

# **MOVIE BOX OFFICE GROSS PREDICTION USING IBM WATSON MACHINE LEARNING**

**AN INDUSTRY ORIENTED MINI REPORT**

Submitted to

**JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY,  
HYDERABAD**

In partial fulfillment of the requirements for the award of the degree of

**BACHELOR OF TECHNOLOGY  
IN  
COMPUTER SCIENCE AND ENGINEERING**

Submitted by

**MADIPADIGE SAI CHARAN**

**19UK1A0520**

**KAGITHA RAMYA**

**19UK1A0551**

**CHENNABOINA MOUNIKA**

**19UK1A0527**

**AIRNENI PRANITHA**

**19UK1A0570**

Under the esteemed guidance of

**Mrs. K. SOWMYA**

(Assistant Professor)



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
VAAGDEVI ENGINEERING COLLEGE**

(Affiliated to JNTUH, Hyderabad)

Bollikunta, Warangal – 506005

**2019– 2023**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
VAAGDEVI ENGINEERING COLLEGE  
BOLLIKUNTA, WARANGAL – 506005  
2019 – 2023**



**CERTIFICATE OF COMPLETION  
INDUSTRY ORIENTED MINI PROJECT**

This is to certify that the Industry Oriented Mini Project entitled “**MOVIE BOX OFFICE GROSS PREDICTION USING IBM WATSON MACHINE LEARNING**” is being submitted by *M.SAICHARAN*(H.NO:19UK1A0520),*K.RAMYA*(H.NO:19UK1A0551),*CH.MOUNIKA* (H.NO:19UK1A0527),*A.PRANITHA*(H.NO:19UK1A0570) in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science and Engineering** to **Jawaharlal Nehru Technological University Hyderabad** during the academic year **2022-23**, is a record of work carried out by them under the guidance and supervision.

**Project Guide**  
**Mrs. K. Sowmya**  
(Assistant Professor)

**Head of the Department**  
**Dr. R. Naveen Kumar**  
(Professor)

**External**

## ACKNOWLEDGEMENT

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved **Dr.P.PRASAD RAO**, Principal, Vaagdevi Engineering College for making us available all the required assistance and for his support and inspiration to carry out this Industry Oriented Mini Project in the institute.

We extend our heartfelt thanks to **Dr.R.NAVEEN KUMAR**, Head of the Department of CSE, Vaagdevi Engineering College for providing us necessary infrastructure and thereby giving us freedom to carry out the Industry Oriented Mini Project.

We express heartfelt thanks to Smart Bridge Educational Services Private Limited, for their constant supervision as well as for providing necessary information regarding the Industry Oriented Mini Project and for their support in completing the Industry Oriented Mini Project.

We express heartfelt thanks to the guide, **Mrs. K. Sowmya** Assistant professor, Department of CSE for his constant support and giving necessary guidance for completion of this Industry Oriented Mini Project.

Finally, we express our sincere thanks and gratitude to my family members, friends for their encouragement and outpouring their knowledge and experience throughout the thesis.

**M. SAI CHARAN (19UK1A0520)**  
**K. RAMYA (19UK1A0551)**  
**CH. MOUNIKA (19UK1A0527)**  
**A.PRANITHA (19UK1A0570)**

# ABSTRACT

Predicting society's reaction to a new product in the sense of popularity and adaption rate has become an emerging field of data analysis. The motion picture industry is a multi-billion-dollar business, and there is a massive amount of data related to movies that is available over the internet. This study proposes a decision support system for movie investment sector using machine learning techniques. This research helps investors associated with this business avoid investment risks. The system predicts an approximate success rate of a movie based on its profitability by analyzing historical data from different sources like IMDb, Rotten Tomatoes, Box Office Mojo, and Metacritic. Using Support Vector Machine (SVM), Neural Network and Natural Language Processing, the system predicts a movie box office profit based on some pre-released features and post-released features. This paper shows Neural Network gives an accuracy of 84.1% for pre-released features and 89.27% for all features, while SVM has 83.44% and 88.87% accuracy for pre-released features and all features respectively, when one away prediction is considered. Moreover, we figure out that budget, IMDb votes, and no. Out screens are the most important features which play a vital role in predicting a movie's box-office success.

