

**Project Report on EXPLORATORY DATA ANALYSIS OF MOVIES**

**B.TECH** *3RD YEAR*

**(Branch – CSE)**

**Submitted To: Submitted By:**

MR. SURAJ PATIL VED BHATKAR(B205)

(Assistant Professor) ARIHANT JAIN(B220)

VARAD NILAKHE(B238)

MPSTME, NMIMS Course: B TECH CS

Shirpur Campus Batch: 2019-2023

**ACKNOWLEDGEMENT**

This Project report was completed because of support from the groupmates, although not all of them can be mentioned.

We are greatly indebted to our good supervisor **Suraj Patil sir** for his/her useful and necessary observation, suggestions, contribution, and corrections. We would not have been able to achieve anything in this research without your supervision.

Ved Bhatkar Arihant Jain Varad Nilakhe

# INTRODUCTION

* The film industry plays a major role in the planetary or world- wide economy. It is the symbolic contributor to the global economy.
* Every year more than hundreds to thousands of movies are released to the public audience with the hope that the movies getting released will be the next block buster.
* According to the movie industry statistics, six to seven movies out of ten movies gets unprofitable, only one third of the movie gets success.
* The producers, studios, investors, sponsors in the movie industry are alike interested in predicting the box office success of the movie.
* This paper work is on analyzing the film genre, the releasedate around holidays, the release month of movies, the languages and country with more movies from the movie review dataset.
* There are attributes (country, languages, genre, movie release date, budget and revenue) taken from the dataset and the derived attributes (release monthof the movie derived from release date of movie and profit from budget and revenue) is analysed to determine the movie performance.
* The analysed data is plotted in graphs for statistical observation of the moviesuccess.

# Problem Statement:

* To automate the process of predicting whether the persom has liked themovie or not. In order to do so, we will be using Naïve Bayes Classification.
* Another objective of this study is to investigate that howmuch a movie is liked by the people so the production should do a remake of that movie or not .
* Our goal is to compare the movies which are more liked and less liked by the people and on the basis of that whatcan be done to remake it or not.

# TECHNIQUE USED

* It is a [classification technique](https://courses.analyticsvidhya.com/courses/introduction-to-data-science-2/?utm_source=blog&utm_medium=6stepsnaivebayesarticle) based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.
* For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on eachother or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as ‘Naive’.
* Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

# ALGORITHM

* Step 1: Convert the data set into a frequency table
* Step 2: Create Likelihood table by finding the probabilities like Overcastprobability = 0.29 and probability of playing is 0.64.
* Step 3: Now, use [Naive Bayesian](https://courses.analyticsvidhya.com/courses/naive-bayes?utm_source=blog&utm_medium=naive-bayes-explained) equation to calculate the posterior probability for each class. The class with the highest posterior probabilityis the outcome of prediction

# ABOUT THE DATASET:

* + The data contains the name of the movies, their release period, type of genre, music director, the movie director, the budget of the movie and the revenue of the particular movie and the number of screens its beingplayed.

## IMPLEMENTATION

Write the steps with each code and provide output snapshot.

# Source Code:

#a Importing the dataset library(dplyr)

movie<-read.csv("D:\\DM\\PROJECT\\movie.csv") View(movie)

str(movie)

summary(movie) #Data Pre Processing

#b Handling the missing data movie$Budget.INR. =

ifelse(is.na(movie$Budget.INR.),ave(movie$Budget.INR., FUN

= function(x) mean(x, na.rm = 'TRUE')),movie$Budget.INR.)

movie$Revenue.INR. = ifelse(is.na(movie$Revenue.INR.),ave(movie$Revenue.INR., FUN = function(x) mean(x, na.rm = 'TRUE')),movie$Revenue.INR.)

