BANGLA MOVIE SUCCESS PREDICTION USING MACHINE LEARNING

BY

Sonia Nasrin

ID: 152-15-5522

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering.

Supervised By

Syed Akhter Hossain

Professor and Head

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

MAY 2019

APPROVAL

This Project titled **"Bangla Movie Success Prediction Using Machine Learning"**, submitted by Sonia Nasrin, ID No: 152-15-5522 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 3rd May, 2019.

BOARD OF EXAMINERS

Dr. Syed Akhter Hossain
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman

Md. Tarek Habib
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Moushumi Zaman Bonny
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Swakkhar Shatabda
Associate Professor
Department of Computer Science and Engineering
United International University

External Examiner

DECLARATION

I hereby declare that, this project has been done by me under the supervision of **Syed Akhter Hossain, Prof and Head, Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Syed Akhter Hossain
Designation
Department of CSE
Daffodil International University

Submitted by:

Sonia Nasrin

Sonia Nasrin
ID: 152-15-5522
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project successfully.

I really grateful and wish my profound my indebtedness to **Syed Akhter Hossain**, **Prof and Head**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of "*Data Mining under Machine Learning*" to carry out this project. His endless patience, scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

I would like to express our heartiest gratitude to the Almighty Allah and Head**,** Department of CSE, for his kind help to finish my project and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank my entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

In present world, machine learning techniques and models are used to predict future. Prediction the movie success before its release has been an immense point of concern for movie industry related all people, especially producer, shake holders and director. Since Bangladeshi movie industry is in threat, they need some kind of assurance that the movie will be successful or not and which factor can improve the profit. The purpose of the work is to make a model which predict Bangla movie success depending on some pre-release factors like actor, actress, director, producer, genre, budget, release date, duration, playback singer, music director that helps Bangladeshi movie industry to determinate specific reason for success so that they can resolve before release.

In this model, data mining process and algorithm are applied for movie classification. Decision tree, Random forest, Logistic Regression, Support Vector machine are used and evaluated on this dataset and also focus on to find out some interesting relationship between features using feature engineering techniques and tools. Here, it is supervised classification which classify movie based on IMdb movie rating, where rating are divided into five classes (flop, bad, watchable, super hit and block buster). I visited all Bangladeshi movie related sites to collect data. As Bangladeshi movie data collection is highly unorganized and unavailable, our dataset is not much more longer. This prototype model has been provided much better performance in this challenging scenario.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLES

Chapter 1

Introduction

1.1 Introduction

In today's world, every sector wants to grab machine learning techniques for their betterment. Machine learning introduce an amazing sector called artificial intelligence and the data mining leads basic level of these techniques. Data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data [1]. It has also opened up exciting opportunities for exploring and analyzing new types of data and for analyzing old types of data in new ways [1]. It can find out hidden relationship of various features (attributes) and features dependency that helps for classification and indicates the feature importance. Data mining technique is a very popular method which is used in countless scenarios such as customer relationship, e-commerce, finance, banking, medical, industry, statistics and so on. Recommendation system using data mining techniques for classification, clustering, predictions. This technique's required lot of data for investigation. The more data we have, the more accuracy we can gain in model. Attributes of data is studied and analyzed very well to gain higher accuracy. Several techniques like F1 measure, Recall, Precision, confusion matrix are used to shoe the accuracy and failure rate of a model.

Using data mining techniques in movie data helps to find out the reason of downward trend of movie quality and helps to rise up from recessional condition that helps movie marker to make successful movie. Producer and shareholder around the world is highly concerned about the success of film. The rate of movie success has a highly impact in actor, actress, director career as well. In this work, the most challenging part was data collection because of unavailability.

I mainly focus on IMdb[2] because it has rich information about movie and its features and movie rating is given by many people so that rating is more accurate. And I also use others sites like: Wikipedia [3], Box office Bangladesh [4] and BMdb [5] sites for collecting data.

I try my heart and soul to prepare dataset and extract information from the data. I found out some interesting information pattern and relation between data. Like, watchable movie doesn't depend on budget, our movie quality is increased than last decent, most of the movie is release in the last part of the year, actor-actress play a vital role and so on which details in step by step.

1.2 Motivation

Movie is the one of the best entertainment source of people. Many researchers in abroad did many analyses in movie database which enhance their movie industry directly. But in Bangladesh this scenario is totally different. The saddest part is that, though movie is the reflection of people thought and create a vital impact on people, our movie quality is going to downward day by day. But nobody worries to save this sector and none are doing any research on Bangla movie. So this scenario highly influences me to start the initiative. That's why I chose the topic for research.

1.3 Rationale of the work

Bangladesh is improving in technological sectors day by day. But movie industries are in threat because of flop movie. People lose their interest to see movie in theater. But once we had a golden era of our film industry. But last 20 years this site impairs its significations. Most of the theaters are shut downed already. People, depending in this sector are being unemployment. So I try to find out the lacking of this downward process using data correlation technique.

Nobody do work in this sector, so its highly motivate me to do the research of the sector. I am probably the first one to do research in Bangla movie site. So it's a very much difficult to manage data and I try my best. Hope that it will help our industry to rethink. I think, movie maker (like director, producer, shareholder, actor-actress) will be benefited by this work.

1.4 Research Questions

I already told the importance of this research. During this work, I try to find out some answers of particular questions. Here they are:

1. Why people lose interest in Bangla movie?
2. Which factors are important for movie success?
3. How can we collect and manage data?
4. How can we handle those data?
5. Which algorithm can give better result?

1.5 Expected Outcome

In this work, I want to find out a movie's success category before its release depending on some particular pre-release factors. The performance of this prototype model will be evaluated currently that we have a construction process of usage of different kinds of ML algorithms which indicates the comparison results depending on the dataset.

This research work will deliver hidden relation between feature and good prediction result. Whether it is good or bad or average, movie maker and also viewer will be benefited and decision making process will be very easy for them.

1.6 Report Layout

Chapter 1 Discusses about data mining, thesis motivation, Rationale of the Study, Research Question and Expected Outcome.

Chapter 2 Focus on the Background history of the work. Also give the idea of related works. Problems and challenges are also mentioned in this part.

Chapter 3 Mention approximately the tools and techniques of my research. It discusses a short view of algorithms, machine learning techniques. Data collection process, data mining process, analysis of data and visualization are also discussed here.

Chapter 4 Focus on the result of the dataset, outcome and evaluation of the algorithms.

Chapter 5 Provides short view of the work. Discuss the future scope of the work that will be performed to enhance my work.

Chapter 2

Background

2.1 Introduction

Many research work is done upon this sector in abroad. Some of them are using text documentation, some are using image processing, and some are using data mining. Here the sample are the dataset is consisting of numeric and categorical values. So data mining process will be applicable in this work.

2.2 Related Works

In 2004, Saraee, MH, White and Eccleston performed prediction of movie rating based on online platform IMdb where contains resources more than 390,000 movies and television shows [6]. They mentioned that their main problem was to extract information from IMdb documentation as the format of the source data is plain text. They use Hollywood data and they first concluded that budget is not a important factor for movie success and there is a downward trend in the movie quality. To find the success category (like, flop, super hit) of movie, mainly classification approach is used. Data hidden pattern, facts and relationship between features can be focused by using data mining and its influence model directly for better accuracy. Data amount and accessibility highly influence in this process as I said before that the more we have, the more accuracy we can gain. However, data access is difficult for copy right issue. The work by Zhang and Skeina worked on new analysis for improving accuracy rate in prediction model [7]. They got better performance for using both IMdb and new data. In 2015, Karl worked to find out a comparative study between two demanding algorithm named Random forest and SVM [8]. The result difference between of two algorithms is minor. Kabinsingha, Chindasorn, Chantrapornchai talk about how they used decision tree with T-score to find out rating of movie and give the suggestion [9]. Hasan and Hammad discussed about how to follow the steps of data preprocessing, data extraction, integration, analysis and feature selection and finally classification [10].

2.3 Research Summary

The research study of the related work ensures me that there is no work on Bangla movie. I can also learn that our research is about Supervised Machine Learning where target values are included in train dataset to learn machine and its classification based problem.

In my work, I use several algorithms (KNN, SVM, RF, LR, Decision tree) to see the behavior of multiple algorithm on my dataset. I have done my research using both clustering and classification approach. Other research use a large number of data to train and test purpose. I spilt my dataset into train (75%)and test (25%) to make the model but my data is not too much because there is no data available.

2.4 Scope of the Problem

The prediction based research work is totally new in Bangladesh on Movie dataset. As our movie industry is destroyed day by day and loses its beauty so the there is a big opportunity of this model to help director, producer and also viewer to find out the essential attribute to success any movie. As the dataset is unique and totally new so there is a huge opportunity to work and extract more and more amazing information by doing research.

2.5 Challenges

- Collecting dataset
- Create model based on this complex data
- Find out correct spelling of a value
- Ensure data quality
- To solve error in model
- Increase accuracy rate

Chapter 3

Research Methodology

3.1 Introduction

As Bangladesh is backward in technology and movie industry faces recessional condition in last docents, data is not available in any platform like online or under organization. It's an unimaginable difficult situation to extract the authentic and sufficient data. I spend most of the time for data collection for this project. It is spent more than 30 minutes for a single sample entry. I recheck every information in several time just for collecting accurate and authentic information.

3.2 Research Subject and Instrumentation

Previously I said, my research domain in machine learning using data mining. In this section I discuss about machine learning and data mining process. Also give a short introduction about algorithms, those I use.