# **CONTENTS**

<b>SNO:</b>	<b>TOPIC</b>
<b>1.</b>	<b>INTRODUCTION</b>
<b>2.</b>	<b>LITERATURE SURVEY</b>
<b>3.</b>	<b>THEORITICAL ANALYSIS</b>
<b>4.</b>	<b>EXPERIMENTAL INVESTIGATIONS</b>
<b>5.</b>	<b>FLOW CHART</b>
<b>6.</b>	<b>RESULT</b>
<b>7.</b>	<b>ADVANTAGES &amp; DISADVANTAGES</b>
<b>8.</b>	<b>APPLICATIONS</b>
<b>9.</b>	<b>CONCLUSION</b>
<b>10.</b>	<b>FUTURE SCOPE</b>
<b>11.</b>	<b>BIBLOGRAPHY</b>
	<b>APPENDIX</b>

# **1. INTRODUCTION**

## **1.1 overview**

Predicting society's reaction to a new product in the sense of popularity and adoption rate has become an emerging field of data analysis, and such kind of analysis can help the movie industry to take appropriate decisions. Can film studios and its related stakeholders use a forecasting method for the prediction of revenue that a new movie can generate based on a few given input attributes like budget, runtime, released year, popularity, and so on. This study marks as a decision support system for the movie investment sector using machine learning techniques. This project helps investors associated with this business for avoiding investment risks. The system predicts an approximate success rate of a movie based on its profitability by analyzing historical data from different sources like Online rating, Director, Budget, Pre-Release business, Genre, etc.

## **1.2 Purpose**

The film industry has grown immensely over the past few decades generating billions of dollars of revenue for the stakeholders. Now people can watch movies online and offline on a variety of mobile devices during leisure or travel through Netflix, YouTube and downloads. A prediction system to assess the box office success of new movies can help the movie producers and directors make informed decisions when making the movie in order to increase the chance of profitability and box office gross success. New social media tools are constantly appearing which are enabling people to gather information on films and post comments about movies. These comments can influence the initial prediction about the box office gross success of a movie which some of the existing research do not consider. Critic reviews often come out a few days before the film is released and may, therefore, help in prediction and at the same time influence the box office revenue.

