevant features, aiding data-driven decision-making for greenlighting projects.

---

3. Objectives

- Analyze and clean the movie dataset.
- Perform detailed Exploratory Data Analysis (EDA).
- Engineer features and prepare data for modeling.
- Train classification models (Logistic Regression, Random Forest, Gradient Boosting).
- Evaluate models with a focus on precision for *Hit* class.

---

4. Dataset Overview

- Over 5000 movie records.
- Features include: Budget, Gross Earnings, Cast Facebook Likes, IMDb Score, etc.
- The IMDb score was converted into categorical target classes: *Flop*, *Average*, and *Hit*.
- Major challenges included missing data and skewed distributions in numeric fields.

---

5. EDA & Insights

- High-budget movies with high social media engagement are more likely to be hits.
- Features like Facebook Likes and Votes showed strong correlation with movie success.
- Visualization of distributions helped identify patterns across different success classes.

---

6. Data Preprocessing & Feature Engineering

- Irrelevant columns removed.
- Missing values imputed using median or mode.

- Categorical features label-encoded; numeric features scaled.
- Log transformation applied to skewed features like Budget and Gross.

---

7. Model Building

- Three models were trained using an 80/20 stratified train-test split:
  - **Logistic Regression**
  - **Random Forest**
  - **Gradient Boosting**
- Evaluation metrics included Accuracy, Precision, Recall, and F1-Score.

---

8. Model Evaluation & Comparison

Model	Accuracy	Precision (Hit)
Random Forest	0.84	0.88
Gradient Boosting	0.83	0.86
Logistic Regression	0.72	0.78

- **Random Forest** emerged as the best-performing model with balanced results.
- Focus was on maximizing precision for *Hit* movies to reduce false positives.

---

9. Business Insights

- Budget and social media engagement are strong predictors of movie success.
- Facebook Likes emerged as a key feature.
- The model can guide investment allocation and early marketing strategies.

---

10. Future Enhancements

- Hyperparameter tuning (e.g., GridSearchCV) for better performance.
- Incorporate sentiment analysis from reviews or trailers.
- Deploy an interactive dashboard using Power BI or Streamlit.

---

11. Conclusion

The project demonstrates that machine learning can effectively predict movie success categories using pre-release data. The Random Forest model offers a reliable tool for production houses to assess project viability and plan investments. Next steps include deploying the model via a web app for real-world use.