3.2.1 Machine Learning

Machine learning is the technique that is used by device to learn intelligently and automatically without any help of programmers. It is the combination of mathematics, statistics and algorithms.

Machine learning algorithms are often categorized into several groups. They are:

Supervised learning algorithm try to predict the value of unseen or unknown target feature from a bunch of learning data where the target is known. It learns pattern from training data that has target variable and predict the test data which does not have label (target) variable and finally evaluate the result to find out the accuracy percentage. Our work is under the section

Unsupervised machine learning algorithm are applicable where we do not have classifier no labeled data in our dataset. It tries to discover the hidden structure and features from unlabeled dataset.

Semi-supervised learning algorithms deals in between supervised and unsupervised learning, as dataset has both labeled and unlabeled data in training sectors. Generally, it contains small amount of label or target data and a large number of unlabeled data. The system which uses this method are able to improve accuracy.

Reinforcement machine learning algorithm is learning method that interact with its environment by producing actions and discovers errors or rewards [11]. In this article shows, trial, error search and producing operations and discovers errors and rewards and it allow software to automatically determine the ideal behavior within a specific context in order to maximize its performance [11].

3.2.2 Data Mining

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems [2]. It retrieves the information from raw dataset. In today's world billions of data is generated per seconds. Information is hiding in raw data. Data is not perfect as well. Missing, outliers, duplicate and dummy data are mixed with informative data. In data mining process, this unusable data is through away and retrieve knowledge. Discovering the knowledge is like that:

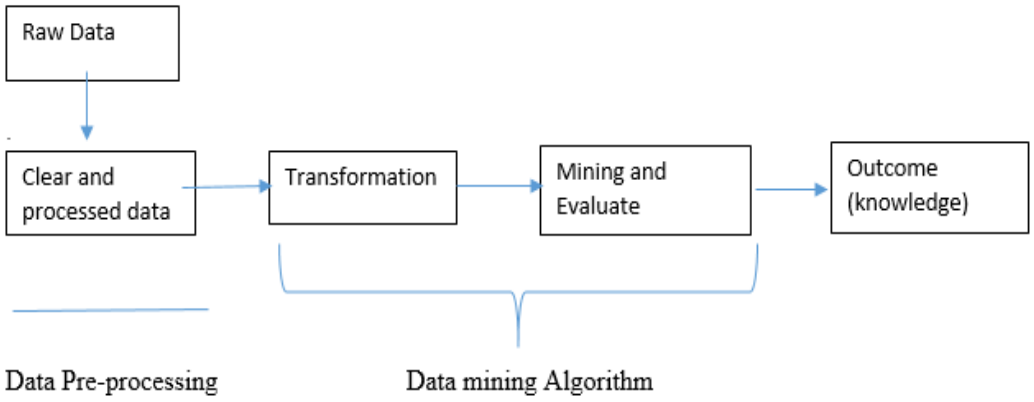


Figure 3.1 KDD (Knowledge discover databases) process in data mining.

3.3 Data Collection Procedure

Data collection is one of the most challenging task of my work. Because of lacking information in internet, lacking of trusted site it becomes challenging and tough. Sometimes I have use VPN for accessing in a particular site for gathering data. First, I have to find the movies name that release in particular year. Then I collect some data about particular movie from Wikipedia and IMdb because they contain rich information. Movie name, music director, genre, playback singer, release time are easy to find out in Wikipedia and IMdb but main problem I faces to collect data of movie rating, budget, movie success and award. For collecting those attributes, I search almost all relatable link found in google. I like to mention some sites name with dependency in below:

- IMDB (dependency 50%)
- Wikipedia (dependency 25%)
- Box Office Bangladesh (dependency 15%)
- BMdb (dependency 10%)

Sometimes I used VPN for accessing in a particular site as this site is not available in my regular internet connection. Sometimes I had to check a particular data (like: Rating and budget, gross) in different site because there are some variation of this particular data in different site. Also spelling of a particular name in movie cast is different in different sites. So I had to check multiple time to ensure the variation of the name indicates different person or not. So It took 35-45 minutes to gather a sample (single movie data) features. So its cost me like 85-90 hours to gather the data of 100 movies. But unfortunately most of the important features are not found in any site like (gross, profit, budget). Some data is not included due to lacking of trusted source as well as information. So I have to drop those features which have massive missing data. Lacking of important data is also a curse or my model for lower accuracy.

3.4 Statistical Analysis

Statistical analysis of dataset is a process of performing different statistical operations which gives us a short of overall description of data at a glance. Its operation depends on research questions and hypotheses. It seeks quality and quantity of data. Statistical analysis will be: ANOVA, ANCOVA, MANOVA and MANCOVA. If it is relationship based, then the analysis will be correlation, regression, covariance-based etc.

In my work, I use describe and correlation method for survey and observation of data using panda library of python language. In section 3.5.3 it is described in details.

3.5 Implementation Requirements

To achieve my goal in this research I use several steps that are connected to one another for extracting information and creating model. Working processes are:

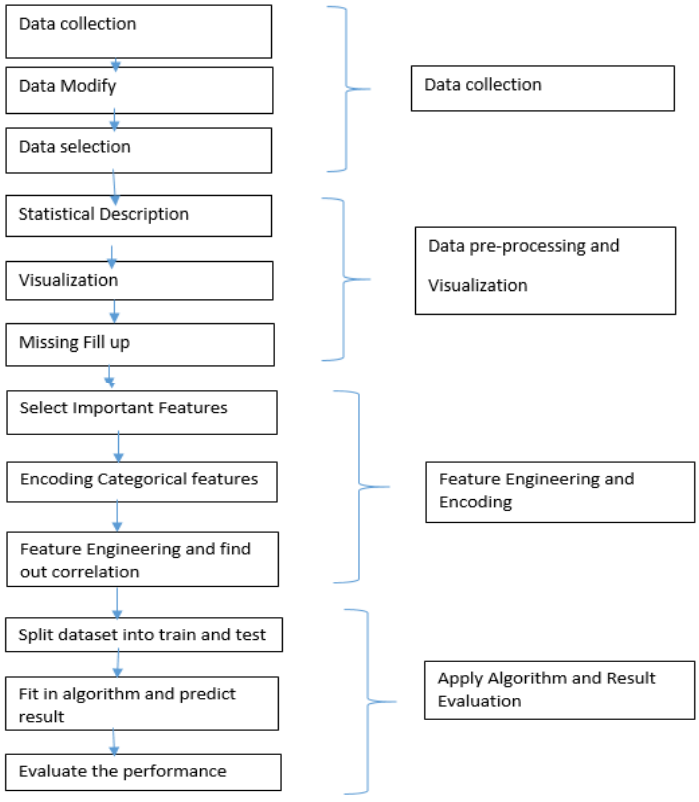


Figure 3.2 Working process for model creation.

I am working with Python language in Jupyter (web based platform). I use pandas for data preprocessing, matplotlib library for visualization and scikit learn for analysis complex data.

3.5.1: KDD process in Data Mining

Density of data is increasing day by day, but its often far to get perfect data. Data is mixed up with error, inconsistency, anomalies which can biased and misguided actual information. So KDD process helps to manage the data for getting actual information.

Steps of KDD Process:

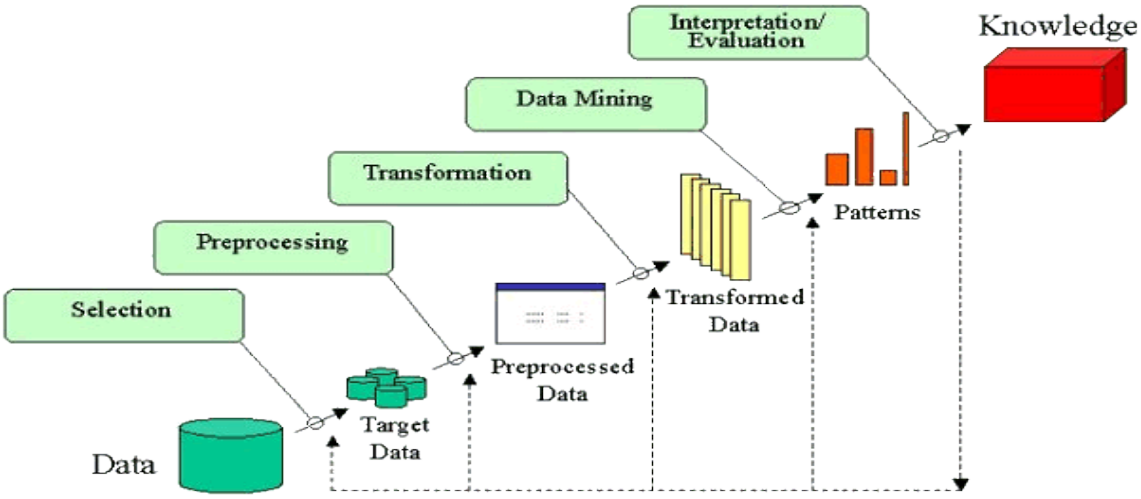


Figure 3.3 KDD process in DM

- **Data Cleaning:** Detect and eliminate error, inconsistency, anomalies of data.
- **Data Integration:** Blending information
- **Data Selection:** Recover and select data for analysis
- **Data Transformation:** Transfer data in machine suitable format so that machine can read and operate
- **Data Mining:** Mining data using intelligent process and methods for bringing out pattern.