## **2.LITERATURE SURVEY**

### **2.1 Existing Problem (OR) Problem Statement**

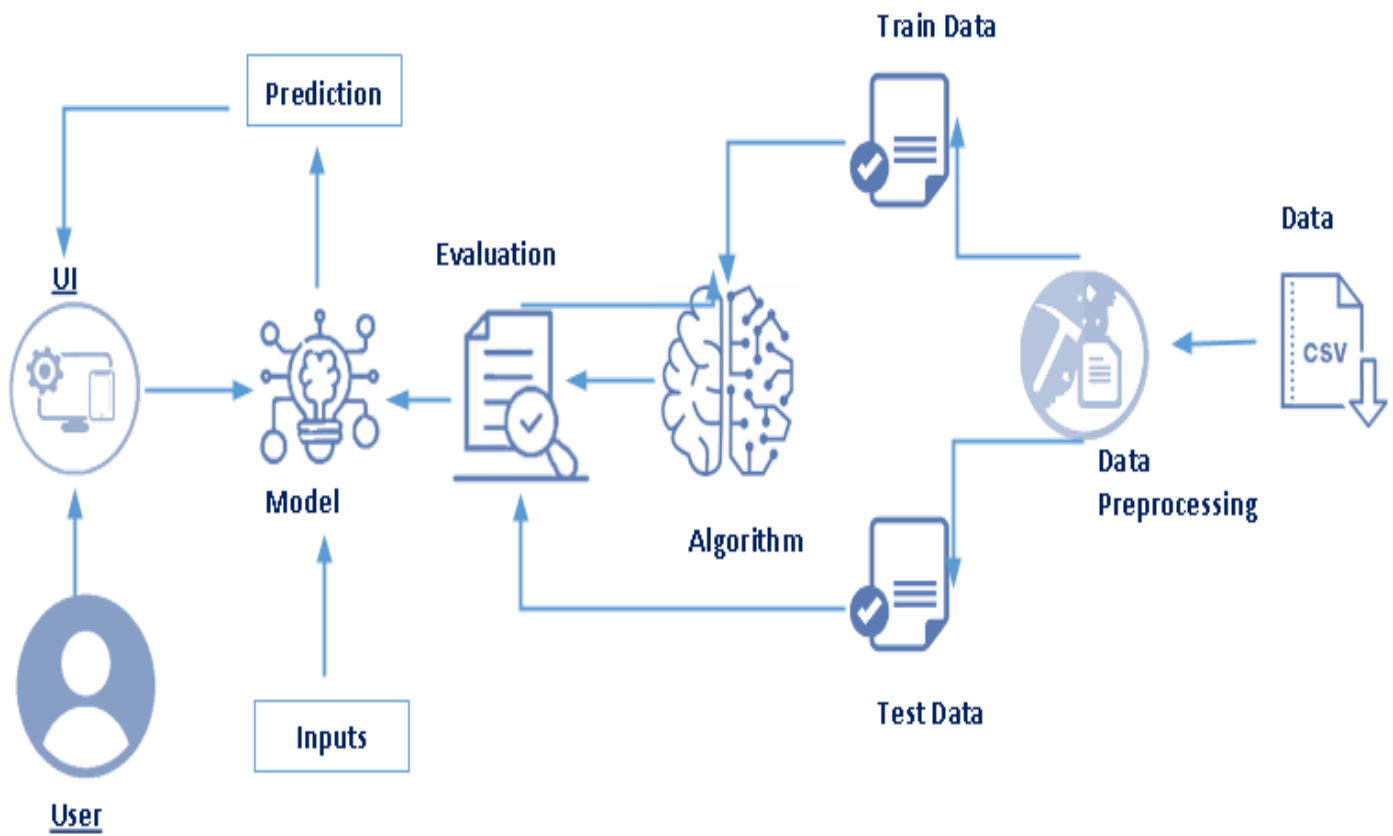
Given two datasets containing various attributes, use the features available in the dataset and define a supervised classification algorithm which can identify whether the movie gross getting correct predicted values or not. This data set contains many movie gross records. The data set records were collected all over the world.

### **2.2 Proposed Solution**

This is a classic example of supervised learning. We have been provided with a fixed number of features for each data point, and our aim will be to train a variety of Supervised Learning algorithms on this data, so that when a new data point arises, our best performing classifier can be used to categorize the data point as a positive example or negative. Exact details of the number and types of algorithms used for training is included in the 'Algorithms and Techniques' sub-section of the 'Analysis' part. This project focuses on the related works of various movies to calculate box office gross prediction such that algorithms were implemented using Jupyter that is a machine learning software written in Python. Various attributes that are essential in the prediction of movie gross were examined and the dataset of movies were also evaluated. This project compares various classification algorithms such as Random Forest, Support Vector Machine and KNN classification Algorithm with an aim to identify the best technique. Based on this study, Random Forest with the highest accuracy outperformed the other algorithms and can be further utilized in the prediction of movie box office gross recommended to the user. Later by using Flask app create html files and create a user interface to display the movie box office gross prediction values.

### 3.THEORITICAL ANALYSIS

#### 3.1 Block Diagram





## 3.2 Hardware / Software Designing

The following is the Hardware required to complete this project:

- Internet connection to download and activate
- Administration access to install and run Anaconda Navigator
- Minimum 10GB free disk space
- Windows 8.1 or 10 (64-bit or 32-bit version) OR Cloud: Get started free, \*Cloud account required.