#Designing the histogram hist(movie$Number.of.Screens, col = "red")

#Designing All the Plots library(ggplot2)

#9

#a Display bar chart ggplot(data=movie,aes(Genre))+geom\_bar()

#b Design Histogram

ggplot(data = movie, aes(Budget.INR.))+geom\_histogram(bins

= 10)

#c Design Scatter Plot

ggplot(data =movie, aes(x = Budget.INR., y = Revenue.INR., col = Genre))+geom\_point()

#d Design Box Plot

ggplot(data =movie, aes(fill = Genre, x = Budget.INR., y = Revenue.INR.))+geom\_boxplot(notch=TRUE)

#Factoring target feature

movie$Whether.Remake = factor(movie$Whether.Remake, levels = c('No','Yes'),

labels = c(0,1))

#Transforming data into class variable Remake\_class<-ifelse(movie$Whether.Remake=="1", "Remaked", "Not Remaked");

movie <- data.frame(movie, Remake\_class) View(movie)

#Removing Whether.Remake from Data Frame movie <- movie[,-3]

#Splitting the Dataset library(caTools) set.seed(2)

id <- sample(2, nrow(movie), prob=c(0.7, 0.3), replace=TRUE) print(id)

#Training and Testing movie\_train <- movie[id==1,] movie\_test <- movie[id==2,] print(movie\_train)

#Building the model library(e1071)

model <- naiveBayes(as.factor(Remake\_class)~., movie\_train) print(model)

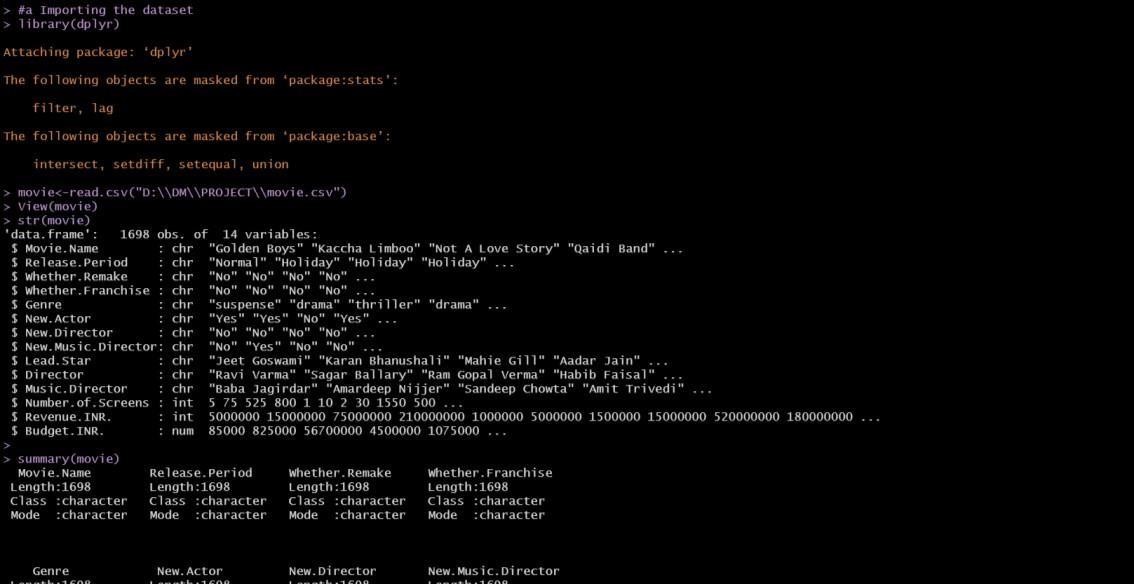
pmodel <- predict(model, movie\_test)

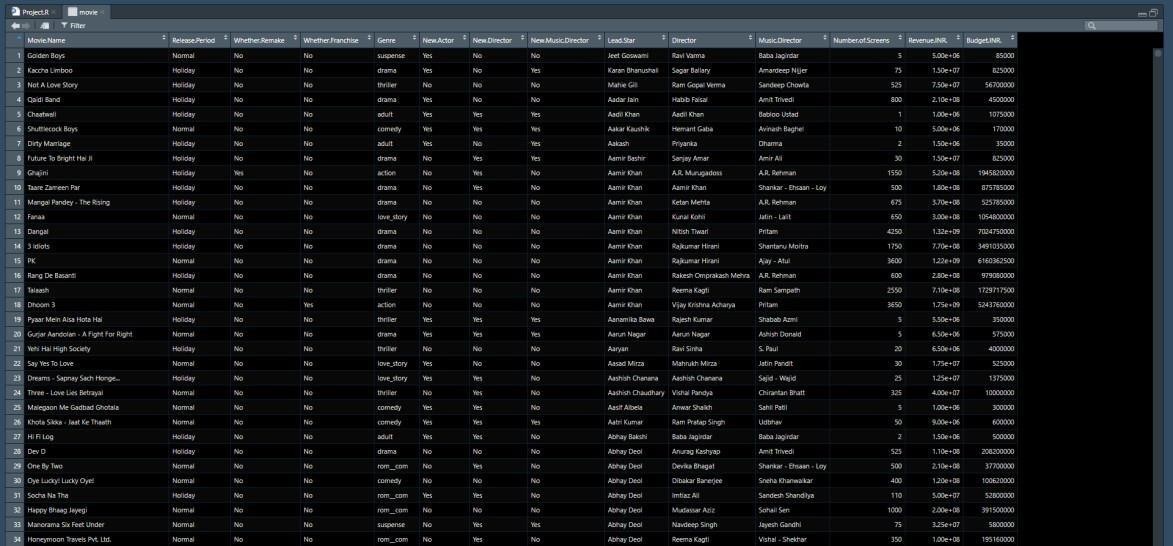
#Building the confusion matrix table(pmodel, movie\_test$Remake\_class)