- **Knowledge Representation:** Represent knowledge in matrix, diagram bar chart, table or other suitable process.

3.5.2 Pre-processing of Data

In data collection phase, I avoid confusing data and try to include correct spelling of name of several features. For proving of authentication I also add link in my data at each sample. All processes are already discussed in section 3.3.

After data collection, at first Excel file is created in Comma Separated Value (CSV) format.

I use Python and Jupyter as mentioned before. Using pandas library for manipulate data. I load necessary libraries and dataset in the platform.

I drop 'link' feature as it has no contribute in machine learning. I avoid gross feature as its confusing and lot of missing values. I do not use 'movie classification' feature as its contradict with rating because rating value is more real data that is collected from IMdb where hundreds or thousands on viewer rate a movie so it's more authentic. I classify this rating in five classes in our model later.

Scikit Learn in python is the most beautiful tool for any machine learning work as well.

3.5.3 Analysis phase

First we calculate the amount of missing value using python library. Here we can see that budget and award feature have high number of missing values. As few movies are being awarded for their performance so this missing is natural. But due to unavailability of 'budget' data, we have huge missing. After considering the importance of the feature for Bangla movie dataset, I don't drop it. And including this, other missing values are fill up with suitable methods, details in section 3.5.5. Figure 3.4 indicates the number of missing value.

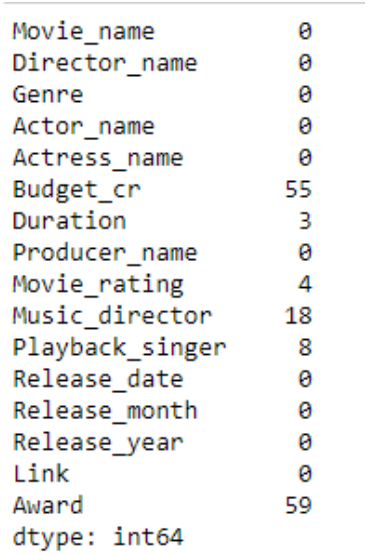


Figure 3.4 Show the missing value amount of dataset.

Secondly, for performing analysis we have to clear the datatype of our dataset features. Figure 3.5 is showing this amount:

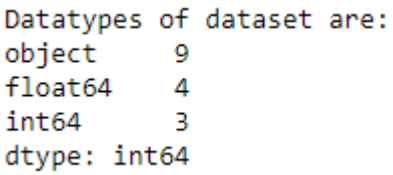


Figure 3.5 Amount of various datatype.

Data description is necessary to find out data anomalies at a second. The short details of data 'describe' methods is given in section 3.4. Figure 3.6 shows my dataset description of numeric values only:

	Budget_cr	Duration	Movie_rating	Release_date	Release_month	Release_year	Award
count	44.000000	96.000000	95.000000	99.000000	99.000000	99.000000	40.000000
mean	2.703182	138.500000	6.678947	14.242424	7.111111	2013.202020	4.700000
std	1.696080	20.157799	1.354339	8.514203	3.251025	4.750896	3.531543
min	0.300000	85.000000	3.500000	1.000000	1.000000	1993.000000	1.000000
25%	1.500000	126.750000	5.700000	7.000000	4.000000	2011.000000	2.000000
50%	2.250000	145.000000	6.800000	13.000000	7.000000	2015.000000	4.000000
75%	4.000000	151.000000	7.800000	20.000000	10.000000	2016.000000	7.000000
max	6.500000	179.000000	9.300000	31.000000	12.000000	2018.000000	17.000000

Figure 3.6 Description of Bangla movie dataset

Descriptive statistics show the following information:

- Frequency: count the number of data in each feature.
- Min: Minimum value of particular feature
- Max: maximum value of particular feature
- Mean: average value of features
- Std: standard values of particular features
- Mode: most frequent value of a particular feature
- 25%, 50%, 75% percentile of data which indicates the outlier of the particular features.

To evaluate and gather the idea of data dependency of one feature to another, correlation is an awesome technique. Though our dataset is not much rich so correlation value can't be high as natural.

	Budget_cr	Duration	Movie_rating	Release_date	Release_month	Release_year	Award
Budget_cr	1.000000	0.209438	-0.188008	0.258709	-0.061236	0.370586	-0.496869
Duration	0.209438	1.000000	-0.523650	-0.048464	-0.031491	0.045153	-0.177988
Movie_rating	-0.188008	-0.523650	1.000000	0.115807	0.073325	-0.109519	0.424317
Release_date	0.258709	-0.048464	0.115807	1.000000	0.053577	0.174857	-0.079936
Release_month	-0.061236	-0.031491	0.073325	0.053577	1.000000	-0.186453	0.053703
Release_year	0.370586	0.045153	-0.109519	0.174857	-0.186453	1.000000	-0.159812
Award	-0.496869	-0.177988	0.424317	-0.079936	0.053703	-0.159812	1.000000

Figure 3.7 Correlation table of dataset

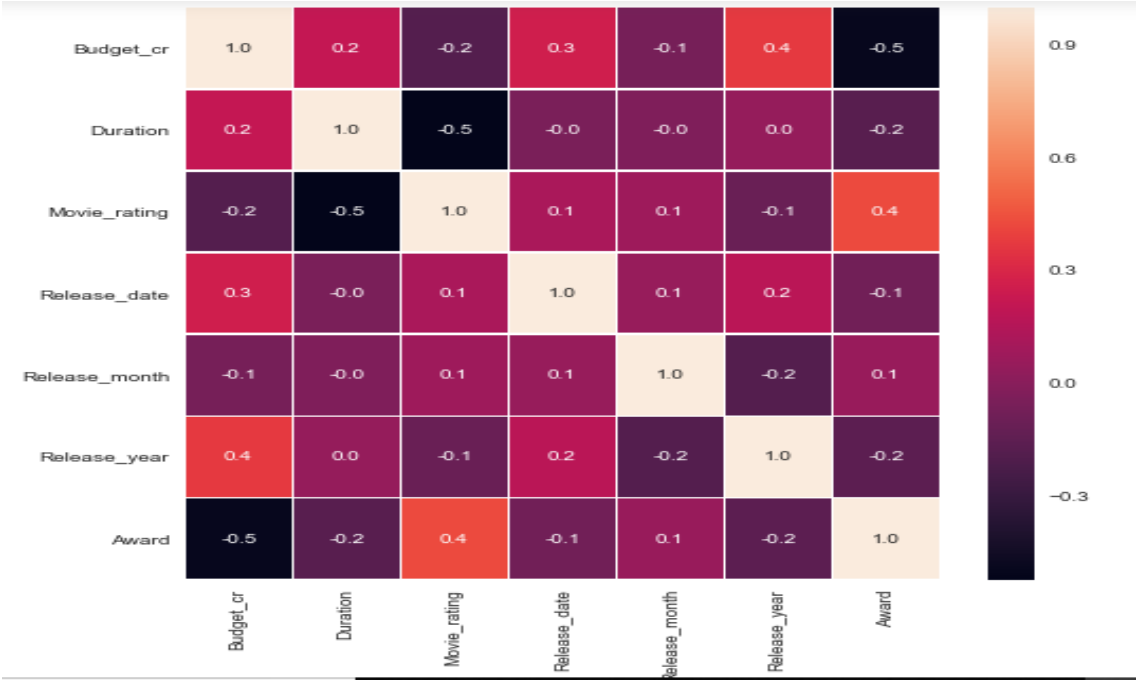


Figure 3.8 Correlation plotting using seaborn library.

Correlation measure the relationship between two features in range of -1 to +1. +1 means total dependency between that particular feature, like same features correlation is +1 and our Bangla movie dataset also indicates that. '0' means there is no correlation and negative values means the negative correlation between features. In our dataset, we can find out easily that the positive correlation is existing between budget and duration, release date

among with movie rating and budget of the movie, award among with movie rating. As I mentioned that this correlation is calculated only on numeric values. Categorical values importance will discuss in visualization section 3.5.

While analysis dataset, we see every features (mainly categorical) have a distinct number of unique values. Its almost 70% up . In figure 3.5.3.5(1-3) shows some features variation (highest) amount for clear it. Other feature has less than 50 unique values.

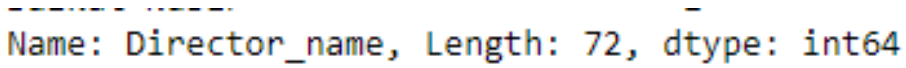


Figure 3.9 Unique values of Director_name feature

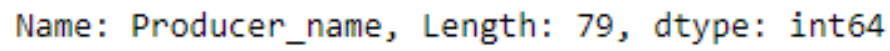


Figure 3.10 Unique values of Producer_name feature

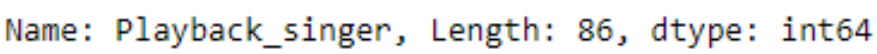


Figure 3.11 Values variation of Playback_singer feature

Using the mean method upon categorical variable, we can get idea that which person's (actor, actress, producer, director) is more successful than others. If success rate is high, then it increases the weight of that particular values.

3.5.4 Encoding Method for categorical values

Machine can read and operate only numerical value. So we need to translate data in machine suitable process for executing. I have large categorical values in my Bangla movie dataset. Its compulsory to encode them with most suitable process. There are several process in encoding methods. They are:

- Label encoding: Machine labeled the categorical data into numeric integer number. It is easy process but it has disadvantage. The main disadvantage is that numeric integer has a order form and machine uses this order form information later. Like 1<2<3<4…. etc that previous value is smaller than the next value. In categorical features it is not suitable or legal. So its misguided machine. The output type of this method is data frame.

