Prediction Models for Box Office Revenue before Theatrical Release of Movies

MSc Research Project
Data Analytics

Altamash Naik
x16135342

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Altamash Naik
Student ID:	x16135342
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Dr. Catherine Mulwa
Submission Due Date:	11/12/2017
Project Title:	Prediction Models for Box Office Revenue before Theatrical Release of Movies
Word Count:	6625

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	10th December 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:
1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Prediction Models for Box Office Revenue before Theatrical Release of Movies

Altamash Naik
x16135342
MSc Research Project in Data Analytics

10th December 2017

Abstract

This research project proposes to develop prediction models in the entertainment domain. There Prediction Models for Box Office Revenue before Theatrical Release of Movies is a model that has been developed using the parameters of social media platforms and necessary attributes of the movies that influence the success of the movie. The data encapsulating these essential attributes for predictive analysis is extracted from Data repository and using R programming language. The research highlights the impact of predictive data analysis in decision making in lm making industry. Machine learning techniques like Decision Tree, Logistic regression, Nave Bayes, Random Forest and SVM are implemented for designing predictive models that utilises this information and determines the protability of the movie using Rapid Miner. The attained results are evaluated and visualized. Graphical representations of these results are implemented in Tableau and Rapid Miner. The model provides great insights for the investors to identify the key fundamentals that are influential in the success of the movie from initial pre-production stages to the release of the movie covering all the major aspects for marketing and promotions.

1 Introduction

1.1 Project Background and Motivation

Prediction Models for the box oce success of a movie before its release will be benecial for its investors. The amount of influence of social media platforms over the entertainment industry has been growing immensely over the years. After identifying the factors the production team can start making informed decision in accordance to the movie since pre-production stages in regards to marketing the film. The Prediction Models are a collaboration of Key factors implemented using the concept of Machine learning techniques to understand social media buzz surrounding.Machine learning techniques such as logistic regression, SVM, Decision Tree, Random Forest and Nave Bayesian method for data analysis. The section followed is by Project Requirement Specification section that encapsulates all the needs of the Prediction Models that should be developed.

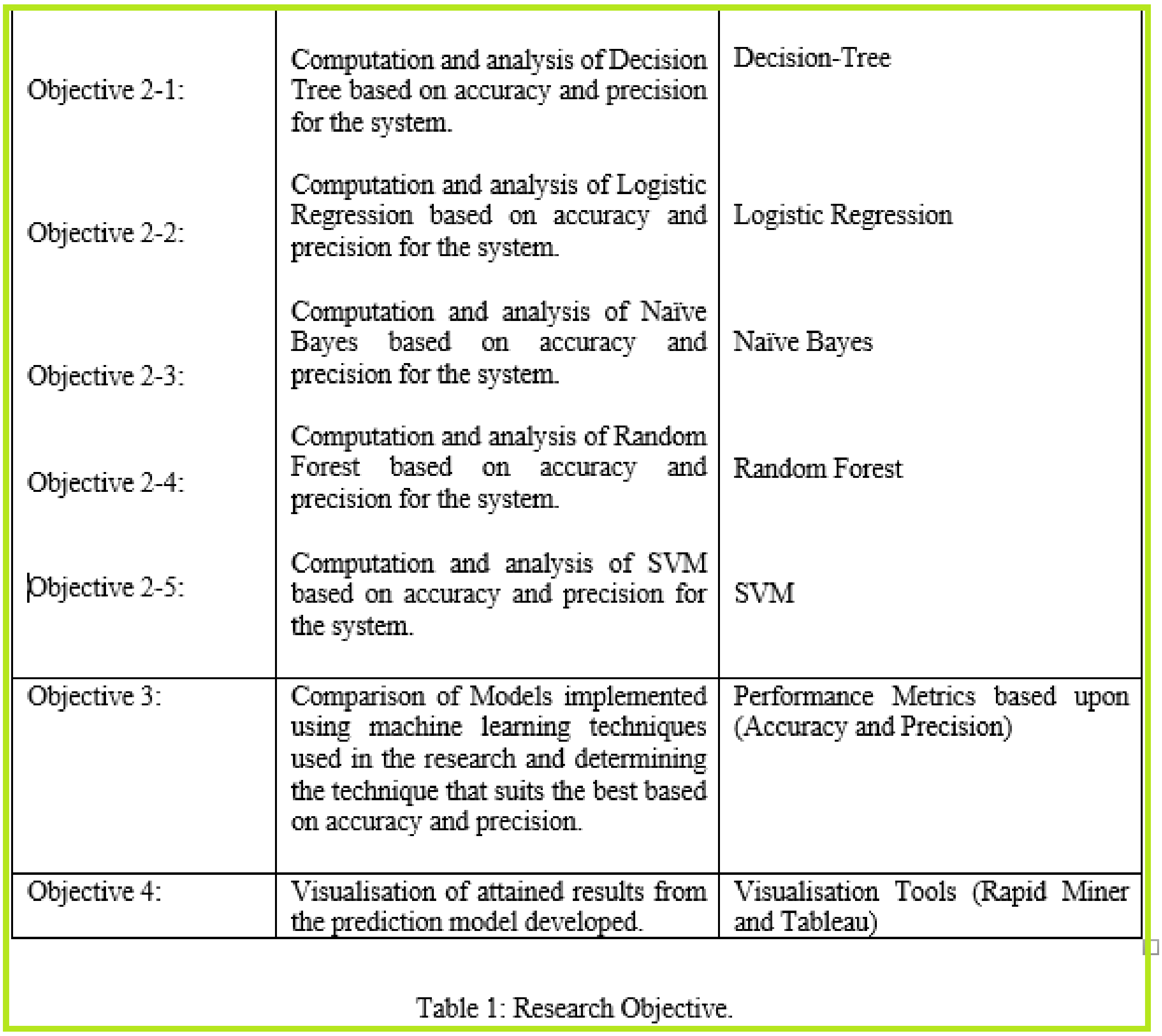
1.2 Project Requirement Specifications

1.2.1 Research Question

RQ: How can prediction of box office revenue help production companies (Walt Disney, 21st century Fox, Paramount Pictures) in supporting and improving decision making process in initial stages of production of movies?

1.2.2 Research Objectives

This table includes all the research objectives that should be completed for successful implementation of the project provided below in Figure 1.



Objective 2-1:	Computation and analysis of Decision Tree based on accuracy and precision for the system.	Decision-Tree
Objective 2-2:	Computation and analysis of Logistic Regression based on accuracy and precision for the system.	Logistic Regression
Objective 2-3:	Computation and analysis of Naïve Bayes based on accuracy and precision for the system.	Naïve Bayes
Objective 2-4:	Computation and analysis of Random Forest based on accuracy and precision for the system.	Random Forest
Objective 2-5:	Computation and analysis of SVM based on accuracy and precision for the system.	SVM
Objective 3:	Comparison of Models implemented using machine learning techniques used in the research and determining the technique that suits the best based on accuracy and precision.	Performance Metrics based upon (Accuracy and Precision)
Objective 4:	Visualisation of attained results from the prediction model developed.	Visualisation Tools (Rapid Miner and Tableau)

Table 1: Research Objective.

Figure 1: Research Objective Table 1

1.2.3 Project Deliverables

This project will result in the following deliverables:
(i) The main contribution is a fully developed and evaluation prediction model for box

office success of the movies. Minor contributions are: (1) Results of the computed and the fully evaluated decision tree model, (2) Results of the computed and the fully evaluated logistic regression model, (3) Results of the computed and the fully evaluated Nave Bayes model, (4) Results of the computed and the fully evaluated Random Forests model, (5) Results of the computed and the fully evaluated SVM model.
(ii) The results of the comparison of the developed models (Objectives 2-1 to 2-5) (Kurt et al.; 2008)
(iii) The visualized results of the prediction model. The rest of the technical report as follows: Chapter 2: Presents a review of predictive models based on Entertainment Industry, Chapter 3: Methodology used, and modification made in the methodology, Chapter 4: Presents the design, processing, implementation of Prediction Models for Box Oce Revenue before Theatrical Release of Movies, Chapter 5: Evaluation and Visualization of Predictive Model developed, Chapter 6: Provides Conclusion of the and Future work.

2 Literature Review of Prediction Models based on Movie Industry

2.1 Introduction

Prediction models can be very beneficial in an industry that invests millions of dollars every- year in its product. Entertainment industry is one such domain. The amount of investment in the industry has only been on the rise over the year. It is a common term in the industry that all the trade insiders always assert on Box Office is always unpredictable. The fact cannot be denied however a pattern can be analyzed on the successful movies based on which a prediction models can help the investors identifying the key factors that play an influential role in the success of the movie. After evaluating the factors and working over them constructive planning can be done to address those factors to improve the chances of making the movie a profitable venture.