**OUTPUT SNAPSHOTS:**

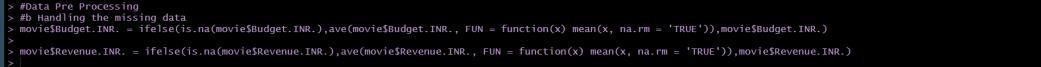
1. **Importing dataset**

Firstly, we will import the dataset.





## Data Preprocessing



1. **Data Visualization**

**Plots:**



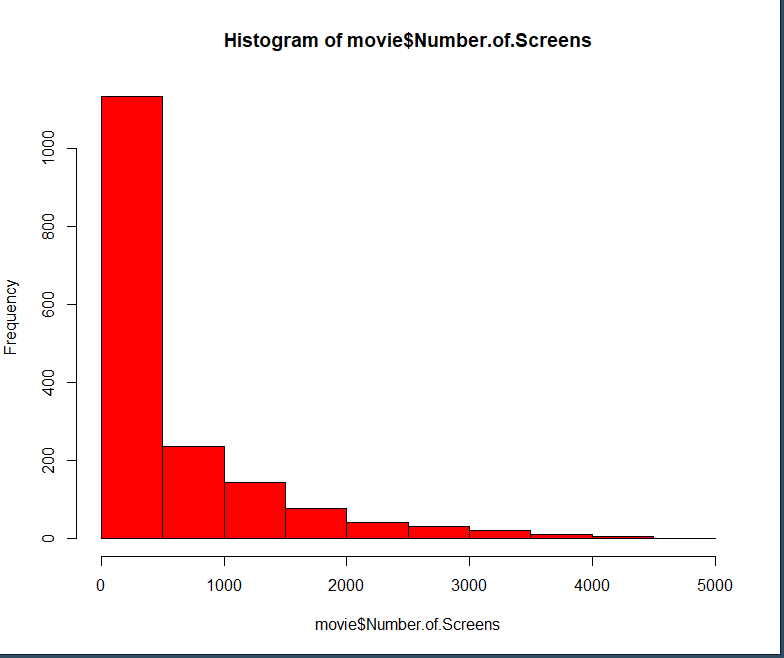


FIG-1 HISTOGRAM

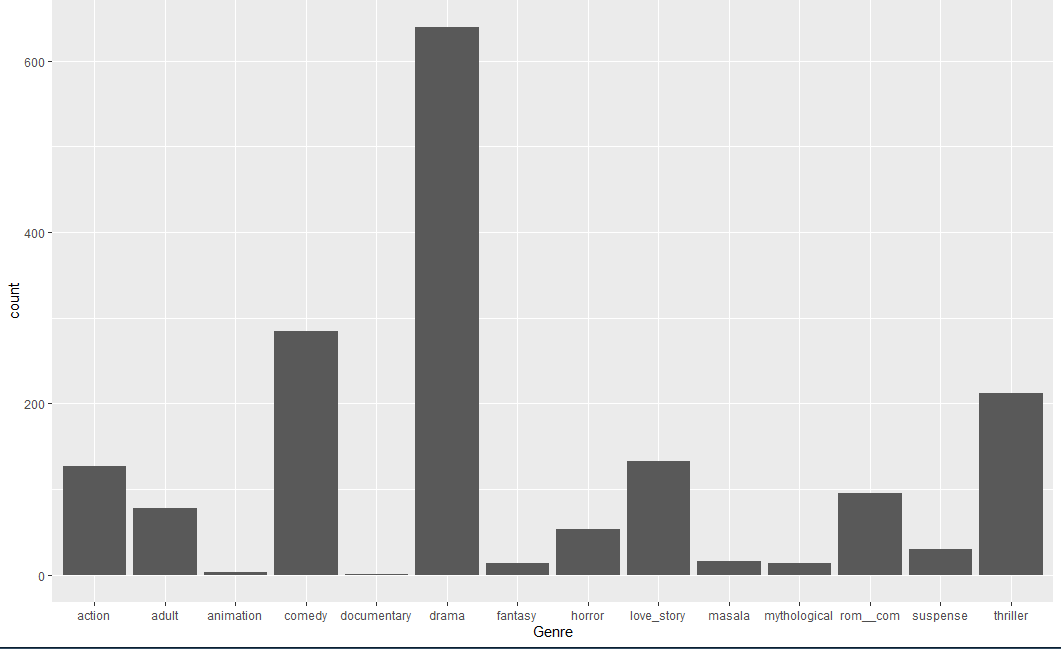