- One Hot encoding: Machine create a separate column for each unique value. So that number of column will be very large but it's the safe method than Label encoding.

	Name	Generation	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6
4	Octillery	Gen 2	0	1	0	0	0	0
5	Helioptile	Gen 6	0	0	0	0	0	1
6	Dialga	Gen 4	0	0	0	1	0	0
7	DeoxysDefense Forme	Gen 3	0	0	1	0	0	0
8	Rapidash	Gen 1	1	0	0	0	0	0
9	Swanna	Gen 5	0	0	0	0	1	0

One-hot encoded features by leveraging pandas

Figure 3.12 One hot encoding example using game generation feature.

- Get Dummy variable: It's a straight forward method in Pandas library. It is just like one-hot encoding but it generates (m-1) features that mean one feature of encoding scheme will be eliminated. It has a built in function in Pandas, we have to initiate and call the function. It is more easy to manage. The main advantage is it can directly use in data frame and algorithm can recognize it automatically. In our work, we use this methods of transform process of categorical values. Figure 3.5.4.2 will show a sample of this method after apply

]:

Award	Movie_name_Agnee-2	Movie_name_Aguner Poroshmoni	Movie_name_Amar Bondhu Rashed	...	Playback_singer_Shafiq Tuhin|Ahmed Imtiaz Bulbul	Playback_singer_Shammi Akhtar|Mita haque	Playback_singer_Shouquat Ali Imon|Imran|Shochi Shams|Porshi|Shahin|Tahsin	Playback
0.0	0	0	0	...	0	0	0	
0.0	0	0	0	...	0	0	0	
0.0	0	0	0	...	0	0	0	
0.0	0	0	0	...	0	0	0	
7.0	0	0	0	...	0	0	0	

Figure 3.13 our dataset (first 5 samples) transformation using get dummy encoding.

3.5.5 Feature Engineering

Feature engineering is the most important part of DM which has a positive influence in model accuracy.

Firstly, As our model is classification base then I have been binned movie rating into several classes for make it easy to handle. IMdb rates movie between 1-10. Table 3.5.5.1 is given below:

Table 3.1 Shows movie success class depends on rating

New Class	Movie rating range	Class of movie success
0	0.0 - 4.0	Flop
1	4.0 - 5.9	Bad
2	5.9 - 7.0	Watchable
3	7.0 - 8.0	Super hit
4	8.0 – 10.0	Block buster

Secondly, to fill up missing values depending on statistical analysis, I fill ‘Movie_rating’ feature with the mean value, Award with ‘zero’, budget with median and categorical

features with mode values. I use median because mean value is highly effected by outlier. In that case, median is the best method for use. So which feature has outlier, that is filled up median value for increasing accuracy.

3.5.6 Visualization

Now I like to share the visualization diagram of features that I found in my model which give a clear concept of data.

The frequency of budget values are shown in figure 3.5.6.1. Here we can see that when the budget range is around 3.0 core, the movie will be more successful (block buster) then others in in perception of Bangladesh. Here movie classification variation is also presents (almost all class of movie is here) but success frequency is high in that budget. But budget cannot create any factor in the class of watchable movie. So for getting 6-7 rate, it does not depend on budget.

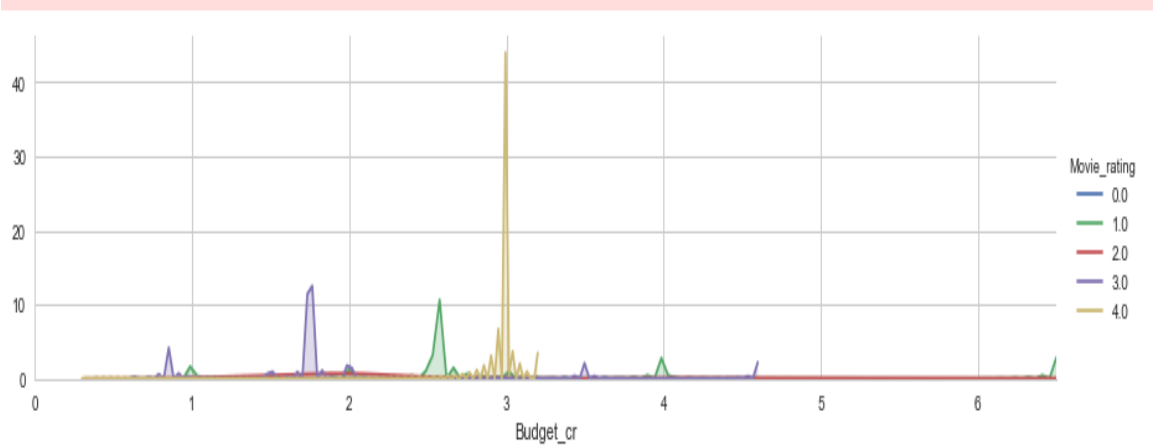


Figure 3.14 Budget frequency with movie rating using seaborn

Most of the movie's budgets are around 1.5 cores to 2.3 cores.

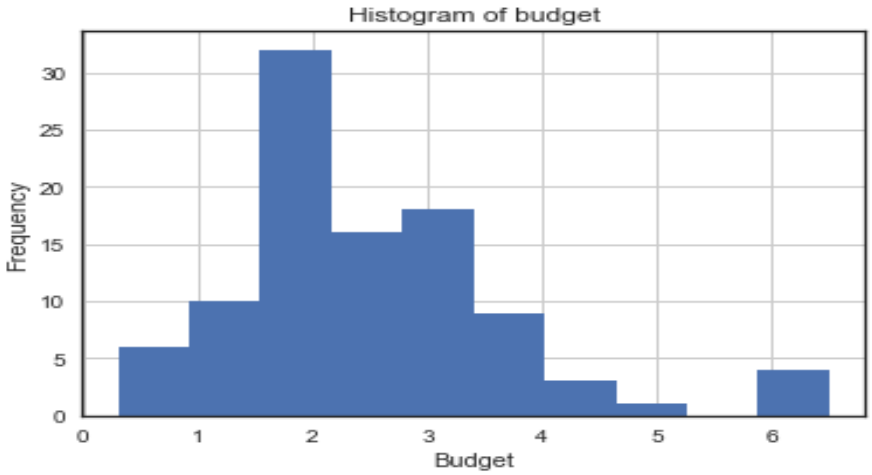


Figure 3.15 Budget frequency with movie rating using histogram.

The visualization of 'Award' feature tells us that the mainly watchable movie will get highest number of award on various category. After watchable movie class, second high frequency is existing for super hit movie class.

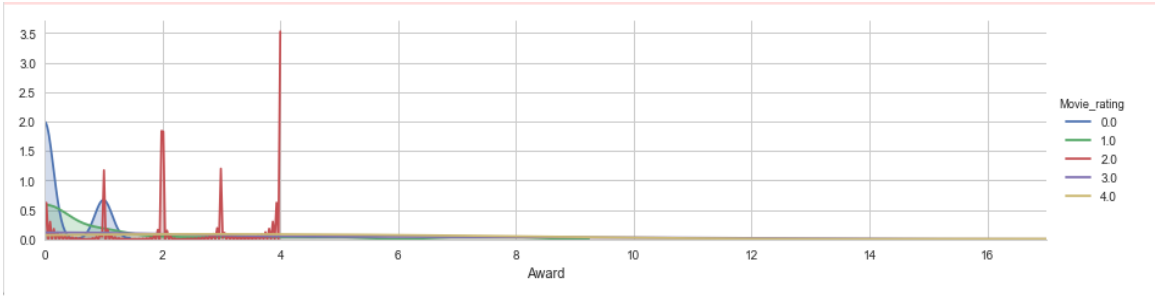


Figure 3.16 show the frequency of 'Award' along with movie rating. using seaborn library

Duration of the movie is also an important factor for movie success. Budget is also a influencer for duration.

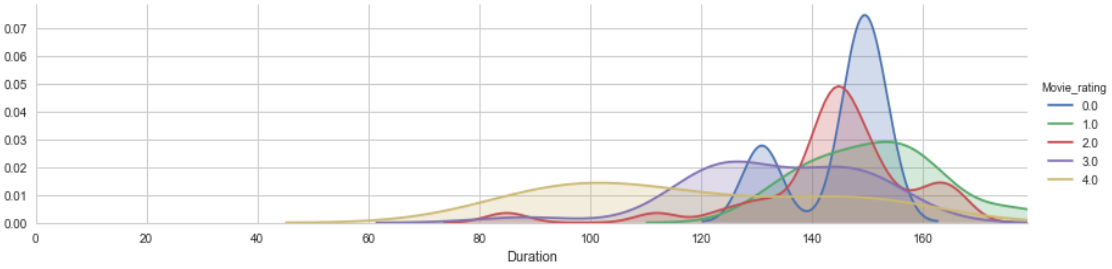


Figure 3.17 The duration frequency with movie rating using sns

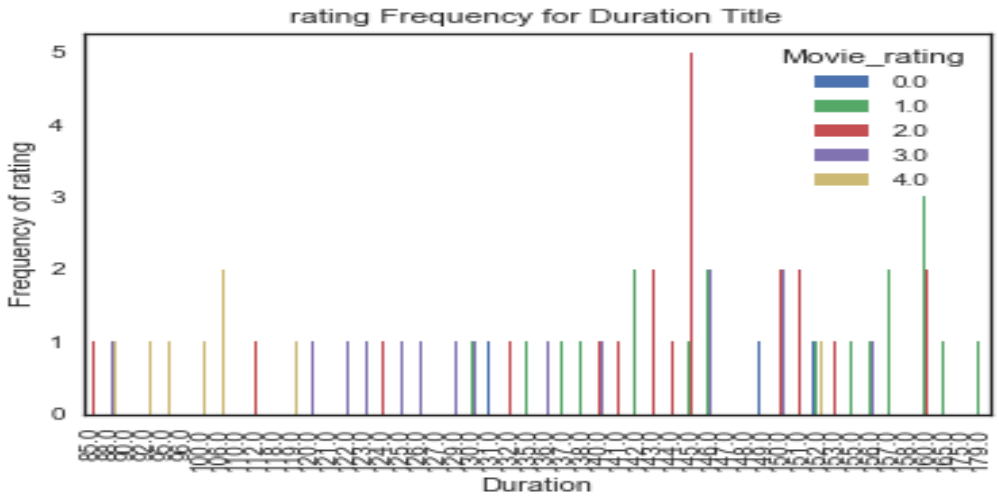


Figure 3.18 Duration Bar chart.