2.2 Investigating different predictive models for Movies

There can be multiple factors that play a significant role in the success of the movie. The impact of each factor can vary in different cases for the movies these factors include star cast, Director, production house even the social media popularity of these cast can be an influential factor in many cases. A predictive model developed by (Lash and Zhao; 2016) is based upon on a similar concept it examines the movie and bases its analysis on the factors like Star power, Release date and hybrid Features using Regression analysis focusing dependent upon 4 major machine learning techniques Decision Trees, Logistic Regression, Random Forest and Nave Bayesian. There is another prediction model which performs analysis of the success of a movie including these features as well as ratings of the movies using ROC curve and AUC accuracy by (Parimi and Caragea; 2013) the major drawback of this model being that it does not consider the concept of Return on Investment (ROI) which is a key component in the Revenue of the movie. There have also been multiple systems that have been designed for the customer that help customer in selecting the movies that an individual should prefer dependent upon his preferences and interest which can be termed as movie recommendation system (Wang et al.; 2014)

2.3 Identifying Gaps in Predictive Models

The multiple prediction models developed on Movies a brief description of these models and analysis of gaps in these models will consolidated the development of this predictive model.

2.3.1 An improved collaborative movie recommendation system using computational intelligence

An improved collaborative movie recommendation system using computational intelligence This model has been specifically designed focusing on the needs of the customer (movie goers). It is based on the concept of online movie recommendation system which utilizes historical data generated by the individual and precisely capturing close neighbours for the individual. However, with increasing number of movie releases and the individual views the distinguishing lines starts to blur thus impacting the accuracy of prediction of the system. This system counters this problem by using hybrid model which is a combination of improved K-means clustering method and Genetic Algorithm. It also employs PCA (Principal Component Analysis) and data reduction techniques which improves the accuracy of the system.

Identified Gap in this System: The major flaw of this system is it is customer oriented which is not helpful from the investors perspective. However, this model can act as a base for analysis to identify the trends in the market and the genre that can be successful.

2.3.2 Mining Online reviews for movie performance

This is model is more concise towards the perspective of the investors. This model functions on sentimental analysis using concepts of two major machine learning like logistics regression and SVR. The approach of this system is based upon S-PLSA (A Probabilistic Approach to Sentiment Mining). It uses the reviews of the movie released and then help in predicting the collection based on these reviews. (Dave et al.; 2003)

Identified Gap in this System: This system predicts the box office revenue of the movie after the movie is released benefiting the customers as well as the distributors of the movie more in contrast to the investors as it mines the reviews that are generated after the theatrical release of the movie. (Yu et al.; 2012) which also turns ineffective for the investors. Thus it cannot be used in the initial pre production stages of the movie.

2.3.3 Predicting Box office success of movies using Neural Networks

A predictive model based on neural networks is a model more on the similar lines. The attributes used for this model are MPA rating, Genre, Sequel, Competition, Screen Count and Star value. It is a system that does nit consider the social media platform data for prediction models.It does not apply hybrid functions or feature engineering only utilizing its core dataset. (Sharda and Delen; 2006)

Identified Gap in this System: Even though the accuracy of this system is commendable a question can be raised over the attributes used for the system as it does not cover up social media which has become one of the most significant factors in determining the success of the movie. Also, another flaw of the system is that it only uses one technique can does not compute and compare other machine learning techniques restricting its reach.

2.3.4 The Who, What, and When of Protability

Early predictions of movie success (Lash and Zhao; 2016) is among the most developed predictive models for the investors in movie industry. It is based upon the concept of feature engineering and then implements multiple machine learning techniques over the classified database. This machine learning techniques by this system are Logistic Regression, Random Forest and Nave Bayesian. It also includes a major factor for social media analysis i.e. Facebook and the necessary attributes for predictions. The system is dependent upon the feature engineering of ROI derived from gross and budget attributes.

Identified Gap in this System: This prediction model does not includes information about Twitter and YouTube parameters for social media analysis. The number of computations of every machine learning technique is restricted as well.

2.4 Conclusion (Filling the identified gaps)

As it can be observed based upon the identified gaps there are no prediction models in the system that keeps an account of Twitter Count or YouTube which have become one of the salient Features of Social media platform. Apart from that even the models implementing machine learning technique do not provide extensive computation and comparative analysis in the techniques used. Based upon these identified gaps Chapter 3 proposes a methodology that should be implemented in this case followed by Chapter 4 Design, Processing and Implementation of Prediction Models for Box Oce Revenue before Theatrical Release of Movies

3 Research Methodology Approach Used

3.1 Introduction

The project is based on Knowledge Discovery and Data Mining i.e. Kdd process which is based upon 5 distinct phases. However, the methodology has been modified according to the needs of the project.The modified Methodology is below Figure 2. As the majors phases of this methodology effectively collaborates along with the phases of the project.

3.2 Modified Methodology Approach used.

Final Database: The final database is a culmination of 3 distinct datasets that have been extracted from multiple sources Our main dataset has been extracted from Data.world which includes major attributes of the prediction model. YouTube Extracted Data and twitter extracted data are two other social media buzz influential factors that play a determining role in social media analysis.
Cleaning Pre-processing of Data: These datasets have been joined together using R programming and cleaning of the data has been performed using Rapid Miner. The pre-

processed data has also been used to develop certain attributes for laying the foundation of Feature Engineering.

Feature Engineering for Machine Learning Techniques: Based on the available attributes we implement various machine learning techniques new attributes are generated.

Application of Machine Learning Techniques: The processed data is then passed through various machine learning techniques that are computed individually to determine the best possible attributes that should be accounted for each technique implemented.

Predictive Model Visualisation of patterns trends of Predictive Model: Based on this model module patterns and trends are evaluated. Visualizations are developed using 2 different technologies Tableau and Rapid Miner. (Williams and Huang; 1996)

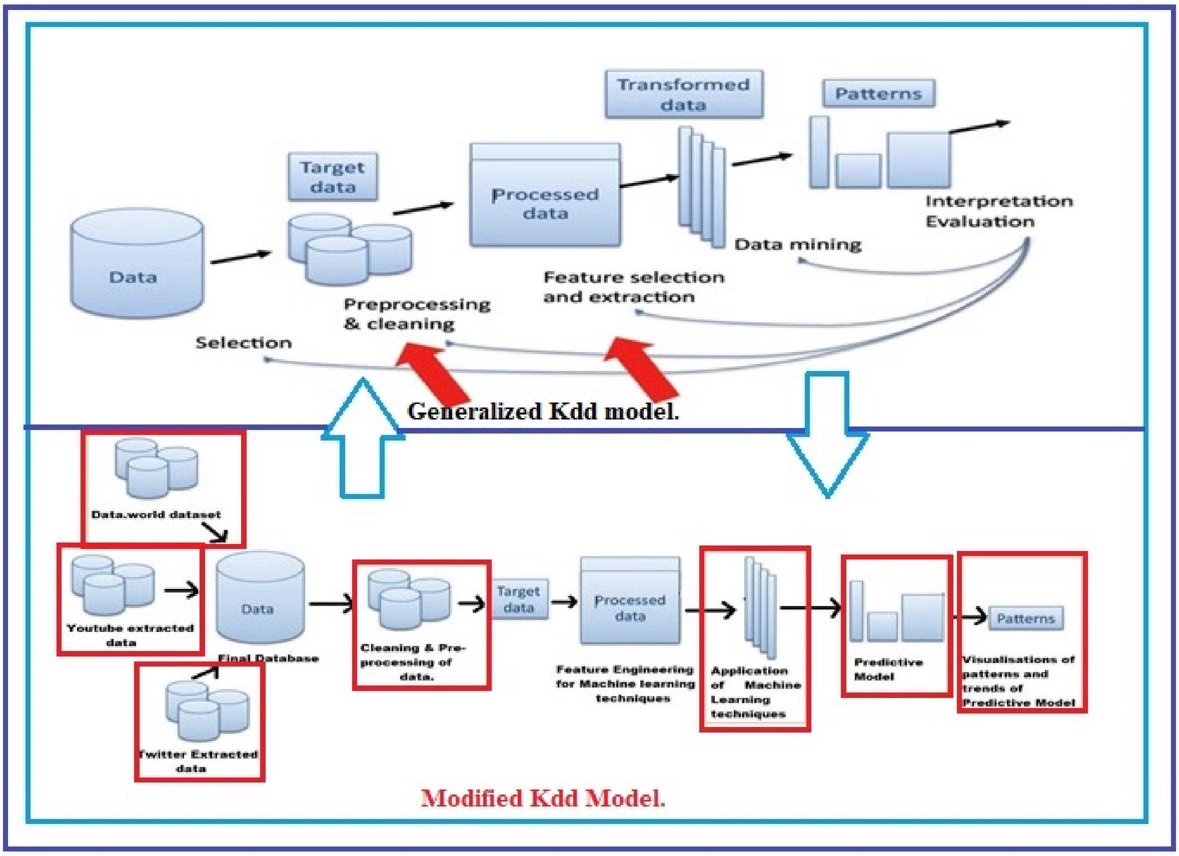


Figure 2: Modified Kdd Methodology

3.3 Data Preparation Process Flow Diagram

The most primitive and fundamental part of the project are datasets there are multiple datasets that have been used in the projects that can be broadly categorised in two different sections such as Structured Dataset: A well organised with predefined attributes which can be extracted without the utilization of any functions can be referred as

Structured Dataset: In this project data.world serves repository for structured data. https://data.world/popculture/imdb-5000-movie-dataset (Kayacik et al.; 2005).