FIG-2 Bar Chart



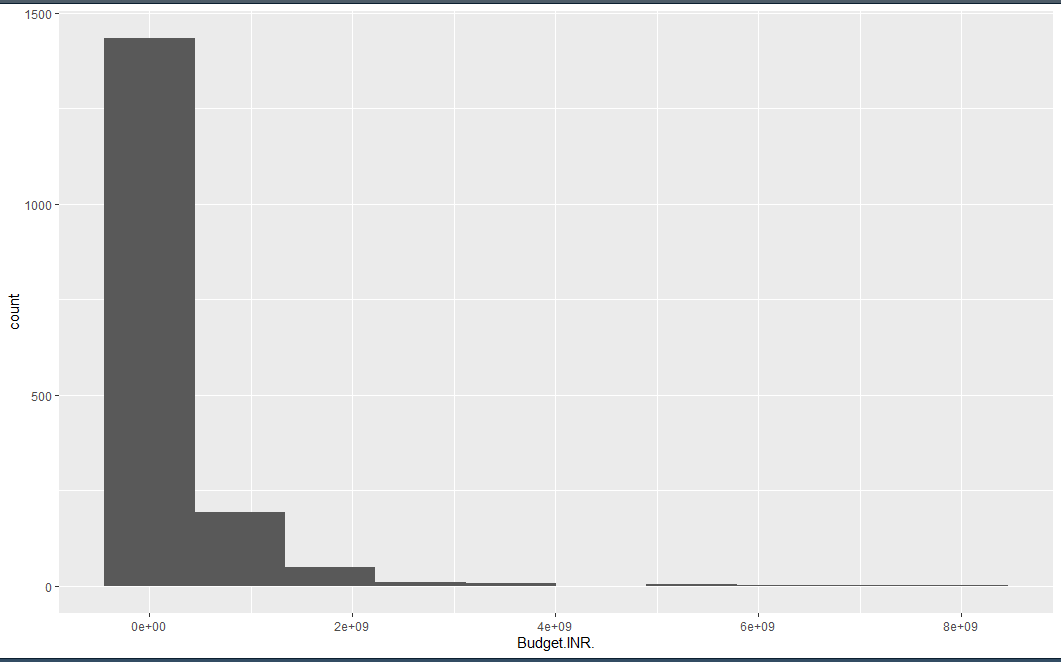


FIG-3 Histogram



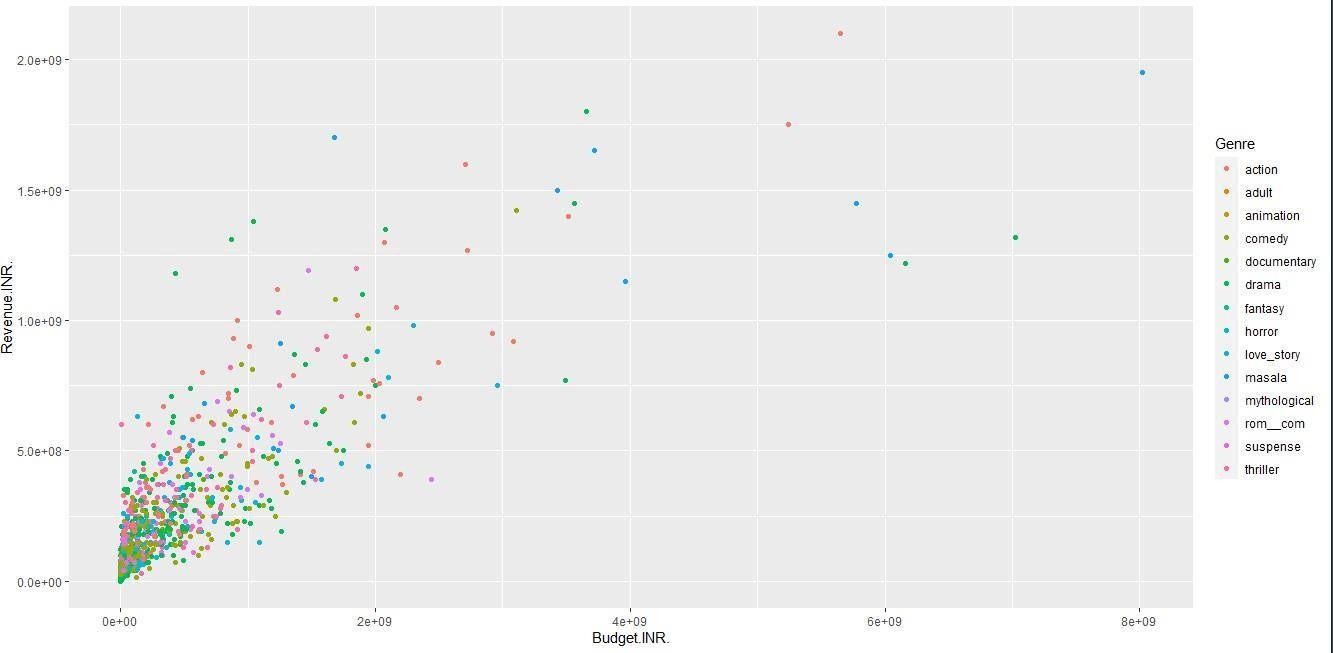


FIG-4 Scatter Plot



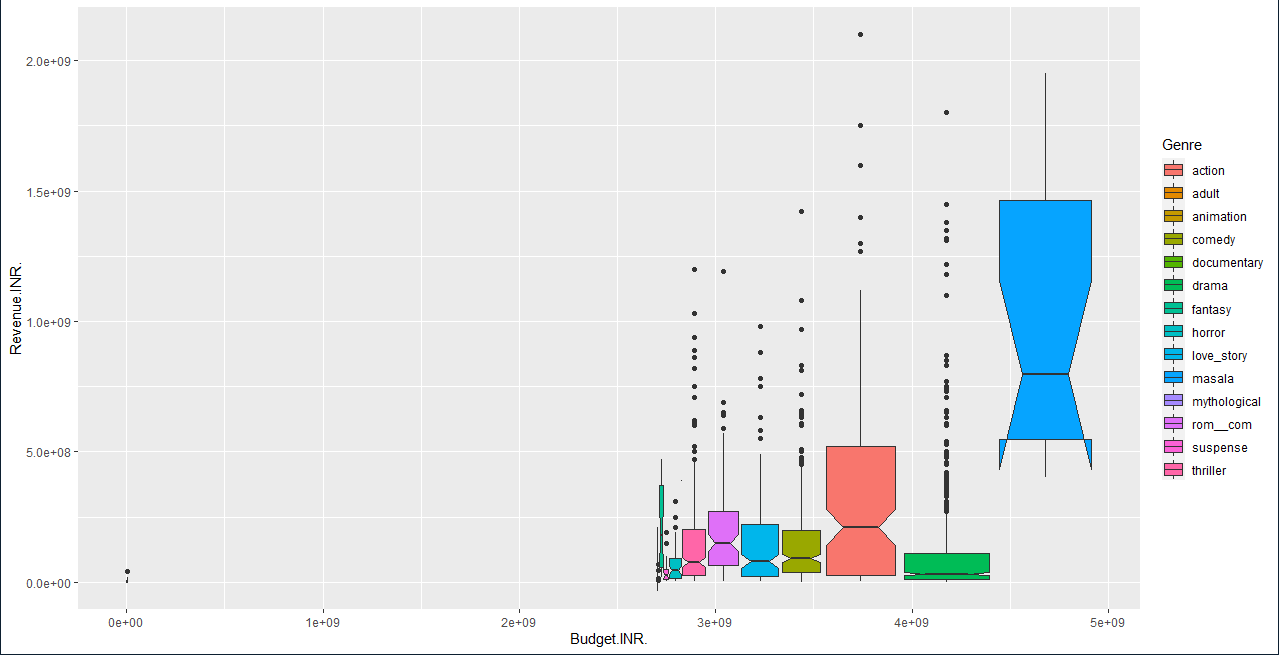
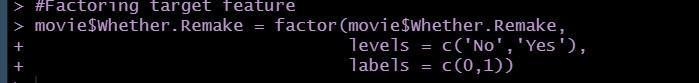
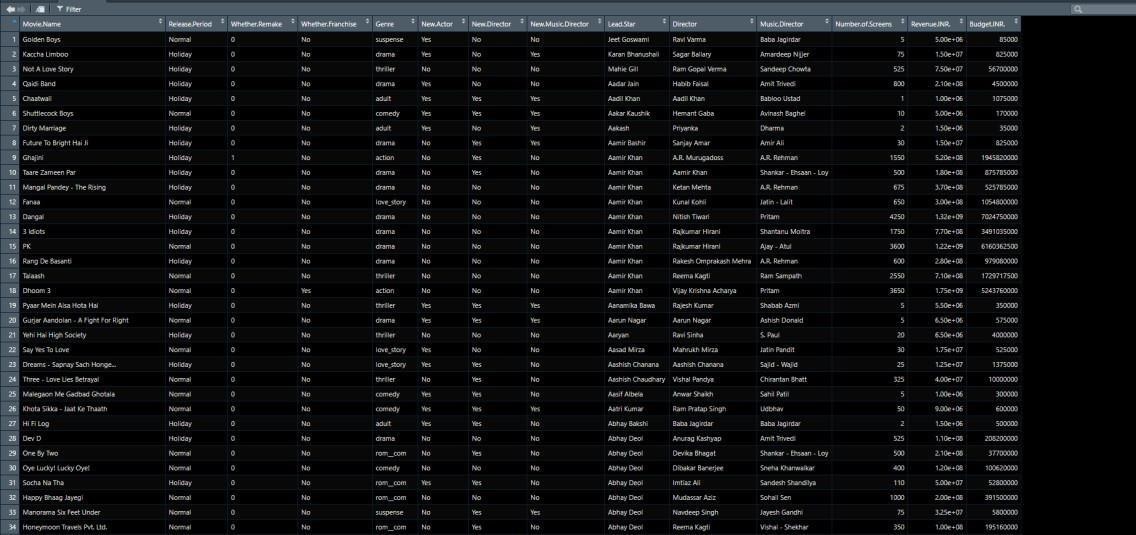
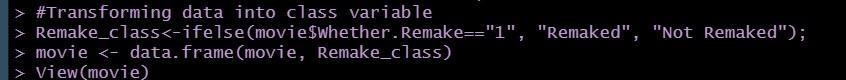


FIG-5 Box Plot











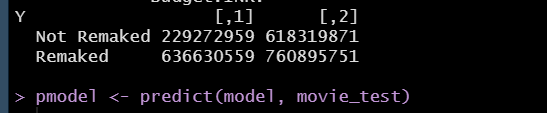
## Splitting the Dataset



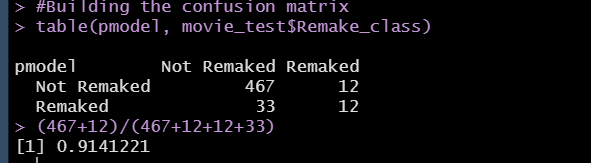
1. **Training and Testing**



1. **Building the model**



1. **Building the Confusion Matrix**



## ANALYSIS/CONCLUSION

* We concluded and analyzed whether the movie will be a remake or not according to the number of screens it was played as the people liked it and the revenue made by the movie by the help ofthe model.
* Movie played on more the number of screens more the revenue made by it.

# REFERENCES

1. <https://en.wikipedia.org/>
2. <https://www.kaggle.com/>