Here we can see that duration is not a factor for increasing movie rate. Lower rating movies duration is high, that's a interesting part.

In the histogram of duration, it is observed that most of the movies duration is in between 144min to 150min approximately.

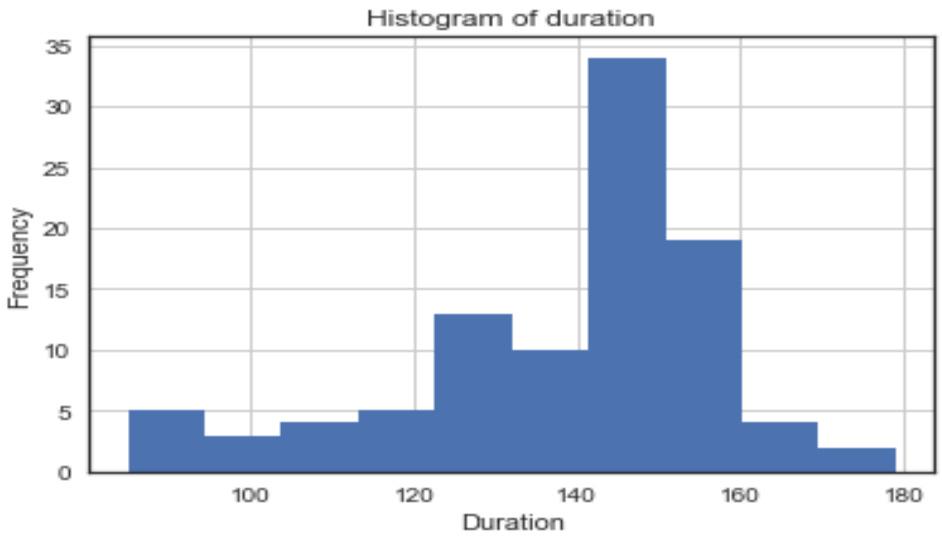


Figure 3.19 Duration frequency with histogram.

The bar plot visualization using matplotlib indicates the variation of producer name is huge. As more than 80% producer produce only one movie so this feature can not help our model to learn any pattern.

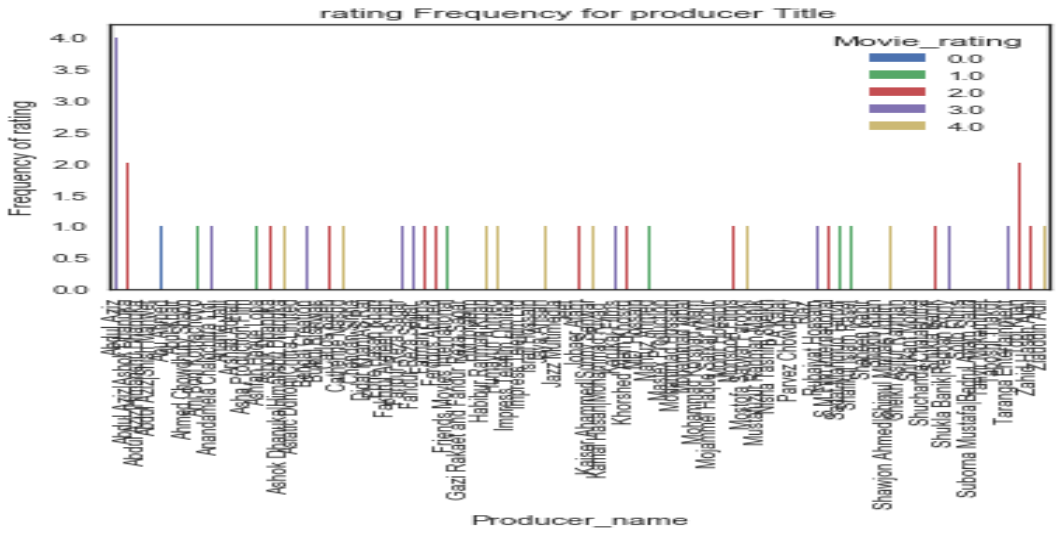


Figure 3.20 Producer name frequency

In this visual part, actor frequency is shown. Here its clear that the variation of actor is much lower then producer feature. So its definitely an important feature and also helps our

model to learn and training season. In the part we can see that, the frequency of actor name is high that means that actor acts in multiple movie.

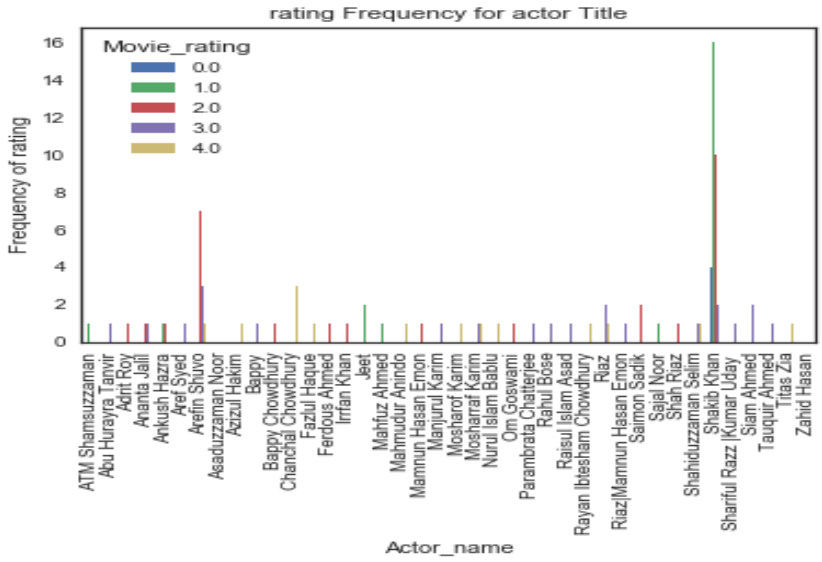


Figure 3.21 Actor_name frequency

Actress frequency is shown in figure 3.22 where the variation length of actress qui name is similar to actor length. Its an important factor for movie success prediction.

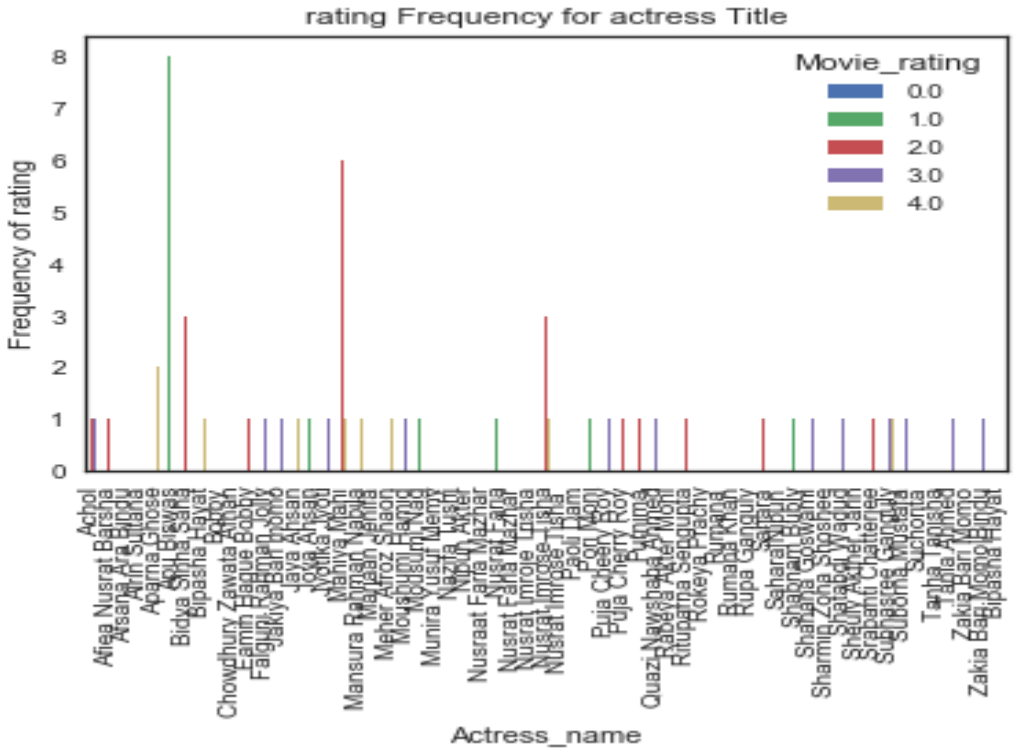


Figure 3.22 Actress _name frequency using bar chart

Director name variation is also huge. Its difficult to find out which movie will be successful depends on director. But in value count sector and mean() method , we see that director success is depends on how many work they did. That means if any director directs several movie then his success of movie rate is increasing.