Unstructured Dataset: Datasets that are extracted based upon certain pre-defined requirenebts of the queries dependent upon the functions in the data using a specific extraction method. For this prediction Model R programming is used for extracting 2 unstructured datasets which are Twitter and YouTube.

(i) Twitter Unstructured Dataset: This data is extracted using two different libraries and 1 function.

twitteR: An R package that helps in providing access to the Twitter API. This functionality majorly focuses on the APIs which are more useful in data analysis.

read.csv: This library helps us in reading a csv file attained from twitter.

rbind: This function is used in this case to bind the number of followers attained for the actors in accordance to our structured dataset

(ii) YouTube Unstructured Dataset: This dataset is extracted by installing two significant packages devtools and httr and two major libraries curl and jsonlite a brief description about them are as follows.

devtools: This package provides us tools that helps us in developing R packages easier.

httr: This package provides us tools that helps us in working with http and urls in R.

jsonlite: This library acts as a robust parser generator for R.

curl: This helps in downloading the file from url. (iii) Combing datasets: The main dataset is a combination of these structured and unstructured datasets. This dataset is formed using R programming. It is performed using two main libraries tidyr and dplyr

tidyr: It is used for tidying of data not for generally reshaping or aggregating it.

dplyr: It helps in using the data frame as an object.

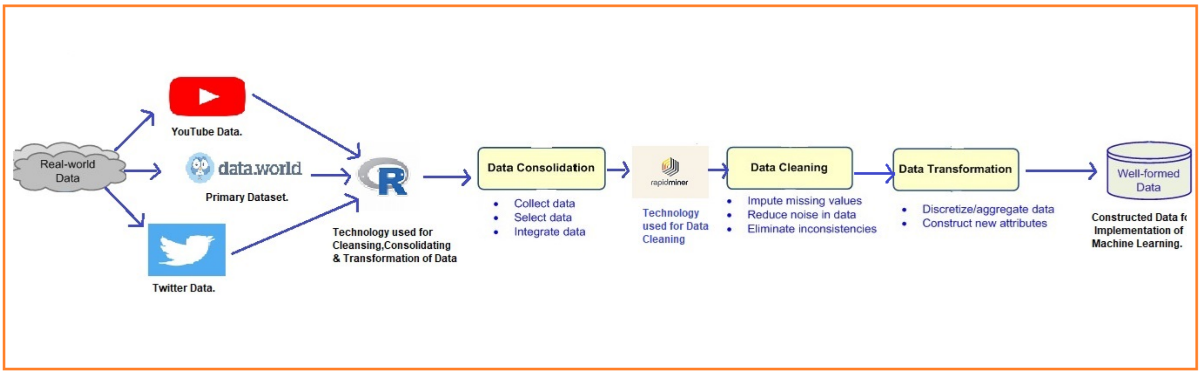


Figure 3: Data Preparation Process Flow Diagram

This Figure 3 encapsulates a systematic process description of all the steps undertaken for data preparation for the Predictive Model.

3.4 Feature Selection Process

ROI attribute: This attribute is generated on the basis of gross and budget of the movie. It is generated using an automated formula of ROI=(Gross-Budget)/Gross. in excel itself the values attained are integer format based on which classes are formulated. Label classification: The labels are classified under these 3 sections if ROI is less 0 it is considered as a Flop, if ROI is between 0 and 0.2 it is considered as Average and if ROI is greater then 0.2 it is considered as a successful using the if else looping structure in (Scott and Matwin; 1999).

3.5 Architectural and Technical Design of the Prediction Models for Box Office Movie Design

The architectural design of the project has been based major machine learning technologies for Predictive Analysis and technologies utilized on this. Different technologies have been used in distinct phases of the project. For data preparation R programming is used to collaborate multiple datasets and transforming the data in accordance to the machine learning technique. While cleaning of data is performed using Rapid Miners function of Replace Missing Values in which the missing values in all the attributes are replaced using average values. After that multiple machine learning techniques are applied over the dataset to attain results for the prediction models. Visualization is the next phase of the architecture in which visualization of the attained results and the datasets are performed using Rapid Miner and Tableau. The architectural design has been provided in Figure 4.

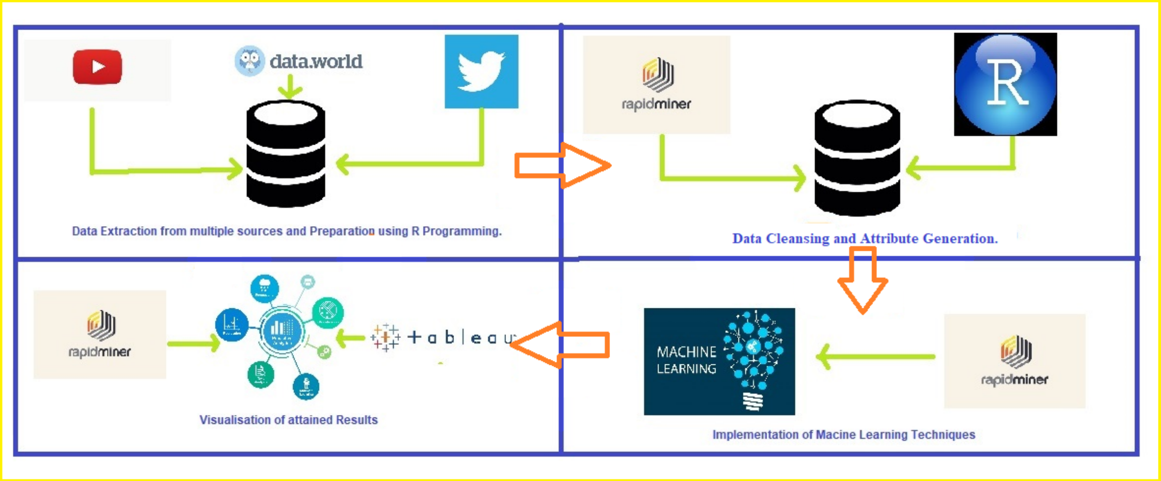

data.world
Data Extraction from multiple sources and Preparation using R Programming.
rapidminer
R
Data Cleansing and Attribute Generation.
rapidminer
+ a b l e a u
Visualisation of attained Results
MACHINE LEARNING
rapidminer
Implementation of Macine Learning Techniques

Figure 4: Architectural Design of the Predictive System

There have been multiple machine learning techniques that have been implemented in this case all of them have been performed to identify the strength and weakness of each techniques for which each technique has been implemented with its possible combination on attributes thus understanding the impact of attributes on each of the technique implemented and overall the techniques are compare together to develop the best possible result, technique and attributes to be considered. Thus, adhering to the attributes mentioned in section Research Objectives Refer Table 1 of Research Objectives (Section 1.2.2 (Sabhnani and Serpen; 2003).

Technology Used: The entire implementation of the project is based upon 3 major technologies Rapid Miner, R programming and Tableau

3.6 Conclusion

As compared to CRISP DM, SEMMA and OLAP this methodology provides Feature Selection which is a significant part of the project. There have been a few extra phases in our methodology process according to the development of the system. These phases have been included in our Modified Kdd methodology. (Azevedo and Santos; 2008).

4 Implementation, Evaluation, Analysis and Results of Prediction Models for Box Oce Revenue before Theatrical Release of Movies

4.1 Introduction

Developing of environment for implementation of the machine Learning techniques have been performed using Rapid Miner. In this case the finalized data set is attained after attribute generation from R programming. Cleansing of the data has been performed in Rapid Miner using a Replace Missing Values functions. In this dataset multiple factors have been encapsulated that play a pivotal role in the success of the movie these factors have been extracted from multiple attributes of the distinct dataset that have been collaborated in the main dataset extracted from distinct resources. These attributes help us in predicting the success of the moviebased upon these factors after computing the machine learning techniques individually and comparing them overall. Depending upon the Machine learning technique the combinations of the attributes differ. However, there are certain default attribute set up that we consider initially based upon which we make further changes. Before delving into the details of each methods let us cover up some General Functions used in Rapid Miner, Default Attributes and generalized setup of the machine learning techniques.

