



**Project Report On**

**Box Office Performance Prediction**



Submitted in partial fulfillment for the award of

Post Graduate Diploma in Big Data Analytics (PG-DBDA) From Know-IT(Pune)

**Guided by: Trupti Joshi Ma’am**

**Prasad sir**

## Submitted By:

Ajay Hoke (230343025019)

Amisha Kotkar (230343025027)

Sakshi Patil (230343025037)

Anjali Sul (230343025050)







# CERTIFICATE

**TO WHOMSOEVER IT MAY CONCERN**

**This is to certify that**

Ajay Hoke (230343025019)

Amisha Kotkar (230343025027)

Sakshi Patil (230343025037)

Anjali Sul (230343025050)

#### Have successfully completed their project on

**Box Office Performance Prediction**

**Under the guidance of Prasad sir**







# ACKNOWLEDGEMENT

This project **“Box Office performance prediction using ML and Spark”** was a great learning experience for us and we are submitting this work to CDAC Know-IT (Pune).

We all are very glad to mention the name of **Prasad Sir** for his valuable guidance to work on this project. His guidance and support helped us to overcome various obstacles and intricacies during the course of project work.

We are highly grateful to **Mr. Vaibhav Inamdar** Manager (Know-IT), C-DAC, for his guidance and support whenever necessary while doing this course Post Graduate Diploma in Big Data Analytics (PG-DBDA) through C-DAC ACTS, Pune.

Our most heartfelt thanks goes to **Mrs. Bakul Joshi** (Course Coordinator,PG-DBDA) who gave all the required support and kind coordination to provide all the necessities like required hardware, internet facility and extra Lab hours to complete the project and throughout the course up to the last day here in C-DAC Know-IT, Pune.

**From:**

Ajay Hoke (230343025019)

Amisha Kotkar (230343025027)

Sakshi Patil (230343025037)

Anjali Sul (230343025050)







**TABLE OF CONTENTS**

#### ABSTRACT

1. **INTRODUCTION**

#### SYSTEM REQUIREMENTS

* 1. Software Requirements
  2. Hardware Requirements

#### FUNCTIONAL REQUIREMENTS

1. **SYSTEM ARCHITECTURE**

#### METHODOLOGY

1. **MACHINE LEARNING ALGORITHMS**

#### DATA VISUALIZATION AND REPRESENTATION

1. **CONCLUSION AND FUTURE SCOPE References**





## Abstract

The project aims to create an advanced predictive model for predicting a movie's box office performance.

It analyzes factors like genre, release date, budget, and box office collections to uncover patterns affecting

a film's commercial success.

Python is the main programming language for data manipulation, preprocessing, and machine learning. Apache Spark is used for large-scale data analysis, ensuring efficiency.

Tableau is employed for data visualization, creating interactive dashboards and charts to showcase trends and key variables influencing box office performance.

The model has the potential to revolutionize how the entertainment industry approaches marketing and distribution, providing data-driven insights for cinematic success.

"Predictive Box Office Performance" merges data science with entertainment, reflecting our commitment to data-driven solutions for real-world challenges.

By pioneering this predictive model, we aim to reshape the film industry's strategies and enhance its creative endeavors.

.





## INTRODUCTION

* In the dynamic realm of the entertainment industry, the performance of a movie at the box office is a culmination of numerous intricate factors that together shape its success or failure. The ability to predict the box office performance of a movie has been an enigmatic challenge that filmmakers, producers long, and investors strive to decipher. The ever-evolving landscape of cinema, characterized by its reliance on captivating storytelling, exceptional talent, and innovative direction, presents an intriguing opportunity to explore the relationship between various film attributes and their influence on box office earnings.
* The project at hand delves into the captivating domain of Box Office Performance Prediction, aiming to leverage the power of data-driven insights and machine learning techniques to develop a predictive model. By analyzing key features such as budget, cast composition including actors and actresses, directorial prowess, and even linguistic attributes like spoken languages, the project endeavors to unravel the hidden patterns that impact a movie's financial success.
* The project will use machine learning algorithms such as clustering, classification and regression to analyze the data and identify patterns and trends in crime rates. The project will use Apache Spark, a distributed computing system, to analyze large datasets efficiently.

#### Datasets and features:

Data used in the project is structured in nature from year 20012 to 2017. It was collected from [www.kaggle.com.](http://www.kaggle.com/) The main goal of the analysis is to be build accurate and robust regression model to predict the box office performance of movies. This research uses Random Forest, Decision Tree, XGBoost , LightGBM.







## SYSTEM REQUIREMENTS

#### Hardware Requirements:

* Platform – Windows
* RAM – 8 GB of RAM,
* Peripheral Devices – Mouse, Keyboard, Monitor
* A network connection for data recovering over network.

#### Software Requirements:

* Python 3
* Machine Learning
* Apache Spark
* Tableau
* Google Collab and Jupyter
* **OS – Windows**





## FUNCTIONAL REQUIREMENTS

#### Python 3:

 Python is a general purpose and high level programming language.