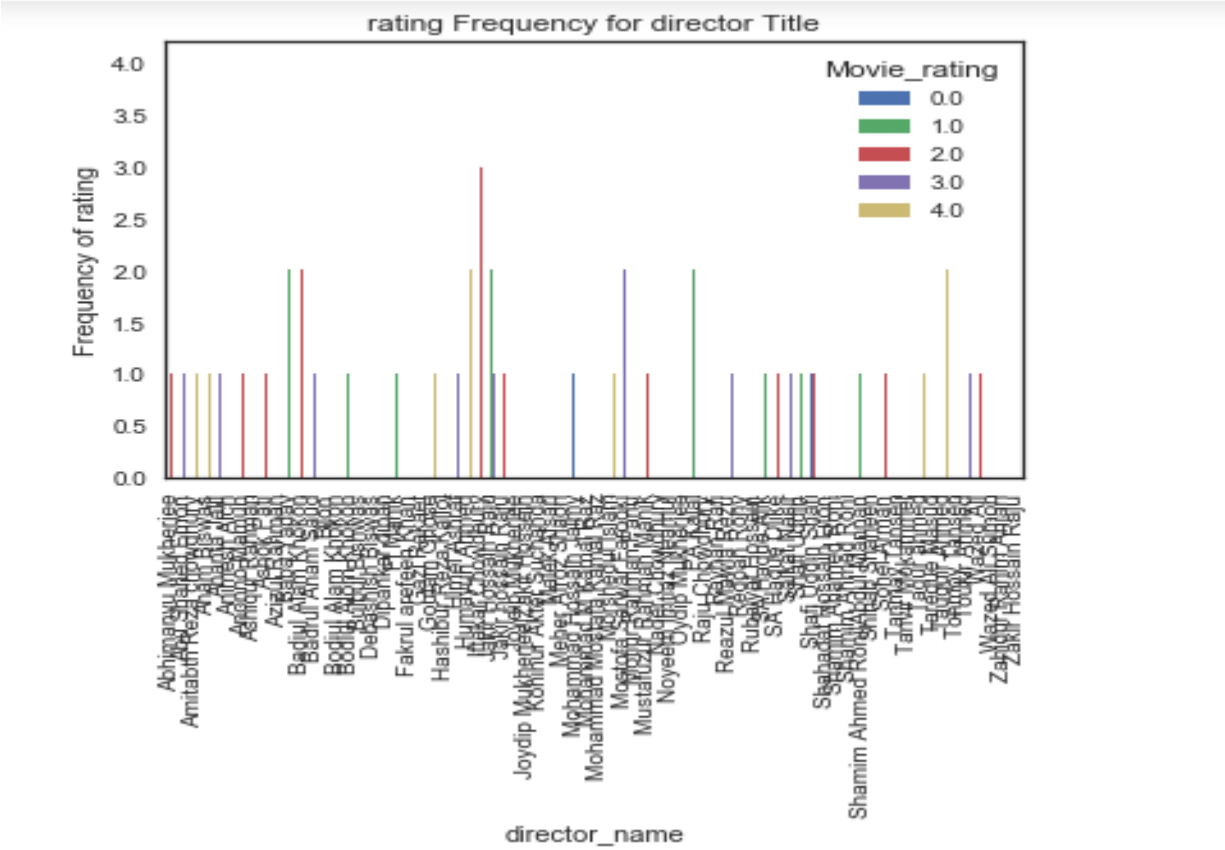


Figure 3.23 Director_name frequency using bar chart

Genre is an important term for movie recommendation system. Its also important for growing people interest in a particular movie. The sequence of movie for highest frequency is like : Drama > Action > Romantic > Crime

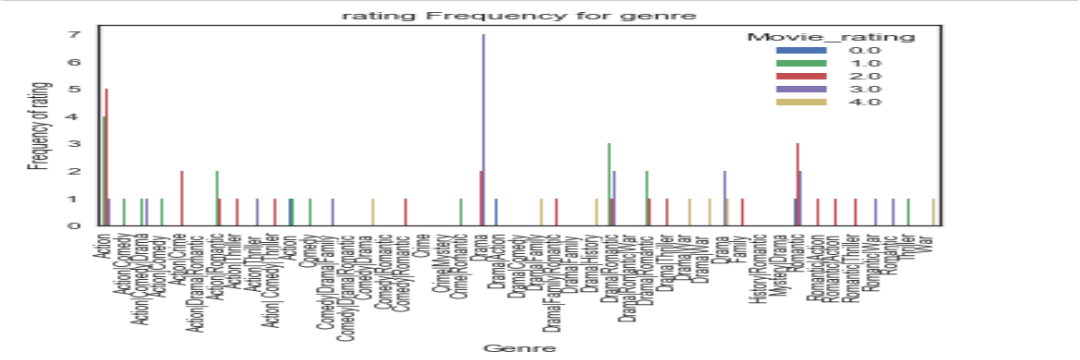


Figure 3.24 Genre of movie frequency

Music director is also a important indicator for movie success. Figure 3.5.6.11 shows the variation and frequency of music director.

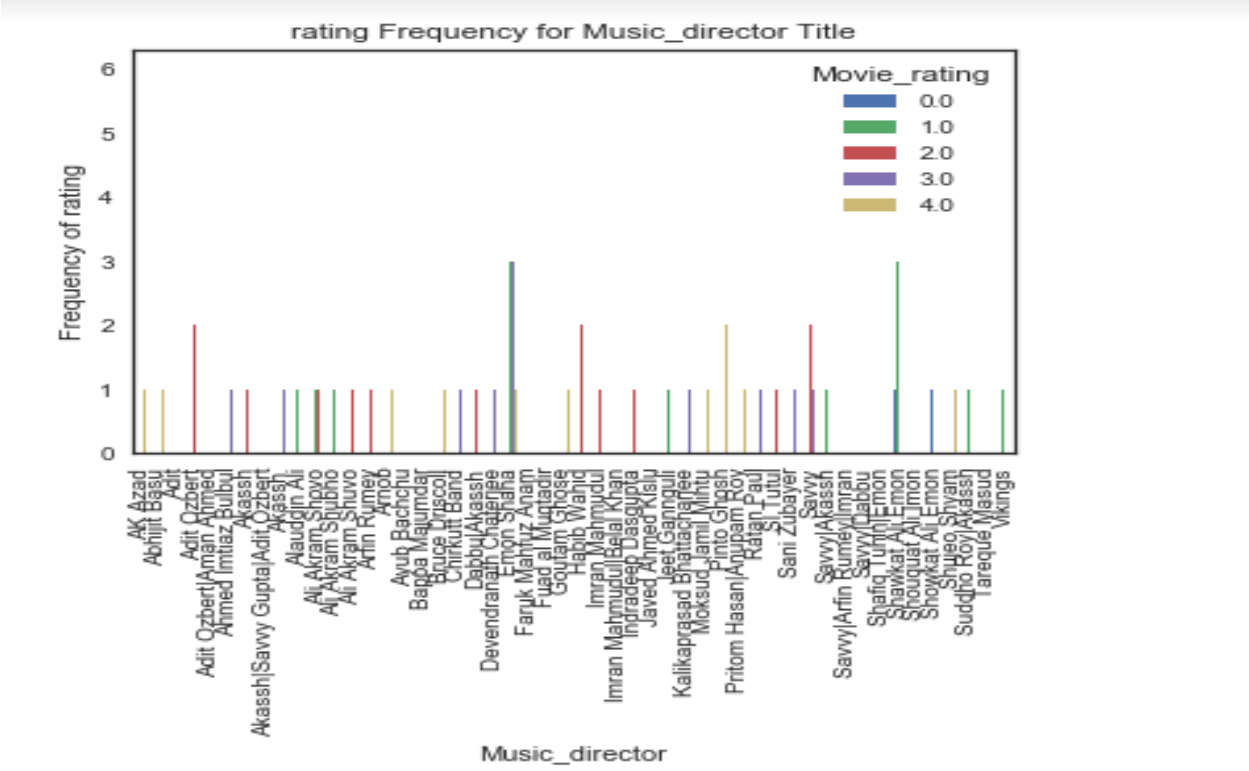


Figure 3.25 Music_director frequency

People are more interest in playback singer of a movie. Favorite singers' movie will be watch by viewer. And it also a good factor or movie success. But due to huge variation of

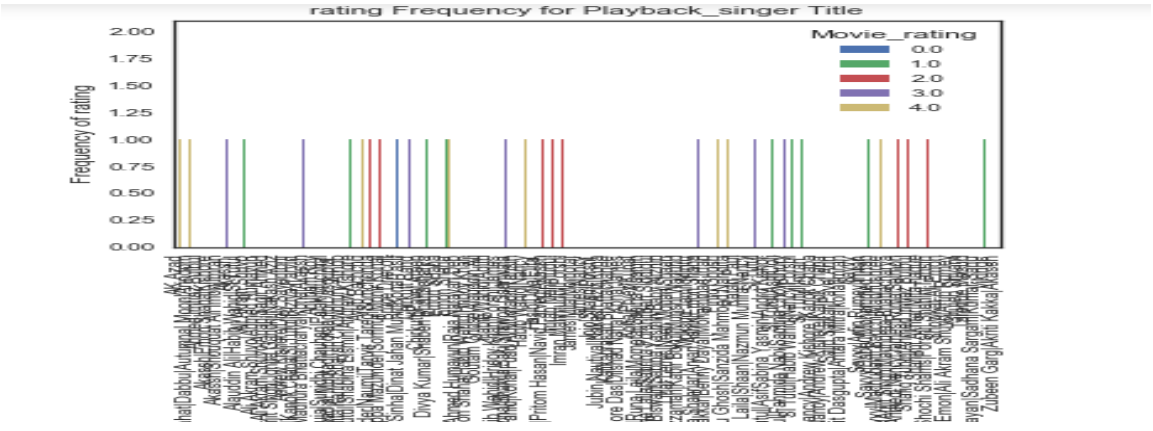


Figure 3.26 Variation of playback singer and its frequency

data, this field does not play a role in our model so much well. Its situation will down our accuracy as well.