4.2 Analysis and Evaluation of Prediction Models on Performance Metrics

The implementation of the Prediction Model is done in Rapid Miner. The evaluation is based upon the accuracy performance metrics. Visualizations of the attained results is performed in Rapid Miner and Tableau through different charts.

4.3 Implementation and Evaluation of Box Office Prediction Model

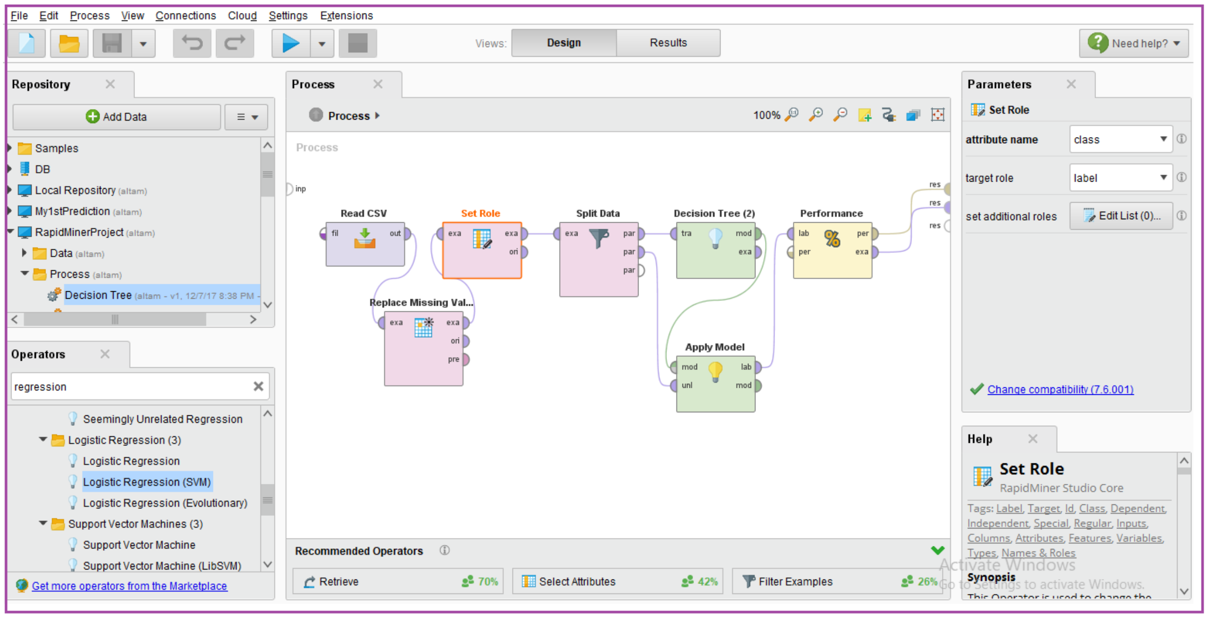
The model is developed using certain General Functions in Rapid Miner a brief description of which is provided below and a set of Default attributes (Defined in Configuration manual).

Read CSV: This Operator helps in reading an ExampleSet from the specified CSV file in Rapid Miner.
Replace Missing Values: This Operator calculates the missing values in the dataset and replaces the missing values of the dataset by a specified replacement in this predictive model which has multiple options available in these prediction model it replaces the missing onea with average values of the attribute in which the missing value exist.
Set Role: This Operator changes the role of one or more attributes of the ExampleSet. The default role is considered as regular, other roles are classified as special. In the predictive model class is defined under the label role (label is the special attribute on the basis of prediction is performed).

Split Data: This Operator splits the entire dataset into small subsets based upon the ratio (total should always be 1). This ration accordingly splits the data ad provides the susets to applied machine learning technique as well apply model.
Apply Model: This Operator plays the most significant part in the implementation. This model is first trained on an ExampleSet by the machine learning Operator. To get the prediction on unseen data.
Performance: This operator is used for statistical performance evaluation of classification tasks. In this prediction model we evaluate the performance based upon accuracy

4.3.1 Functions required for Machine Learning in Rapid Miner and their setup

This screen-shot helps us in understanding the connectivity of multiple functions utilized in Rapid Miner and the sequential flow. This screen-shot has been for Decision Tree Algorithm. Similar set up and functions are used for other machine learning techniques as shown in figure 5


Set Role
attribute name: class
target role: label
set additional roles: Edit List (0)...
Change compatibility (7.6.001)

File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Need help?
Repository
Add Data
Samples
DB
Local Repository (altam)
My1stPrediction (altam)
RapidMinerProject (altam)
Data (altam)
Process (altam)
Decision Tree (altam - v1, 12/7/17 8:38 PM)
Operators
regression
Seemingly Unrelated Regression
Logistic Regression (3)
Logistic Regression
Logistic Regression (SVM)
Logistic Regression (Evolutionary)
Support Vector Machines (3)
Support Vector Machine
Support Vector Machine (LibSVM)
Get more operators from the Marketplace
Process
100%
Read CSV
Set Role
Split Data
Decision Tree (2)
Performance
Replace Missing Val...
Apply Model
Recommended Operators
Retrieve 70%
Select Attributes 42%
Filter Examples 26%
Parameters
Help
Set Role
RapidMiner Studio Core
Tags: Label, Target, Id, Class, Dependent, Independent, Special, Regular, Inputs, Columns, Attributes, Features, Variables, Types, Names & Roles
Synopsis
Activate Windows
Go to Settings to activate Windows.

Figure 5: Accuracy of Decision Tree Best Model.

4.3.2 Implementation, Computation, Analysis and Evaluation of Decision Tree based on accuracy and precision for the system

Decision Tree Concept: It is supervised machine learning algorithm which uses the concept roots and nodes fro its implementation. It splits the data as t moves forward in its analysis and te nodes keep on expanding accordingly (Smith and Tansley; 2003).
Advantages of Decision Tree: Easy in explaining and setting of rules. Interpretation of a complex Decision Tree model can be easily simplified using visualizations.
Disadvantages of Decision Tree: It the number of label class variables are greater than the calculation and evaluation become very complicated in decision tree.

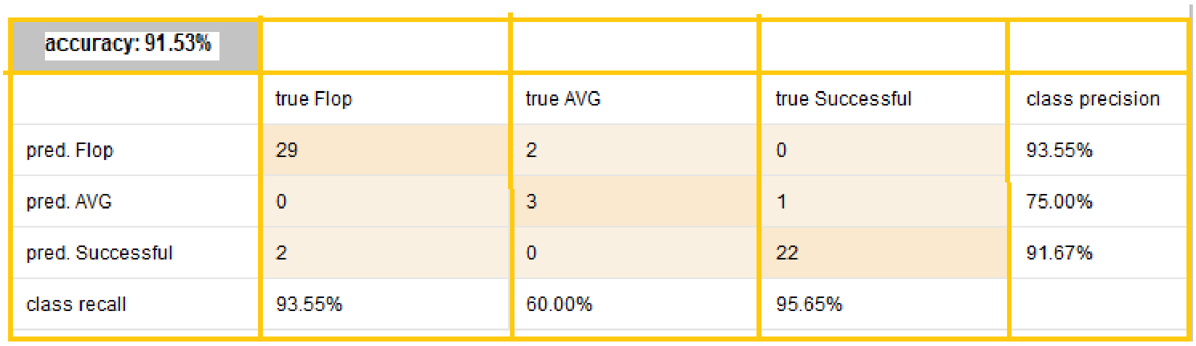
Operator Used for Decision Tree in Rapid Miner.: This Operator helps in implementing decision tree in Rapid Miner which helps us in setting up the criterion as gain ratio a maximum depth of 20 and application of pruning and pre-pruning over the model.

Computation of Multiple Decision Tree Models implemented.

Different Decision Tree Models are performed based upon the distinct combinations of attributes and how the changes affect accuracy of the overall prediction. There have been 4 distinct iterations performed in this machine learning technique the main iteration is based upon the Default attributes and changes and the accuracy's have been mentioned below.

Table 1: Decision Tree Model Iterations Accuracy

Model	Attributes Considered	Attributes Type	Accuracy
Decision Tree	Default	General as earlier	91.53%
Decision Tree 1.1	Without Facebook Attributes	General as earlier	86.61%
Decision Tree 1.2	Without Budget Attribute	General as earlier	78.31%
Decision Tree 1.3	Without ROI Attribute	General as earlier	67.80%



accuracy: 91.53%				
	true Flop	true AVG	true Successful	class precision
pred. Flop	29	2	0	93.55%
pred. AVG	0	3	1	75.00%
pred. Successful	2	0	22	91.67%
class recall	93.55%	60.00%	95.65%	

Figure 6: Accuracy of Decision Tree Model.