Minimum System Requirements To run Office Excel 2013, your computer needs to meet the following minimum hardware requirements:

- 500 megahertz (MHz)
- 256 megabytes (MB) RAM
- 1.5 gigabytes (GB) available space
- 1024x768 or higher resolution monitor

The following are the software required for the project:

- Google Colaboratory Notebook and Jupyter Notebook
- Spyder and Pycharm Community
- Microsoft Excel 2013

## **4.EXPERIMENTAL INVESTIGATIONS**

Coming to analysis or investigations three supervised learning approaches are selected for this problem. Movies is taken that all these approaches are fundamentally different from each other, so that we can cover as wide an umbrella as possible in term of possible approaches. For each algorithm, we will try out different values of a few hyper parameters to arrive at the best possible classifier. This will be carried out with the help of grid search cross validation technique. There are several Machine learning algorithms to be used depending on the data you are going to process such images, sound, text, and numerical values. The algorithms that you can choose according to the objective that you might have may be classification algorithms and Regression algorithms.

### **(1) Support Vector Machine**

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. Support Vectors are simply the coordinates of individual observation. The goal of a support vector machine is not only to draw hyperplanes and divide data points, but to draw the hyperplane the separates data points with the largest margin, or with the most space between the dividing line and any given data point.

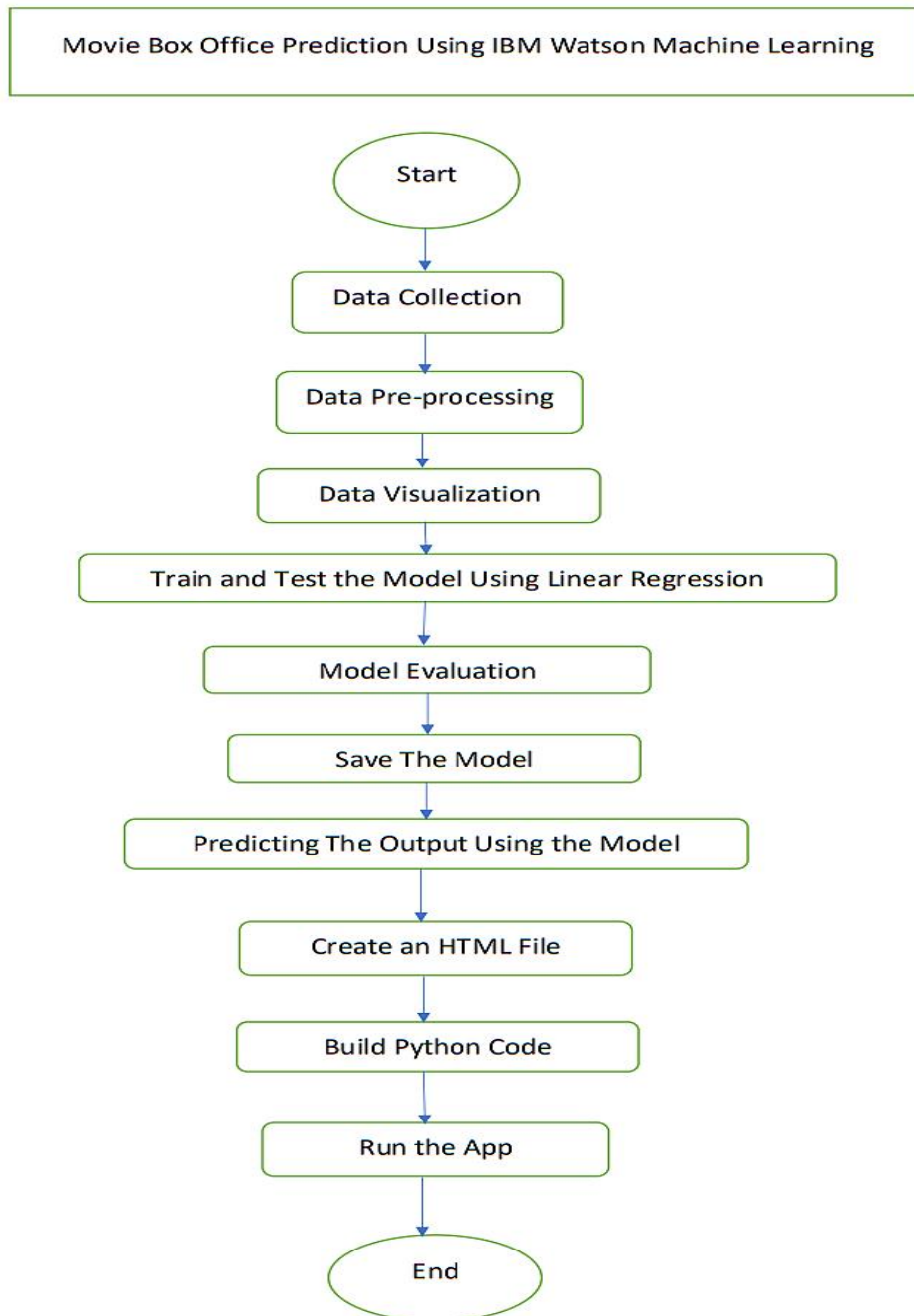
### **(2) Random Forest Classification**