Release Day is a important factor. As we see that most of the movie is released in the middle part of a month. Figure 3.5.6.13 is given

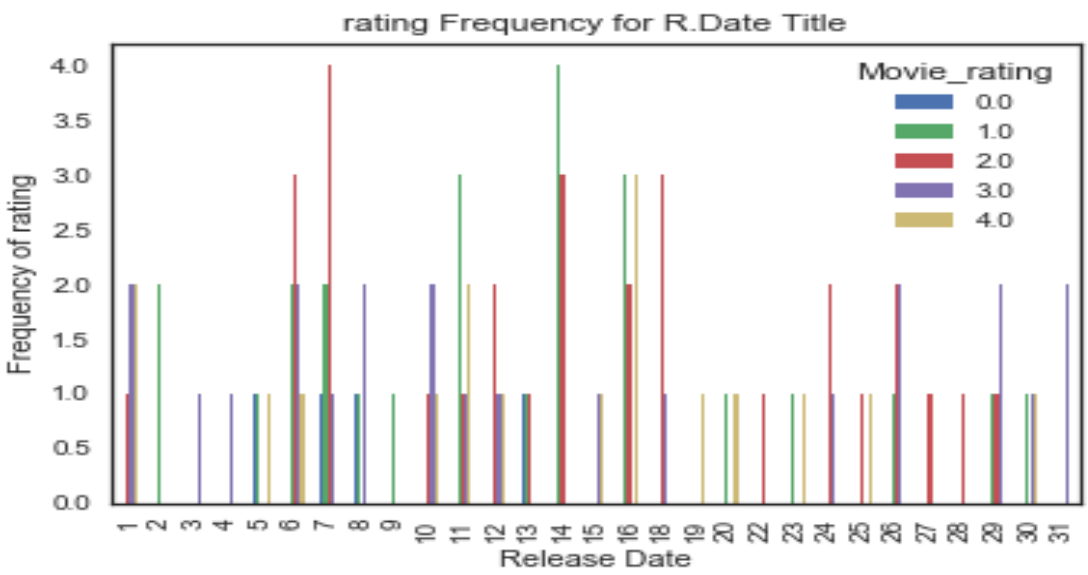


Figure 3.27 Release date frequency in bar chart

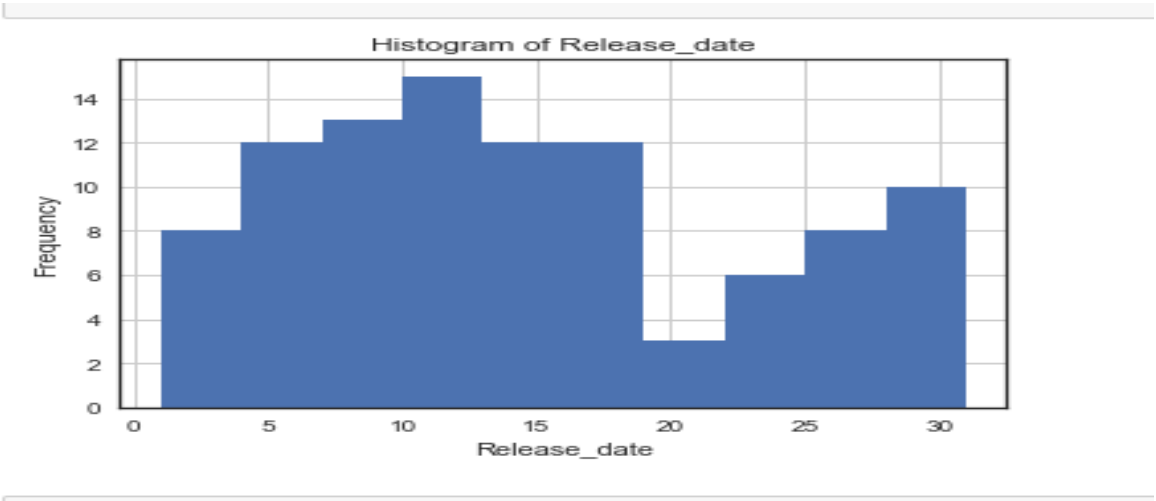


Figure 3.28 Release date frequency in histogram (which indicates most of movie release in 10th to 13th day of a month)

Here another visualization tells us, most of the movies are released at the November and December month of a year.

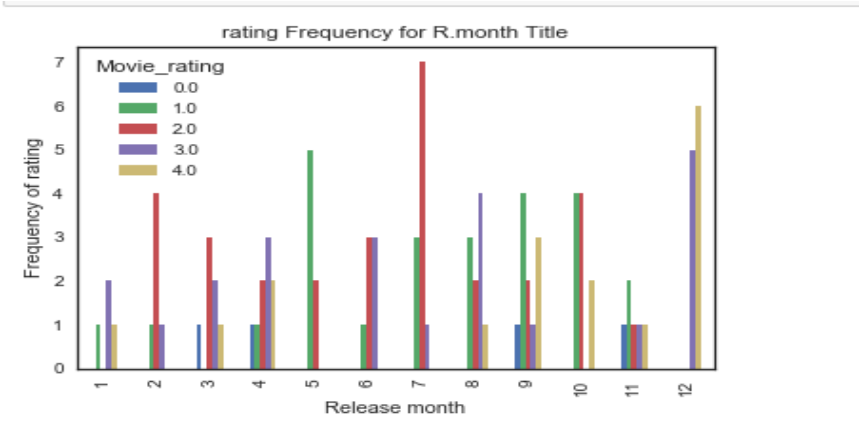


Figure 3.29 Release month bar chart

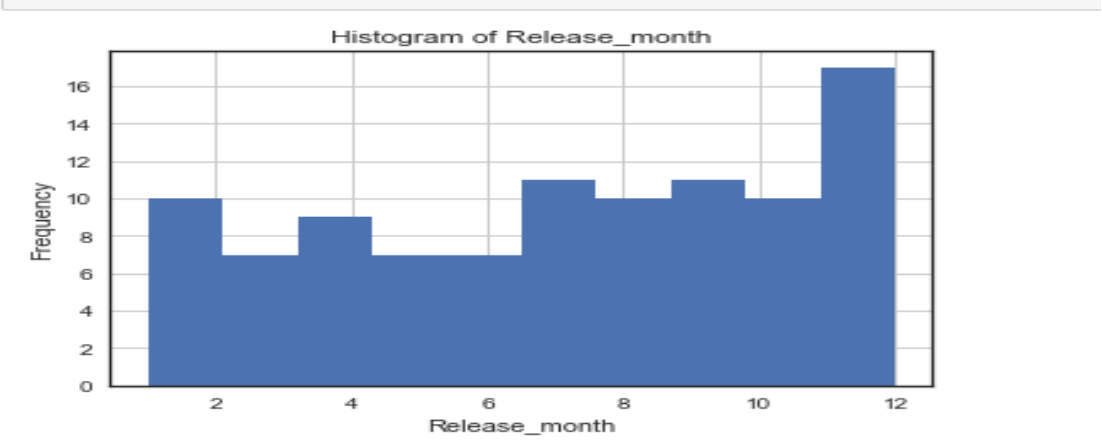


Figure 3.30 Release Month frequency in histogram

3.5.7 Algorithms

In this research work I use five different algorithms for evaluate results and increasing accuracy purpose. They are, Decision tree, Random Forest, Support Vector Machines, Logistic Regression, K-nearest neighbor algorithms.

I split my Bangla movie dataset into train and test data. I split 25% data for testing purpose and 75% data for training purpose. 'Movie rating' feature is used as target label.

Chapter 4

Experimental Results and Discussion

4.1 Introduction

As mention the difficulties of the work, that's why the amount of data of my work is very small that's why my model accuracy is not so high. I am working on it continuously.

4.2 Experimental Results

The performances of used algorithm is shown in the Table 4.2.1. Where we see the highest accuracy is 70% up by using decision tree classification.