Evaluating of Decision Tree Models: The highest accuracy attained in this case is 91.53% which is a significant increase than the previous model. Early predictions of movie success (Lash and Zhao; 2016) in which the decision tree accuracy is 73%.The accuracy is mentioned in the Figure 6 above.

Visualisation of Decision Tree model: This visualisation of Decision Tree is projected for the number Twitter followers across the globe filtered specifically for top 10 countries across the globe. Encapsulating the impact of the social media platform has in the country which can be effectively utilized for marketing the movie in the country and creating a positive buzz for the movie that will overall have an impact on the success of the movie. It is Geographical Map of the Globe covering up Top 10 Countries with the maximum no of twitter followers as shown in Figure 7. It also provides the number if followers every country over along with the name of the country.

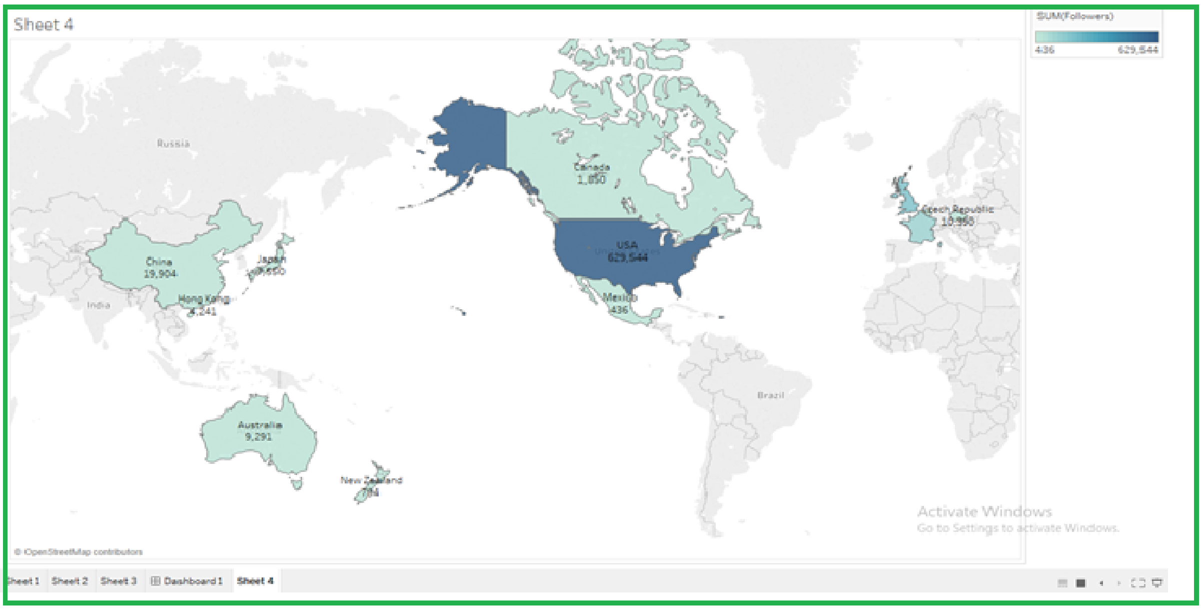


Figure 7: Visualizations for Decision Tree Models.

4.3.3 Implementation, Computation, Analysis and Evaluation of Logistic Regression based on accuracy and precision for the system

Logistic Regression Concept: It is a supervised classification machine learning algorithm. Logistic Regression is a linear classifier which is dependent upon the logit boost of the attributes of the dataset toward the primary target class. There are two types variables dependent and independent where the dependent variables are derived from independent variables. (Hsieh et al.; 1998).

Advantages of Logistic Regression: It is not necessary for the data variable to be normally distributed which makes it robust. A direct linear relationship is not considered in the dependent and the independent variables.

Disadvantages of Logistic Regression: Information is biased in this model as it only evaluates binomial classes. Thus the implementation of this in this techniques is only based upon 2 classes. Thus is the target class is polynomial this implementation fails miserably.

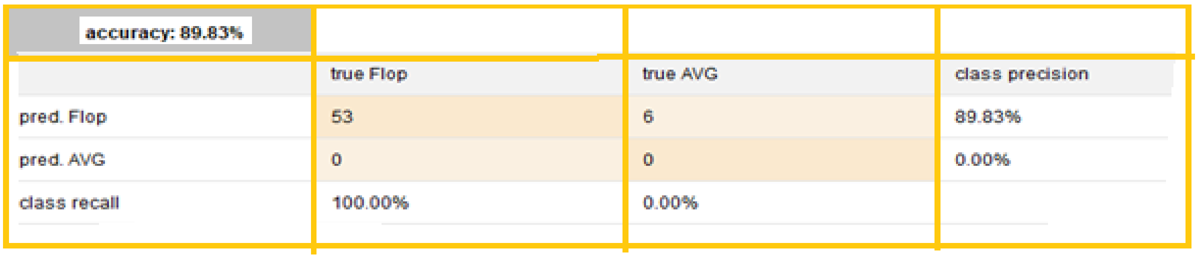
Operator used for Logistic Regression in Rapid Miner: This Operator helps in implementing Logistic Regression in Rapid Miner which helps us in setting up the Solver as default and standardizing it, adding intercept, computing p values, removing collinear columns and meanlipulation i.e. mean manipulation of missing values over the model.

Computation of Multiple Logistic Regression Models implemented:

Different Logistic Regression Models are performed based upon the distinct combinations of attributes and changes affect accuracies of prediction. The iterations performed in logistic regression technique restricts the evaluation of target class as binomial attribute only which is a major drawback

Table 2: Logistic Regression Model Iterations Accuracy

Model	Attributes Considered	Attributes Type	Accuracy
Logistic Regression	Default	General as earlier	89.83%
Logistic Regression 1.1	Without ROI Attribute	General as earlier	88.13%
Logistic Regression 1.2	Without Budget Attribute	ROI as polynomial	86.14%



accuracy: 89.83%

	true Flop	true AVG	class precision
pred. Flop	53	6	89.83%
pred. AVG	0	0	0.00%
class recall	100.00%	0.00%	

Figure 8: Accuracy of Logistic Regression Model.

Evaluation of Logistic Regression Models: The highest accuracy attained in this case is 89.93% which is better than the previous model in which it was implemented Early predictions of movie success (Lash and Zhao; 2016) in which the decision tree accuracy is 79.3%. however the improvement cane be rendered of no use as it predicts only 2 classes. As this prediction model fails to cater the needs of the data.The accuracy is mentioned in the Figure 8 above.

Visualisation of Logistic Regression Model: This visualisation is based upon three major parameters of the dataset ROI, Budget and Gross mapped together on a 3D surface plot implemented using Rapid Miner Visualisations as seen in the Figure 9 below. In this plot the Z axis represents the count while the distinct colours are used to identify the attributes used for the visualisation. It is evident that the Gross scores over the 2 attributes while the ROI lags behind.