Random Forest or Random decision forests are an ensemble method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees.

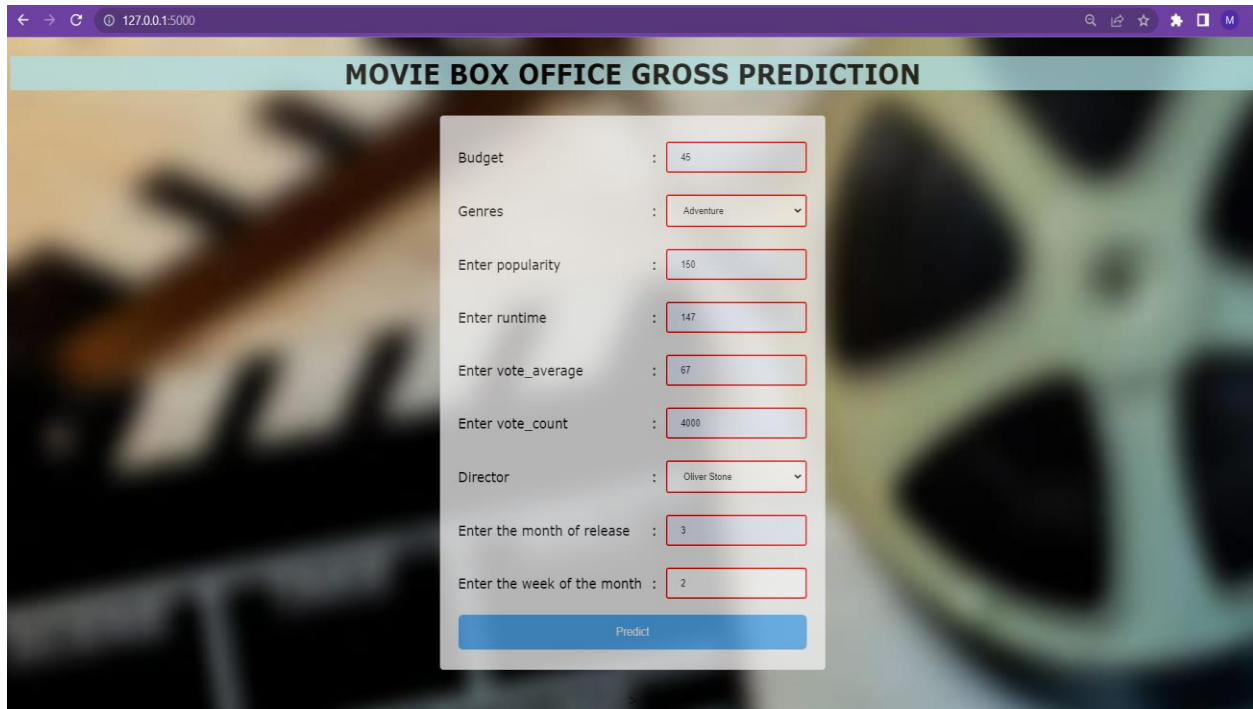
### **(3) KNN Classification algorithm or K-Nearest Neighbour algorithm**