Table 4.2.1 Accuracy rate of different algorithms

Serial No	Algorithms Name	Accuracy rate (%)
1	**Decision Tree Classifier**	**70.67**
2	**Random Forest**	**52.56**
3	**K-nearest Neighbor**	**56.0**
4	**Logistic Regression**	**50.15**
5	**Support Vector Machines**	**36.0**

4.3 Descriptive Analysis

To get the better accuracy of model and reduce the complexity the dataset, I use different algorithm as shown above. I also use cross-validation method for improving testing accuracy otherwise model will be over fit because our dataset is short. I also set random

state number in model so that the result will be same. Using python library, I also see different values of Random state and neighbor number of KNN deliver different accuracy. So I use loop to find the best value of those so that accuracy can improve.

4.5 Summary

The comparison between others movie related research work accuracy and my work accuracy is not acceptable because of dataset problem. Dataset is unique, hard to collect and complexity is high. Others works are done with Hollywood, Bollywood datasets which is organized, clear and thousands of samples. So considering this situation my work accuracy is not bad at all and I also work further to improve it.

Chapter 5

Summary, Conclusion, Recommendation and Implication for Future Research

5.1 Summary of the study

Categorical features of any movie dataset is very important factors for model creation and accuracy. But the problem of Bangla movie dataset is due to insufficient amount of data the variation of any feature is huge. Around 80% unique values under a feature. That's why model quality is decrease naturally. Because the first work of any movie worker will not be well in general and they are not continuing their work. Maximum movie related workers did only one movie. It not only creates bad effect of quality but also viewers lose their interest. Another important point is that we have to update more and more movie data on internet so that viewer can see and feel interest. Online advertisement will be the one of the best solution in the situation.

5.2 Conclusion

As a beginner in ML field, I am so glad to work on the most challenging and unique dataset that's can help our movie industry also. I tried my best to estimate a powerful model depends on this real dataset. My model will helps not only movie related people but also viewers to select their interested movie as they can predict movie quality. As it's a new work I faced many problems, not only data collection but also model creation. Some errors solution are not found in internet or answer will not work on my model. Finally, I can create it and will do more work in future.

5.3 Recommendations

To start work in this field, I first attend many seminars on the related field, read more and more paper and articles. I also learn in deals of ML and DM process. Read and learn in

one's own language is most beneficial for all. Though very few work are started and article are written in Bangla, a newbie try to cover them all. Hard working and patience and passion is required to do research in any new sector and on dataset. It a blessing to me that my supervisor Syed Akhter Hossain always helps me as I needed and encourage me all difficult time.

5.4 Implication for Further Study

- Enlarge Dataset
- Increasing number of attribute
- Apply other algorithm for better evaluation
- Apply advanced encoding system to reduce complexity
- Apply more feature engineering techniques for improving model.
- Try to add movie plot so that model can tokenize data and is model can also perform as a movie recommendation system.

APPENDIX

DM- Data mining

VPN – Virtual private Network

KDD – knowledge Discover in Database

ML – Machine Learning

DM - Data Mining

SVM – Support Vector Machine

KNN – K Nearest Neighbors

RF – Random Forest

LR – Logistic Regression

Sklearn –Scikit learn

REFERENCES

[1] Pang-Nine Tan, Vipin Kumar, Michael Steinbach (2017). " Introduction to Data Mining" ISBN 978-81-317-1472-0

[2] "Internet movie database," IMDb.com, Inc., [Online]. Available: http://imdb.com. [Last Accessed 25-03-2019 at 11.00 am].

[3] "Wikipedia",[Online] Available: https://www.wikipedia.org/ [Last Accessed on 27-03-2019 at 07.00 pm]

[4] "Bangladesh Movie box office",[Online] Available https://bdboxoffice.com/ [Last Accessed on 25-03-2019 at 07.00 pm]

[5] "Bangla Movie Database", [Online] Available: https://bmdb.co/ [Last Accessed on 15-03-2019 at 2.00 am]

[6] M. Saraee, S. White & J. Eccleston (2004). "A data mining approach to analysis and prediction of movie ratings", 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9.

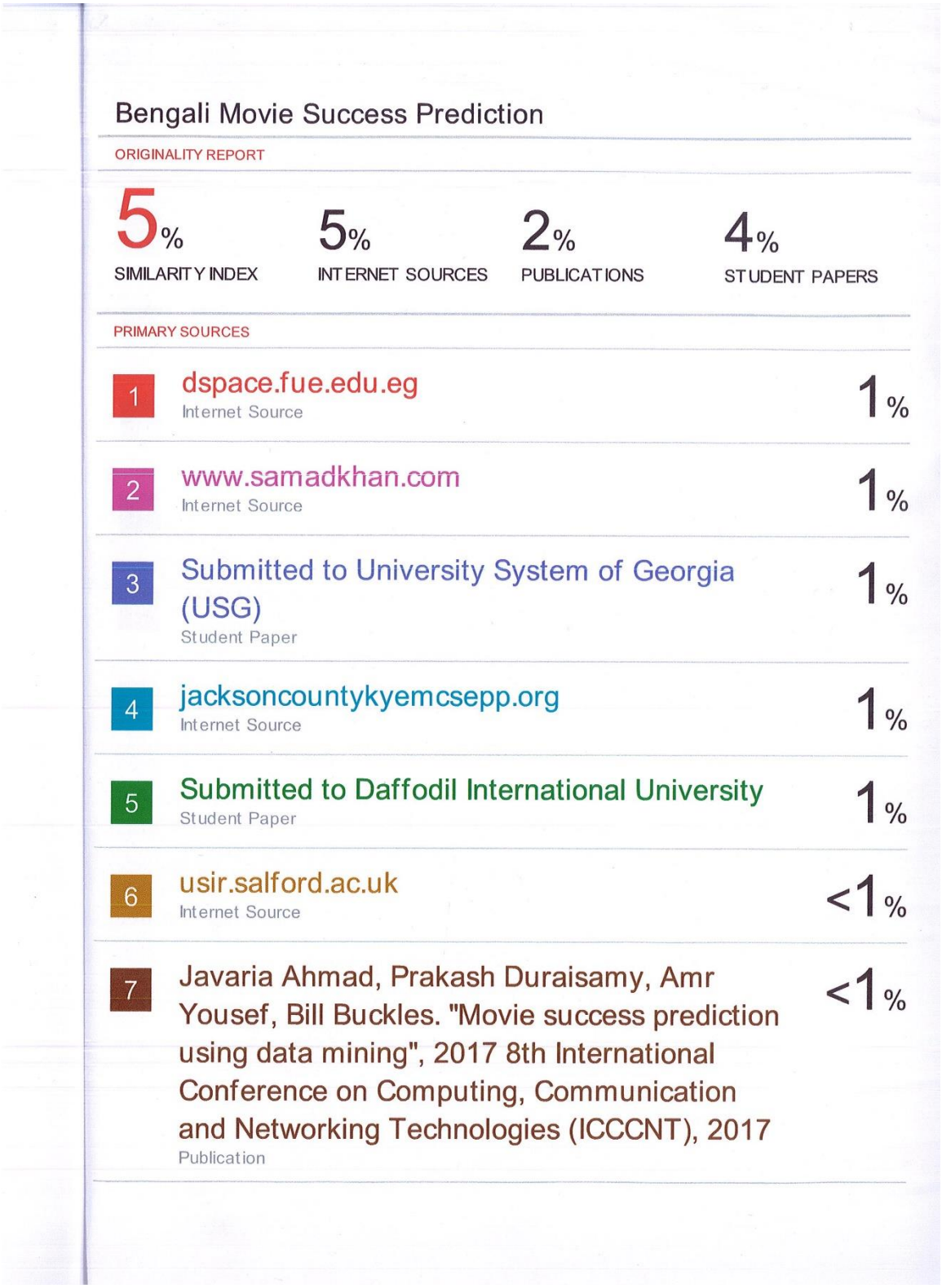
[7] W. Zhang and S. Skiena. "Improving movie gross prediction through news analysis". In Web Intelligence, pages 301-304, 2009.

[8] Karl Persson ECTS, "Predicting movie ratings, A comparative study on random forests and support vector machines",
pp.1-28, 2015

[9] S. Kabinsingha, S. Chindasorn, C. Chantrapornchai, "A Movie Approach and application Based on Data Mining" IJEIT, Volumn 2, Issue 1, July 2012

[10] Hassan and Hammad, "Prediction of Movie Popularity Using Machine Learning Techniques", IJCSNS, Vol.16 No.8, pp.127-131, 2016

[11] "Machine Learning Defination" [Online] Available: https://www.expertsystem.com/machine-learning-definition/ [Last Accessed on 01-04-2019 at 8.00 am]



Bengali Movie Success Prediction

ORIGINALITY REPORT

5%	5%	2%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.fue.edu.eg Internet Source	1%
2	www.samadkhan.com Internet Source	1%
3	Submitted to University System of Georgia (USG) Student Paper	1%
4	jacksoncountykyemcsepp.org Internet Source	1%
5	Submitted to Daffodil International University Student Paper	1%
6	usir.salford.ac.uk Internet Source	<1%
7	Javaria Ahmad, Prakash Duraisamy, Amr Yousef, Bill Buckles. "Movie success prediction using data mining", 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2017 Publication	<1%

8	www.diva-portal.org Internet Source	<1%
9	www.ijariit.com Internet Source	<1%
10	baadalsg.inflibnet.ac.in Internet Source	<1%
11	myassignmenthelp.com Internet Source	<1%
12	unside.t4you.in Internet Source	<1%
13	www.mdpi.com Internet Source	<1%
14	www.ijesmr.com Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		