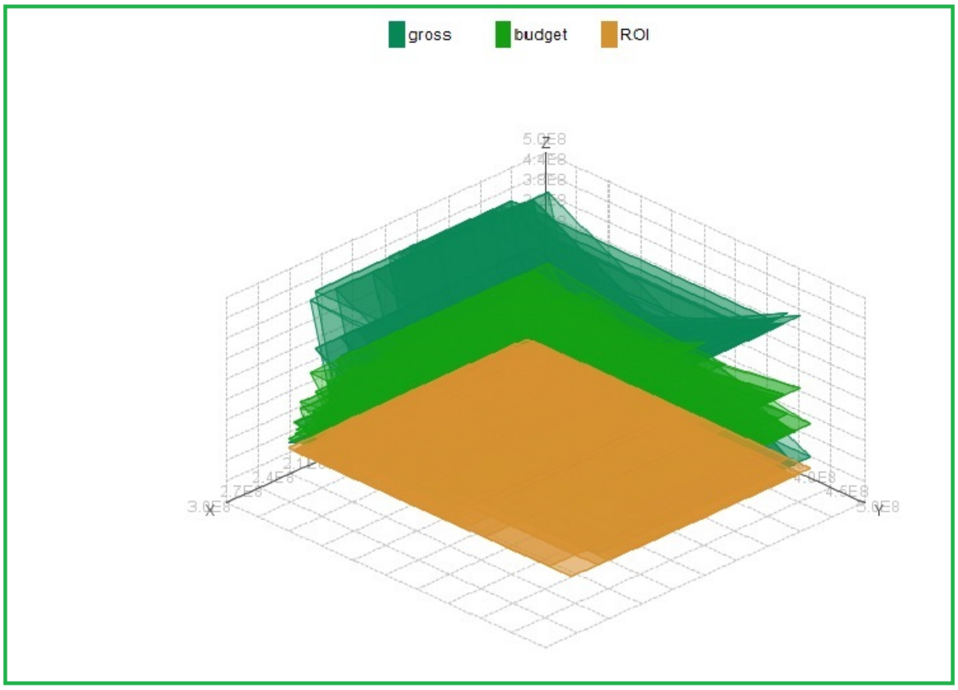


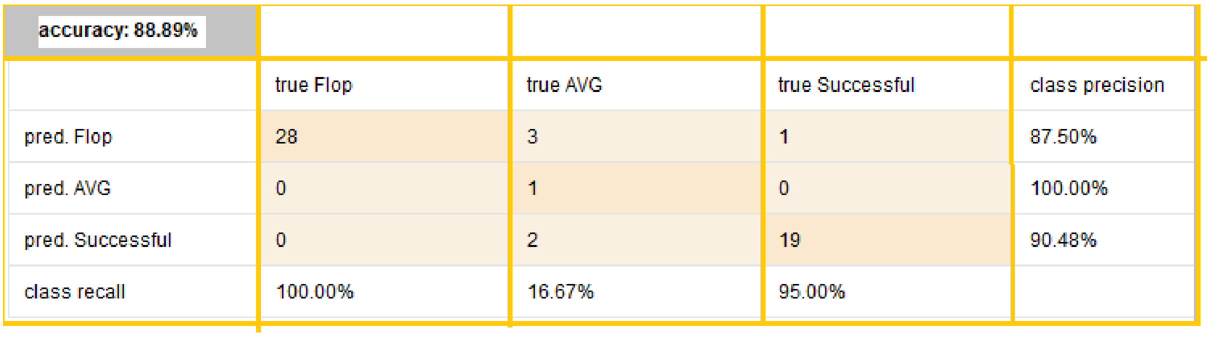
Figure 9: Visualization of Logistic Regression Model.

4.3.4 Implementation, Computation, Analysis and Evaluation of Random Forest based on accuracy and precision for the system

Random Forest Concept: It is a classification machine learning algorithm which has been derived from Decision Tree it utilizes multiple decision tree and evaluates the best option out of them. (Liaw et al.; 2002).
Advantages of Random Forest: The biggest advantage of Random Forest used both for classification as well as regression
Disadvantages of Random Forest: The major flaw of Random Forest is that it at times leads to over fitting or under fitting in Rapid Miner.
Operator used for Random Forest in Rapid Miner: This Operator helps in implementing Random Forest in Rapid Miner. The set-up of Random Forest is pretty like decision tree providing criterion as gain ratio a maximum depth of 20 and application of pruning and pre-pruning over the model along with the options of no of trees to be selected along with no size of pre-pruning and minimal size of split.
Computation of Multiple of Random Forest Models: There have been 3 different Random Forest models implemented. However as mentioned earlier in the Disadvantage section the model over fits for Flop target class which becomes the biggest drawback of the model.

Table 3: Random Forest Model Iterations Accuracy

Model	Attributes Considered	Attributes Type	Accuracy
Random Forest	Default	General as earlier	88.89%
Random Forest 1.1	Change in ROI Attribute	ROI as polynomial	80.00%
Random Forest 1.2	Change in ROI Attribute	ROI as polynomial	66.90%



accuracy: 88.89%				
	true Flop	true AVG	true Successful	class precision
pred. Flop	28	3	1	87.50%
pred. AVG	0	1	0	100.00%
pred. Successful	0	2	19	90.48%
class recall	100.00%	16.67%	95.00%	

Figure 10: Accuracy of Random Forest Model.

Evaluation of Random Forest Models: The highest accuracy attained in Random Forest Model implemented in this case is 88.89% which is not as good as the previous model implemented by (Lash and Zhao; 2016) had accuracy of 90.21%. So, Random Forest us lacking behind in this case. The biggest issue being over fitting of the taregt class Flop.As shown in the figure 10

Visualization of Random Forest Model: This visualization has been deployed upon the three main classes ion which prediction has been performed thus ,making it one of the most important visualizations of the project. As mentioned earlier in Feature Selection Section. It is quite obvious that the Successful movies rank a lot higher as compared to Average movies. While the average movies generally hover in the middle range. As can be observed in Figure 11. The lower frame is occupied by the Flop movie domains. The x axis of the graph is ROI of the movies while the y axis is the gross the plotting thus totally dependent upon the success ratio of the movie the higher the ROI and the higher the gross the more successful is the movie. It is a Bubble graph where the different classes are distinguished by distinct colours that are Green for Successful, Red for Average and Blue for Flop. The visualization of this chart is performed in Rapid Miner.

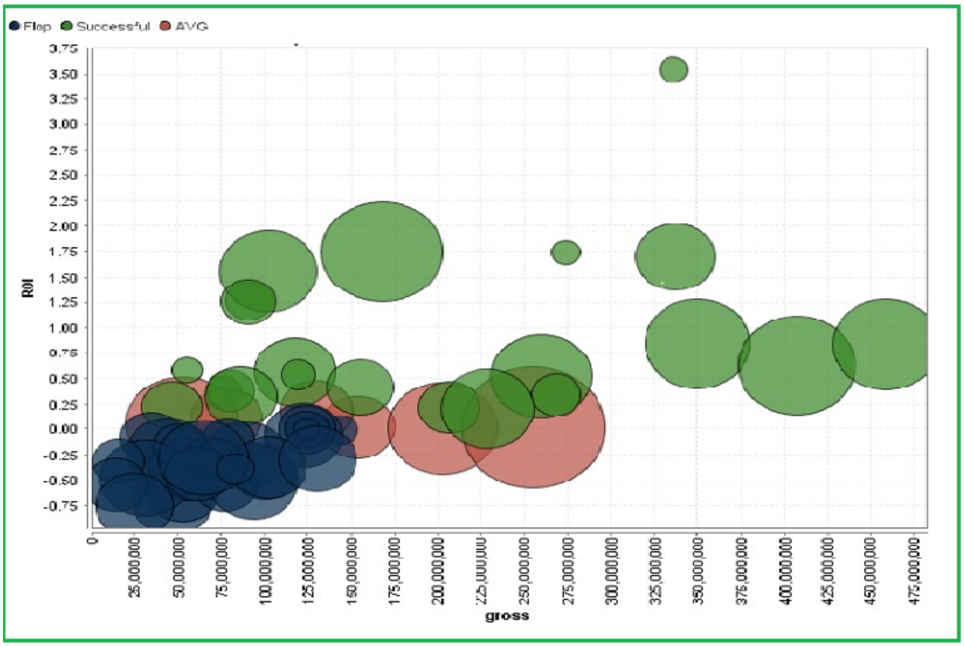


Figure 11: Visualization for Random Forest Model.

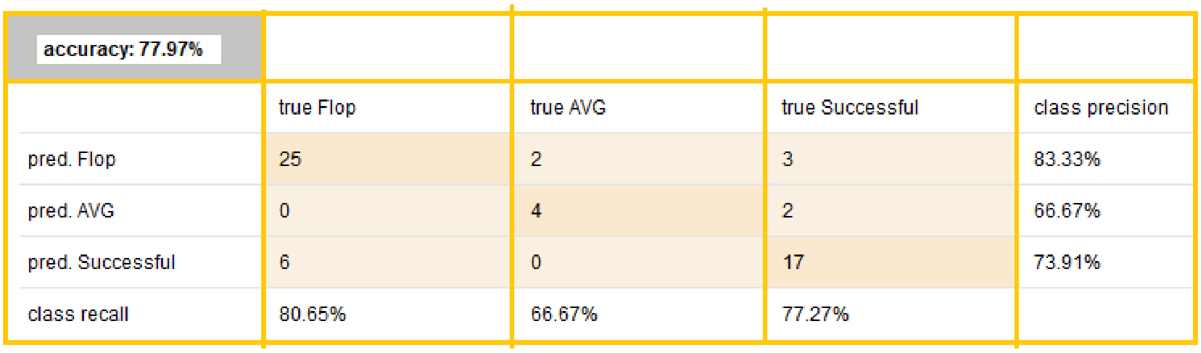
4.3.5 Implementation, Computation, Analysis and Evaluation of Nave Bayes based on accuracy and precision for the system

Nave Bayes Concept: It is a classification machine learning algorithm which are based upon on probability and occurrences of events because of which it enables us with maximum no of iterations. It is dependent upon the concept of conditional probability (Liaw et al.; 2002).
P(H—E)=P(E—H)*P(E)/P(E)
P(H) is the probability of hypothesis H being true. This is known as the prior probability.
P(E) is the probability of the evidence (regardless of the hypothesis).
P(E—H) is the probability of the evidence given that hypothesis is true.
P(H—E) is the probability of the hypothesis given that the evidence is there.
Advantages of Nave Bayes: Naive Bayes Algorithm is a fast, highly scalable algorithm easy to train on small datasets. It provides the options for Binary and Multiclass classification and even further options among them such GaussianNB, MultinomialNB, BernoulliNB.
Disadvantages of Nave Bayes: If the no of probability cases is really high it forms a complicated structure.
Operator used for Nave Bayes in Rapid Miner: This Operator helps in implementing Nave Bayes in Rapid Miner. It allows us to perform Laplace correction that counters the influence of zero probabilities. As the dataset is huge and has multiple attributes there might be multiple cases where even the single attribute and a combination of different attribute either do not provide or have negligent effect on the overall predictive model also the range is considered as Boolean and the default is considered as true. (Patil and Sherekar; 2013).
Computation of Multiple of Nave Bayes Models: Different Nave Bayes Models are performed based upon the distinct combinations of attributes and how the changes

affect accuracy of the overall prediction. As it is quite evident that it provides maximum number of iterations in all the Machine Learning Technique with maximum no of possibilities. There 6 iterations that have been implemented in this scenario.