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

## 5.FLOWCHART



## 6.RESULT



**MOVIE BOX OFFICE GROSS PREDICTION**

Budget	:	<input type="text" value="45"/>
Genres	:	<input type="text" value="Adventure"/>
Enter popularity	:	<input type="text" value="150"/>
Enter runtime	:	<input type="text" value="147"/>
Enter vote_average	:	<input type="text" value="67"/>
Enter vote_count	:	<input type="text" value="4000"/>
Director	:	<input type="text" value="Oliver Stone"/>
Enter the month of release	:	<input type="text" value="3"/>
Enter the week of the month	:	<input type="text" value="2"/>



## 7.ADVANTAGES AND DISADVANTAGES

### ADVANTAGES:

- **Efficiency in workflow:** One of the first desires that probably comes to mind is efficiency. When building your website, you want to be able to reach as many people as you can. The system predicts an approximate success rate of a movie based on its profitability by analyzing data.
- **Reduce costs:** You don't need a large time to wait for the results with in a second the amount will be predicted.
- Using machine learning algorithms to predict movie box office gross in data sets. Various kinds of data sets, have to use this to train classifier algorithms to predict movie gross with good accuracy.

### DISADVANTAGES:

- Any single error in data set can change the entire data.
- Correct accuracy must be needed while doing the project using supervised machine learning algorithms.
- Python code should be correct without any error.

## **8.APPLICATIONS**

This application can further be developed with more idea and implementation and by using different algorithms. The accuracy score of the model can be further improved by using decision tree and also by increasing the data set, K-Nearest Neighbors algorithm is also one of the pertinent methods which can be used to predict the movie gross accurately. It proposes to improve the accuracy further.

## **9.CONCLUSIONS**

Prompt and timely accurate prediction of Movie box office gross plays a vital role in decreasing for producers and stakeholders. In this paper, an attempt is made to predict the presence of Movie gross using Support vector machine, Random Forest, K-NN classification methods of Machine Learning. I developed a computational model for movie box office gross prediction using a combination of features extracted from movie database metadata, budget-revenue relationship graphs, popularity? Revenue relationship graphs, and movie-Revenue relationship graphs. I demonstrated that by using features extracted from these runtime-movies and revenue-movie relationship graphs, we are able to create a more accurate model than using metadata features alone. Among three ML classification methods, SVM and RF performed better than K-NN classification. Although, the accuracy levels for all three methods performed well based on the testing data set.

## **10.FUTURE SCOPE**

There are a variety of extensions that could be made to the existing model proposed in this paper. One alternative would be to model movie gross as a continuous quantity rather than a discrete quantity. Another extension would be to introduce a temporal model for how movie genre and actor popularity change over time, which one might suspect would lead to more accurate gross predictions. Finally, perhaps the biggest improvement that could be made would be to acquire more Movie box office gross data.

## **11.BIBILOGRAPHY**

We referred some books and surfed the internet for the better outcome of the project:

- [1] Simonoff, J. S. and Sparrow, I. R. Predicting movie grosses: Winners and losers, blockbusters and sleepers. In *Chance*, 2000.
- [2] Joshi, M., Das, D., Gimpel, K., and Smith, N. A. Movie Reviews and Revenues: An Experiment in Text Regression. In *Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference*, 2010.
- [3] Sharda, R. and Delen, D. Predicting box-office success of motion pictures with neural networks. In *Expert Systems with Applications*, 2006.

# APPENDIX

## A. COLAB NOTEBOOK

[https://colab.research.google.com/drive/1SJKgm3dQPZFcWqRnXHybyCHUIIES\\_QXx?usp=sharing](https://colab.research.google.com/drive/1SJKgm3dQPZFcWqRnXHybyCHUIIES_QXx?usp=sharing)