 It is use for developing desktop GUI applications, websites and web applications.

 Python allows to focus on core functionality of the application by taking care of common programming tasks.

 Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, Small Talk, and Unix shell and other scripting languages.

#### Apache Spark:

 What is Spark: Apache Spark is an open-source distributed computing system designed for processing large volumes of data.

 Key Features: Spark provides a number of key features that make it well-suited for processing big data, including in-memory processing, support for various data sources and formats, fault- tolerance, and scalability.

 Spark also provides a range of APIs, including SQL, streaming, machine learning, and graph processing, making it a versatile platform for a wide range of use cases.

#### Tableau:

 Data visualization is the graphical representation of information and data.

 It helps create interactive elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

 Tableau is widely used for Business Intelligence but is not limited to it.

 It helps create interactive graphs and charts in the form of dashboards and worksheets to gain business insights.

 All of this is made possible with gestures as simple as drag and drop.





#### Data Cleaning Process:



**Fig: Data Cleaning Process**

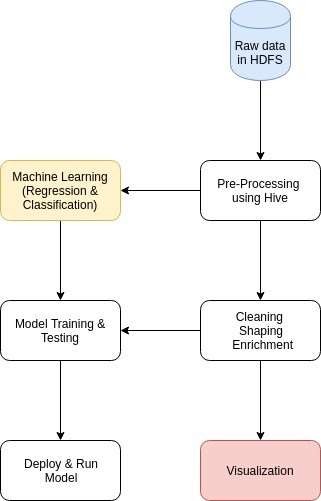
Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleansing may be performed interactively with data wrangling tools, or as batch processing

through scripting. After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleaning differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data.



## SYSTEM ARCHITECTURE

Regression Models

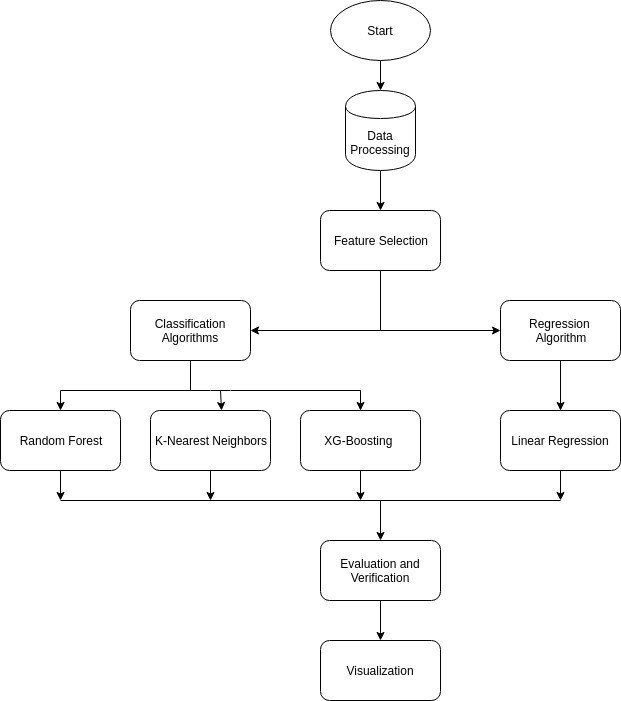


**Pre- Processing**

**Fig: System Architecture Of Box Office Prediction**







LGB

**Fig: Process of Box Office Performance Prediction**







# MACHINE LEARNING ALGORITHMS

In this project we apply various different types of Regression Algorithms such as Random Forest,

Decision Tree, XGBoost , LightGBM During the implementation we analyse the accuracy of all the algorithms.

Machine learning is the research that explores the development of algorithms that can learn from data and provide predictions based on it. Works that study flight systems are increasing the usage of machine learning methods. The methods commonly used include Random Forest, Decision Tree, XGBoost , LightGBM They were mainly used for classification and prediction. In this project we use various machine learning algorithms which are as follows:





## Random Forest:

 The Random Forest (RF) classifier is an ensemble method based on multiple decision trees.

 By combining the Bootstrap aggregating and random space method, RF overcomes the drawbacks of individual decision tree.

 RF is widely used in industry because it can classify high dimensional data in short time with good performance and it has low sensitivity to outliers in the training data.

**Pros:**

 The predictive performance can compete with the best supervised learning algorithms.

 They provide a reliable feature importance estimate

 They offer efficient estimates of the test error without incurring the cost of repeated model training associated with cross-validation.

**Cons:**

 An ensemble model is inherently less interpretable than an individual decision tree  Training a large number of deep trees can have high computational costs (but can be parallelized) and use a lot of memory.

## XGBoost Algorithm:

 **XGBoost** is a powerful machine learning algorithm used for both classification and regression tasks, capable of handling both categorical and numerical data..

**Pros:**

 Very Useful in building classification algorithm.

**Cons:**

 computationally intensive and potentially requiring more time and resources for training compared to simpler models.



## DATA VISUALIZATION AND REPRESENTATION





















# CONCLUSION AND FUTURE SCOPE

## Conclusion:

.





## Future Scope:





# References