Table 4: Nave Bayes Model Iterations Accuracy

Model	Attributes Considered	Attributes Type	Accuracy
Nave Bayes	Default	General as earlier	75.68%
Nave Bayes 1.1	Without YouTube parameters	General as earlier	77.97%
Nave Bayes 1.2	With ROI Attribute included	General as earlier	70.27%
Nave Bayes 1.3	Without Facebook included	General as earlier	48.65%
Nave Bayes 1.4	Without Budget Attribute	General as earlier	64.85%
Nave Bayes 1.5	Without Gross Attribute	General as earlier	72.97%



accuracy: 77.97%

	true Flop	true AVG	true Successful	class precision
pred. Flop	25	2	3	83.33%
pred. AVG	0	4	2	66.67%
pred. Successful	6	0	17	73.91%
class recall	80.65%	66.67%	77.27%	

Figure 12: Accuracy of Nave Bayes Best Model.

Evaluation of Nave Bayes Models: The highest accuracy attained in Nave Bayes Model is 77.97% including al the essential attributes. In contrast to 68.6%(Lash and Zhao; 2016) it is a significant increase in accuracy of the model even after adding the extra parameters. As it can be seen in Figure 12.

Visualization of Nave Bayes Model: The visualization of the Nave Bayes model is based upon You Tube parameters. The major influence in the YouTube platform is View, Like and dislike count of the trailer this helps us in understanding the interests of the customer in watching the movie on the basis of the trailer that has been released. Over the years it has been established that this plays a major role in setting up the movie positive marketing campaign.As it can be seen in Figure 13. This visualization is performed on a multilinear graph with the three parameters distinguished by distinct colours that have been categorized on the side of the line chart. The visualization has been performed in rapid Miner using the Multilinear Chart type. (Patil and Sherekar; 2013).

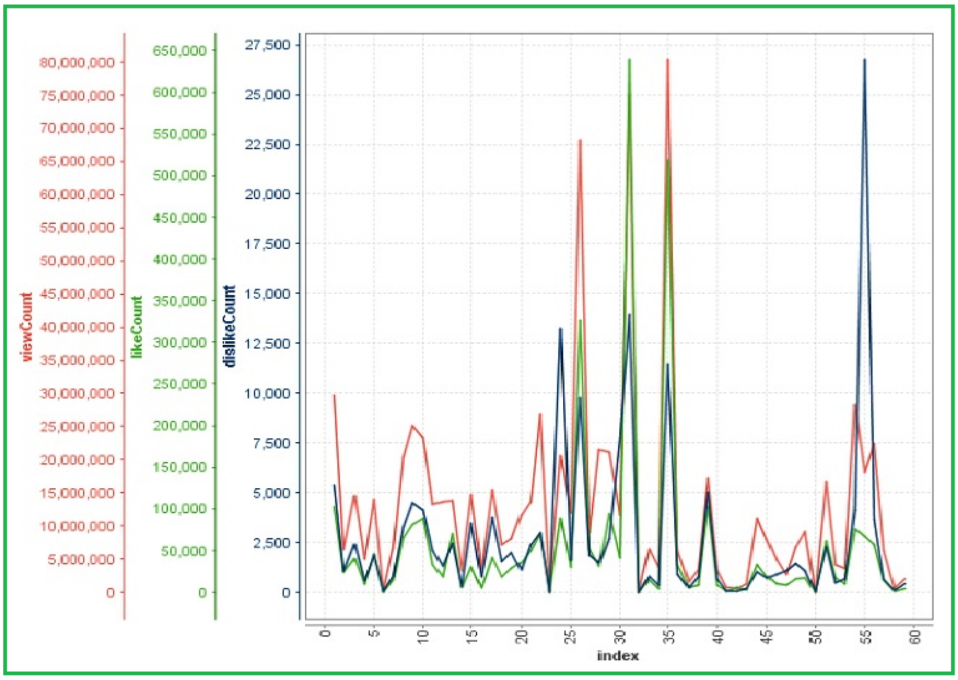


Figure 13: Visualization for Naive Bayes Model.

4.3.6 Implementation, Computation, Analysis and Evaluation of SVM based on accuracy and precision for the system

SVM Concept: It is a classification machine learning algorithm which is used to address multi-classification problems, regression and outlier detection with intuitive model. When the target class is not defined we can perform Support Vector Clustering. In this prediction model there is only one target class that needs to be addressed. The main component of SVM is hyperplane that divides the datasets into different classes. (Chang and Lin; 2011).

Advantages of SVM: SVMs are effective on substantial number of features. The functioning is more efficient when the number of features are more than the number of samples.

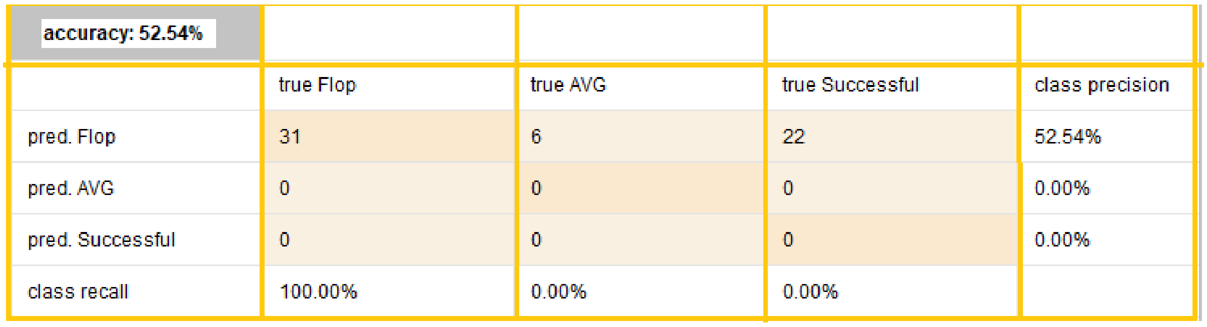
Disadvantages of SVM: With greater number of samples, it starts giving poor performances.It also has another restrictions in regards to its hyperplane.

Operator used for SVM in Rapid Miner: This Operator helps in implementing Random Forest in Rapid Miner. The set-up of SVM in Rapid Miner gives kernel options as well as convergence epsilon, kernel cache and maximum iteration. Scaling is another salient feature provided by the Operator.

Computation of Multiple of SVM Models: Different SVM Models are performed based upon the distinct combinations of attributes and how the changes affect accuracy of the overall prediction. SVM computations differ the accuracy largely there have been 4 distinct iterations the main model is selected SVM 1.1 which has the best accuracy among the other iterations.

Table 5: SVM Model Iterations Accuracy

Model	Attributes Considered	Attributes Type	Accuracy
SVM	Default	General as earlier	49.93%
SVM 1.1	Without YouTube parameters	ROI as polynomial	52.54%
SVM 1.2	With ROI Attribute included	General as earlier	47.54%
SVM 1.3	Without Facebook included	General as earlier	48.49%



accuracy: 52.54%

	true Flop	true AVG	true Successful	class precision
pred. Flop	31	6	22	52.54%
pred. AVG	0	0	0	0.00%
pred. Successful	0	0	0	0.00%
class recall	100.00%	0.00%	0.00%	

Figure 14: Accuracy of SVM Model.

Evaluation of the main SVM Models: This machine learning technique has not been implemented in the previous paper based upon which standard comparisons are being made. The accuracy provided by is SVM machine learning is 52.54%. this technique not implemented in the previous paper adds an extra set of comparative analysis to Prediction Models for Box Oce Revenue before Theatrical Release of Movies. However, this implementation of this technique can be considered as in applicable on the datasets as it fails miserably in all the target classes as it over fits the flop class and under fits the other two classes.As it can be observed in Figure 14.

