INTRO PAGE

I’m unable to remove page borders to all pages.

Either border or header-footer has to be used.

CERTIFICATE PAGE

ABSTRACT

With growing volumes and types of data and piquing interest in using data to produce valuable insights, it has become one of the most important areas of study in today’s era. Huge datasets are available for predictive analysis of several aspects of movies and many domains are available for making predictions. It is beneficial to all varieties of people associated with the art of movie making and watching. Stakeholders like producers can know the risks and advantages of investing in particular movies. Movie watchers can determine if the movie is up to the mark and worth their money. This paper aims to explore the different techniques used for predictive analysis. We also seek to explore what factors are necessary to predict the quality of a movie in terms of its concept and how to establish a relation between different categories.

The objective of this project is to work on the dataset available and identify various factors affecting movie ratings and thereby the quality. These parameters are further used to predict the ratings of the movie before it is released.

We used data analytics to achieve this objective. Data analytics comprises of many different ways to analyse a given dataset. For getting appropriate results from our datasets, we used multiple linear regression to train the model and then predict the results.

ACKNOWLEDGEMENT

We would like to express our gratitude **DR. C P S Prakash**, Principal of DSCE, for permitting us to utilise the resources provided by our college for the project.

We are also very grateful to our respected Vice Principal, HOD of Computer Science and Engineering Department, DSCE , **Dr. Ramesh Babu** for his encouragement and support.

We are immensely thankful to our learned and respected guide, **Dr. Kusuma H**, who constantly guided us throughout all the phases of the project and helped us with the technicalities of the project with her expertise. We deeply express our gratitude for her support.

We would also like to thank our project co-ordinator, **Dr.Vindhya** for providing us the opportunity to work on our projects and giving a wide range of topics to select from. We are very thankful for her constant support.

We would also like to thank other faculty and staff members for their kind co-operation and help.

Lastly, we would like to acknowledge our family members and classmates for providing us with moral support and encouragement.

1

Introduction

* 1. Movie Prediction

Movie making involves huge investment thus movie prediction plays a vital role in the movie industry. Movies is the most convenient way to entertain people. However only few movies get higher success and are ranked high. Many movies are produced by the movie industry in a year.

A movie revenue depends on various components such as cast acting in a movie, director of the movie, film critics’ review, rating for the movie, genre of the movie, etc. Because of these multiple components there is no formula that helps us to provide analysis for predicting how much revenue a particular movie will be generating.

However, by analysing the IMDB score generated by previous movies, a model can be built which can help us predict the expected quality for a particular movie. As we know in today’s world, the movie is one of the biggest source of entertainment and for business purposes. To expand this business further we need the technology through which we can predict the success rate of the movie.

Success rate of movies, models and mechanisms can be used to predict the success of a movie. It will help the viewers whether to watch the movie or not as the quality of the movie will be predicted. Stakeholders such as actors, producers, director etc. can use these predictions to make more informed decisions. They can make the decision before the movie is released.

This proposed work aims to develop a model based upon the data mining techniques that may help in predicting the success of a movie in advance thereby reducing certain level of uncertainty. The excellent way to find detailed information about almost every film ever made is through IMDB.

1.2 Multiple Linear Regression

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable.

In essence, multiple regression is the extension of ordinary least-squares (OLS) regression that involves more than one explanatory variable.

The Formula for Multiple Linear Regression is

​*yi* ​= *β*0 + *β*1​*xi*1 + *β*2​*xi*2 ​+ ... + *βp*​*xip* ​+ *ϵ*

where, for *i* = *n* observations:

*yi*​ = dependent variable

*xi*​ = expanatory variables

*β*0​ = y-intercept (constant term)

*βp* = slope coefficients for each explanatory variable

*ϵ* = the model’s error term (also known as the residuals)​

A simple linear regression is a function that allows an analyst or statistician to make predictions about one variable based on the information that is known about another variable. Linear regression can only be used when one has two continuous variables—an independent variable and a dependent variable. The independent variable is the parameter that is used to calculate the dependent variable or outcome. A multiple regression model extends to several explanatory variables.

The multiple regression model is based on the following assumptions:

* There is a linear relationship between the dependent variables and the independent variables.
* The independent variables are not too highly correlated with each other.
* yi observations are selected independently and randomly from the population.
* Residuals should be normally distributed with a mean of 0 and variance σ.

The coefficient of determination (R-squared) is a statistical metric that is used to measure how much of the variation in outcome can be explained by the variation in the independent variables. R2 always increases as more predictors are added to the MLR model even though the predictors may not be related to the outcome variable.

R2 by itself can't thus be used to identify which predictors should be included in a model and which should be excluded. R2 can only be between 0 and 1, where 0 indicates that the outcome cannot be predicted by any of the independent variables and 1 indicates that the outcome can be predicted without error from the independent variables.

When interpreting the results of a multiple regression, beta coefficients are valid while holding all other variables constant ("all else equal"). The output from a multiple regression can be displayed horizontally as an equation, or vertically in table form.