## B. FLASK CODE

```
import numpy as np
from flask import Flask, request, jsonify, render_template
import pickle
import pandas as pd

app = Flask(__name__) #initialising the flask app
filepath="model_movies.pkl"
model=pickle.load(open(filepath,'rb'))#loading the saved model
scalar=pickle.load(open("scalar_movies.pkl","rb"))#loading the saved scalar file

@app.route('/')
def home():
    return render_template('movieboxoffice.html')

@app.route('/y_predict',methods=['POST'])
def y_predict():
    # For rendering results on HTML
    input_feature=[float(x) for x in request.form.values()]
    features_values=np.array(input_feature)
    feature_name=['budget','genres','popularity','runtime','vote_average','vote_count',
                  'director','release_month','release_DOW']
    x_df=pd.DataFrame(features_values,columns=feature_name)
    x=scalar.transform(x_df)
    # predictions using the loaded model file
    prediction=model.predict(x)
    print("Prediction is:",prediction)
    return render_template("revenuepredict.html",prediction_text=prediction[0])
if __name__ == "__main__":
    app.run(debug=False)
```



## C.HTML FILES

### i. movieboxoffice.html

```
<!DOCTYPE html>
<html lang="en">
<head>
  <title>MOVIE BOX OFFICE GROSS PREDICTION</title>
  <style>
    body{
      opacity : 0.75;
      text-align: center;
      font-family: Verdana, Tahoma, sans-serif;
      font-size: larger;
      background-image: url('https://i.postimg.cc/6q9K4Hrm/filmpicture.jpg');
      background-repeat: no-repeat;
      background-attachment: fixed;
      background-size: cover;
    }
    h1,p {
      animation-duration: 3s;
      animation-name: slidein;
    }

    @keyframes slidein {
      from {
        margin-left: 100%;
        width: 300%;
      }

      to {
        margin-left: 0%;
        width: 100%;
      }
    }
    form{
      padding: 10px;
    }
    td{
      padding: 6px;
    }
    input[type=number]{

      padding: 8px 0px;
      margin: 8px 0;
      border: 2px solid #ccc;
      border-radius: 6px;
    }
    #btn{
      background-color: #52abf3;
      width: 100%;
```

```

border: 10px;
border-radius: 7px;
color: white;
padding: 15px 32px;
font-size: 16px;
cursor: pointer;
}
#btn:hover{
    background-color:#008bf9;
}
table{
    border-radius: 5px;
    background-color: #f2f2f2;
    padding: 20px;
}

#prediction {
color: white ;
}

input[type=text],select,button{
    width: 100%;
    padding: 12px 20px;
    margin: 8px 0;
    box-sizing: border-box;
    border: 2px solid red;
    border-radius: 4px;
}

</style>
</head>
<body>
    <!-- <div class="box"> -->
    <h1 style="background-color:powderblue;">MOVIE BOX OFFICE GROSS PREDICTION</h1>
    <!-- </div> -->

    <!-- <div class="box"> -->
    <center>
    <form action="{{ url_for('y_predict')}}" method="post">
        <table>
            <tr>

                <td>
                    <label for="budget">Budget</label>
                </td>
                <td>:</td>
                <td>
                    <input type="text" id="budget" name="budget" placeholder="Budget in $ Million"
required="required"/>
                </td>
            </tr>
            <tr>
                <td>
                    <label for="genres">Genres</label>

```

```

        </td>
        <td>:</td>
        <td>
            <select id="genres" name="genres" >
                <option>Select the genres</option>
                <option value="6">Drama</option>
                <option value="3">Comedy</option>
                <option value="0">Action</option>
                <option value="1">Adventure</option>
                <option value="10">Horror</option>
                <option value="4">Crime</option>
                <option value="16">Thriller</option>
                <option value="2">Animation</option>
                <option value="8">Fantasy</option>
                <option value="14">Science Fiction</option>
                <option value="13">Romance</option>
                <option value="7">Family</option>
                <option value="12">Mystery</option>
                <option value="5">Documentary</option>
                <option value="18">Western</option>
                <option value="17">War</option>
                <option value="9">History</option>
                <option value="15">TV Movie</option>
                <option value="11">Music</option></select>
            </td>
        </tr>
        <tr>
            <td>
                <label for="Enter popularity">Enter popularity</label>
            </td>
            <td>:</td>
            <td>
                <input type="text" id="Enter popularity" name="Enter popularity" placeholder="Enter the popularity"
required="required" />
            </td>
        </tr>

        <tr>
            <td>
                <label for="Enter runtime ">Enter runtime </label>
            </td>
            <td>:</td>
            <td>
                <input type="text" id="Enter runtime " name="Enter runtime " placeholder="Enter runtime"
required="required"/>
            </td>
        </tr>

        <tr>
            <td>
                <label for="Enter vote_average">Enter vote_average</label>
            </td>
            <td>:</td>
            <td>
                <input type="text" id="Enter vote_average" name="Enter vote_average" placeholder="Enter vote_average"
required="required"/>
            </td>
        </tr>