Visualization of SVM Model: A visualization that maps all the defaults attributes used in Random Forest is visualized using Rapid Miner. This visualization is performed 3 D sticks maps.As it can be observed in Figure 15. All the attributes are distinguish ably visualized using distinct colours. The major advantage of this map is to identify outliers in the model. However, since the data is cleansed and mapped accordingly this map does not shows any outliers. (Russom et al.; 2011).

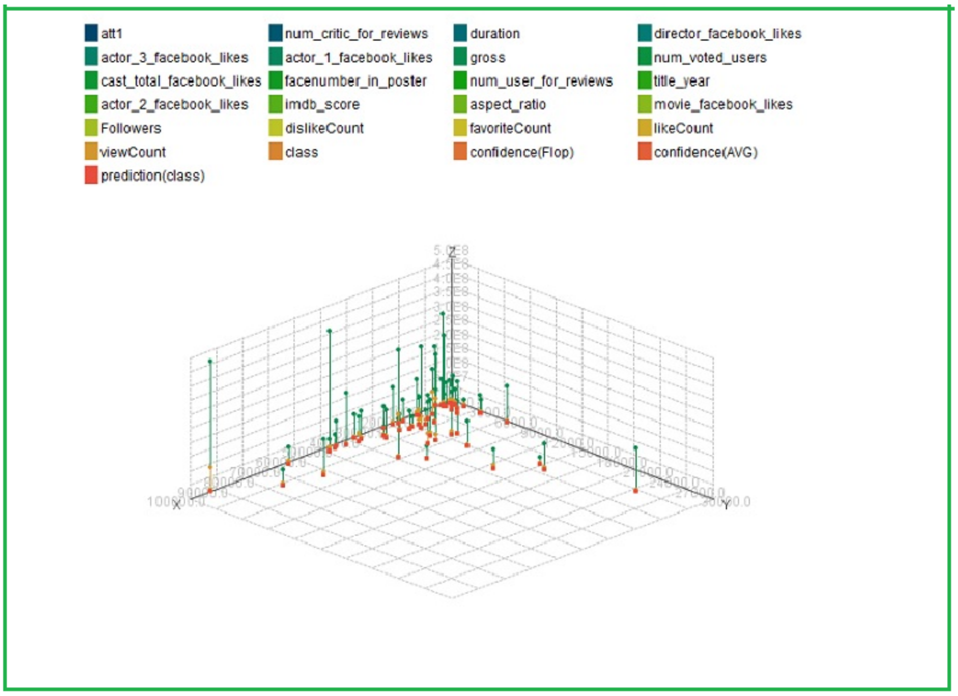


Figure 15: Visualization for SVM models.

4.4 Conclusion

This chapter cover ups all the machine learning techniques that have been implemented. These techniques are analyzed, computed, evaluated and visualized. Thus answering the Research Question in (Section 1.2.1) and all the Research Objectives mentioned in Section (1.2.3) have been tackled. While Decision Tree and Naive Bayes model were successfully implemented. Random forest Models lead to over fitting of a class. On the other hand logistic regression failed to compute the third class and only considered binomial class which was totally against the datasets predictive analysis. SVM even after considering all the classes either under fit or over fit the classes making the technique unsuccessful. Overall if all the techniques are compared Decision Tree Models and Naive Bayes Models were successfully implemented with good accuracy.

5 Conclusion and Future Work

Conclusion: The Prediction models provided detailed information about the attributes that plays a significant role in the success of the movie. There have been multiple prediction models implemented and a analyzed in this domain have been specified in Section 2.2. However, there are certain significant attributes that were not accounted for previously in those models which have been identified in Section 2.3 in detail. The identification of gaps and building the model countering these gap is the motivation for the project. It has been implemented using certain major social media parameters that has never been considered earlier which differentiates this model from others. YouTube and Twitter are social media platform that help us in analyzing the buzz around the movie which were not considered earlier comprehensively together. YouTube provides an insight about the popularity of the movie while Facebook and Twitter analyze the popularity of people

involved in the movie making process. The pivotal features in regards to the movie were extracted directly from a data repository data.world while programs were used to analyze social media portals such as YouTube and Twitter which were extracted directly using R programming language in CSV format. These datasets are filtered and combined using R programming two main attributes based upon the existing datasets are created that help in developing the target class. ROI attribute is created using Excel on the formula mentioned in section 3.3.1 while the target class is created using the loop structure in R Studio. The dataset is cleansed using Rapid Miner and multiple Machine learning techniques are implemented over the dataset to attain the desired result and visualizations which are performed either in Tableau or Rapid Miner. The performance for the attained results are evaluated on metrics such as accuracy and prediction class. The visualizations present a graphical representation of distinct attributes of the datasets and their influence in the prediction model. The system adheres to the objectives specified in Section 1.2. The data preparation process in the Prediction Models for Box Oce Revenue before Theatrical Release of Movies focused upon certain aspects that were not considered earlier in the revious prediction models. The technologies learned while implementing the project Rapid Miner and the functioning and implementation of Machine Learning Techniques. Libraries and looping structure of R programming were also learned. A detailled evaluation of Prediction Models for Box Oce Revenue before Theatrical Release of Movies was implemented and evaluated in the technical report covering up the concept and the minor nuances of the model.
Future Works: The Prediction Models for Box Oce Revenue before Theatrical Release of Movies can be further improvised by adding certain new attributes and implementation in different technologies to improvise the accuracy of the prediction models which can add new dynamics to the Prediction Models. Releasing a movie on a big holiday weekend can give a big push to the revenue of the movie thus becoming a key factor for analysis. A similar model can be implemented for release of games, comic books and novel. A totally different Prediction model can also be designed for Superhero Genre movies as the attributes of these movies are distinct popularity of these such as popularity of the Comic characters of the movies, Comic book adapted for the movie, Rating of the comic book, etc. Another novel feature can be analyzed for future models is the revenue the production houses generate through merchandise sales and game launches these sales can bring in significant amount of revenue in animation movies.

6 Acknowledgement

The author appreciates the contribution of Dr Catherine Mulwa in development of the project her guidance and support. Dr Jason Roche contribution in the initial stages of the project proposal development has also been very valuable in the concept and idea development of the project. The author is very grateful to data.world data repository for providing the primitive dataset around which the other datasets were build and extracted.

References

Azevedo, A. I. R. L. and Santos, M. F. (2008). Kdd, semma and crisp-dm: a parallel overview, *IADS-DM* .

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines, *ACM transactions on intelligent systems and technology (TIST)* **2**(3): 27.

Dave, K., Lawrence, S. and Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, *Proceedings of the 12th international conference on World Wide Web*, ACM, pp. 519–528.

Hsieh, F. Y., Bloch, D. A. and Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regression, *Statistics in medicine* **17**(14): 1623–1634.

Kayacik, H. G., Zincir-Heywood, A. N. and Heywood, M. I. (2005). Selecting features for intrusion detection: A feature relevance analysis on kdd 99 intrusion detection datasets, *Proceedings of the third annual conference on privacy, security and trust*.

Kurt, I., Ture, M. and Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease, *Expert systems with applications* **34**(1): 366–374.

Lash, M. T. and Zhao, K. (2016). Early predictions of movie success: the who, what, and when of profitability, *Journal of Management Information Systems* **33**(3): 874–903.

Liaw, A., Wiener, M. et al. (2002). Classification and regression by randomforest, *R news* **2**(3): 18–22.

Parimi, R. and Caragea, D. (2013). Pre-release box-office success prediction for motion pictures, *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, Springer, pp. 571–585.

Patil, T. R. and Sherekar, S. (2013). Performance analysis of naive bayes and j48 classification algorithm for data classification, *International Journal of Computer Science and Applications* **6**(2): 256–261.

Russom, P. et al. (2011). Big data analytics, *TDWI best practices report, fourth quarter* **19**: 40.

Sabhnani, M. and Serpen, G. (2003). Application of machine learning algorithms to kdd intrusion detection dataset within misuse detection context., *MLMTA*, pp. 209–215.

Scott, S. and Matwin, S. (1999). Feature engineering for text classification, *ICML*, Vol. 99, pp. 379–388.

Sharda, R. and Delen, D. (2006). Predicting box-office success of motion pictures with neural networks, *Expert Systems with Applications* **30**(2): 243–254.

Smith, L. and Tansley, J. (2003). Decision tree analysis. US Patent App. 10/406,836.

Wang, Z., Yu, X., Feng, N. and Wang, Z. (2014). An improved collaborative movie recommendation system using computational intelligence, *Journal of Visual Languages & Computing* **25**(6): 667–675.

Williams, G. J. and Huang, Z. (1996). A case study in knowledge acquisition for insurance risk assessment using a kdd methodology, *Proceedings of the Pacific Rim Knowledge Acquisition Workshop, Dept. of AI, Univ. of NSW, Sydney, Australia*, pp. 117–129.