1.3 Organisation of the Project Report

To be done at last

**TO BE USED FOR SOME OTHER PURPOSE (EVEN YOU CAN USE IT)**

Vast amount of data, which contains much valuable information about general trends in films. Data mining techniques enable us to uncover information which will both confirm or disprove common assumptions about movies, and also allow us to predict the success of a future film given select information about the film before its release. So here we are developing the software for data analytics through which we can predict the success rate of the movie with high accuracy. Here we are using the R-software to predicting the movie success rate into which first we have downloaded the data set from kaggle.com and after that we are generating the training and test data set. In a dataset, a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. The main attributes selected for building model are critics\_score, imdb\_rating, imdb\_num\_votes, audience\_score. Data points in the training set are excluded from the test (validation) set. Usually, a dataset is divided into a training set, a validation set (some people use 'test set' instead) in each iteration, or divided into a training set, a validation set and a test set in each iteration. These training and test dataset is used to build model for selected set of attributes. On the basis of the generated model prediction have been done and result have been generated. Through the acquired result we can easily conclude that the movie is hit, superhit,

blockbuster or flop. The outcome of this research is therefore twofold, it provides tools and techniques to transform the database data into a format suitable for data mining, and provides a selection of information mined from this refined data.

2

Problem Statement and Proposed Solution

2.1. Problem Statement

To identify the categories that affect the quality of a movie and develop a model that takes the required parameters to predict the quality of an upcoming movie based on IMDB score.

2.2. Existing Models

There are several existing models that have attempted to identify how a movie is affected based on different criteria like-

1. A model identified

2.3 Proposed Solution

The proposed solution for the problem is using data analytics. The model takes in the identified parameters which affect the quality of movies. The multiple linear regression model takes these inputs and predicts the score based on weights assigned by the model during training.

Input the identified parameters

Predict the score of the movie using the weights obtained.

Calculate the weightage for each parameter based on multiple linear regression.

The proposed has two phases – Training phase and Testing phase. The processing in each phase is shown as in fig. 2.2.

Training Phase Testing Phase

Input Parameters

Input Parameters

Estimate weights

MLR Model

Predict the score

Update the weights to correct the model.

Compare predicted output with observed output.

3

Literature Survey

**[1] Sentiment Analysis of Movie Review Comments**

**Authors: K. Yessenov and S. Misailovic**

**Description:** This paper presents an empirical study of efﬁcacy of machine learning techniques in classifying text messages by semantic meaning. They use movie review comments from popular social network Digg as our data set and classify text by subjectivity/objectivity and negative/positive attitude. They propose different approaches in extracting text features such as bag-of-words model, using large movie reviews corpus, restricting to adjectives and adverbs, handling negations, bounding word frequencies by a threshold, and using WordNet synonyms knowledge. They evaluate their effect on accuracy of four machine learning methods-Naïve Bayes, Decision Trees, Maximum-Entropy, and K-Means clustering. They conclude our study with explanation of observed trends in accuracy rates and providing directions for future work.

**[2] Deep Learning for Sentiment Analysis of Movie Reviews**

**Authors: H. Pouransari, & S. Ghili**

**Description:** In this study, they explore various natural language processing (NLP) methods to perform sentiment analysis. We look at two different datasets, one with binary labels, and one with multi-class labels. For the binary classiﬁcation, they applied the bag of words, and skip-gram word2vec models followed by various classiﬁers, including random forest, SVM, and logistic regression. For the multi-class case, they implemented the recursive neural tensor networks (RNTN). To overcome the high computational cost of training the standard RNTN they introduce the low-rank RNTN, in which the matrices involved in the quadratic term of RNTN are substituted by symmetric low-rank matrices. They show that the low-rank RNTN leads to signiﬁcant saving in computational cost, while having a similar accuracy as that of RNTN.

**[3] Rating based Mechanism to Contrast Abnormal Posts on Movies Reviews using MapReduce Paradigm**

**Author: Piyush Gupta, Atul Sharma, Jitender Grover**

**Description:** BigData contains large amount of unstructured data in the form of movie data, facebook data, and industry data and so on. There are number of posts are posted on twitter about movies by different users. Out of these posts some of posts may be inappropriate. These posts contain negative comments as well as positive comments about movies. It is difficult to distinguish large number of positive and negative posts. To overcome this kind of problem we proposed a rating based mechanism that distinguishes abnormal posts with the help of users rating. If rating is positive then post is normal otherwise it is abnormal. To implement proposed mechanism we used hadoop platform and MapReduce paradigm.

**[4] Movie Success Prediction using Machine Learning Algorithms and their Comparison**

**Authors: Rijul Dhir, Anand Raj**

**Description:** The number of movies produced in the world is growing at an exponential rate and success rate of movie is of utmost importance since billions of dollars are invested in the making of each of these movies. In such a scenario, prior knowledge about the success or failure of a particular movie and what factor affect the movie success will benefit the production houses since these predictions will give them a fair idea of how to go about with the advertising and campaigning, which itself is an expensive affair altogether. So, the prediction of the success of a movie is very essential to the film industry. In this proposed research, we give our detailed analysis of the Internet Movie Database (IMDb) and predict the IMDb score. This database contains categorical and numerical information such as IMDb score, director, gross, budget and so on and so forth. This research proposes a way to predict how successful a movie will be prior to its arrival at the box office instead of listening to critics and others on whether a movie will be successful or not. The proposed research provides a quite efﬁcient approach to predict IMDb score on IMDb Movie Dataset. We will try to unveil the important factors inﬂuencing the score of IMDb Movie Data. We have used different algorithms in the research work for analysis but among all Random forest gave the best prediction accuracy, which is better in comparison to the previous studies. In the exploratory analysis, we found that number of voted users, number of critics for reviews, number of Facebook likes, duration of the movie and gross collection of movie affect the IMDb score strongly. Drama and Biopic movies are best in genres.