```

```

</tr>

<tr>
  <td>
    <label for="Enter vote_count">Enter vote_count </label>
  </td>
  <td>:</td>
  <td>
    <input type="text" id="Enter vote_count" name="Enter vote_count" placeholder="Enter vote_count"
required="required" />
  </td>
</tr>

<tr>
  <td>
    <label for="director">Director </label>
  </td>
  <td>:</td>
  <td>
    <select id="director" name="director" >
      <option>Select the director</option>
      <option value="2108">Steven Spielberg</option>
      <option value="2323">Woody Allen</option>
      <option value="1431">Martin Scorsese</option>
      <option value="377">Clint Eastwood</option>
      <option value="1851">Ridley Scott</option>
      <option value="1894">Robert Rodriguez</option>
      <option value="2051">Spike Lee</option>
      <option value="2107">Steven Soderbergh</option>
      <option value="1810">Renny Harlin</option>
      <option value="2169">Tim Burton</option>
      <option value="1654">Oliver Stone</option>
      <option value="1904">Robert Zemeckis</option>
      <option value="1930">Ron Howard</option>
      <option value="1034">Joel Schumacher</option>
      <option value="156">Barry Levinson</option>
      <option value="1480">Michael Bay</option>
      <option value="2234">Tony Scott</option>
      <option value="245">Brian De Palma</option>
      <option value="667">Francis Ford Coppola</option>
      <option value="1256">Kevin Smith</option>
      <option value="1973">Sam Raimi</option>
      <option value="2025">Shawn Levy</option>
      <option value="1823">Richard Donner</option>
      <option value="320">Chris Columbus</option></select><br>
    </td>
</tr>

<tr>
  <td>
    <label for="Enter the month of release">Enter the month of release</label>
  </td>
  <td>:</td>
  <td>
    <input type="text" id="Enter the month of release" name="Enter the month of release" placeholder="Enter
the month of release " required="required" />
  </td>

```

```

    </tr>

    <tr>
      <td>
        <label for="Enter the week of the month">Enter the week of the month</label>
      </td>
      <td>:</td>
      <td>
        <input type="text" id="Enter the week of the month" name="Enter the week of the month"
placeholder="Enter the week of the month " min="0" />
      </td>
    </tr>

    <tr>
      <td colspan="3">
        <input id="btn" type="submit" value="Predict" >
      </td>
    </tr>
  </table>
  <p id="prediction"> {{prediction_text}}</p>
</form>
</center>
<!-- </div> -->
</body>
</html>

```

## ii. revenuepredict.html

```

<html>
<style>
.idiv{
border-radius:10px;

}
body
{
background-image:url('../static/projector_image/lights.jpg');
background-repeat: no-repeat;

background-position: center;
font-family:sans-serif;
background-size:cover;
}
input{
font-size:1.3em;

```

```

width:80%;
text-align:center;
}
input placeholder{
text-align:center;
}
button{
outline:0;
border:0;
background-color:darkred;
color:white;
width:100px;
height:40px;
}
button:hover{
background-color:brown;
border:solid 1px black;
}
h1{
color:white;
}
h2{
color:lightyellow;
}
}
h1 {
text-shadow: 2px 2px 5px blue;
}
h2{
color: lightyellow;
}
h2{
text-shadow: 2px 2px 5px orange;
}
</style>
<head>
<title > Movie Box Office Gross Revenue</title>
</head>
<body>
<div class='idiv'>
<br/>
<h1 align="right">Movie Box Office Gross Revenue : </h1>
<br/>
<h2 align="right">The Revenue predicted is $ {{prediction_text}} million </h2>

<br/>
<br/>
<br/>
</div>

</body>
</html>

```