Movie Revenue Prediction

Capstone Project Report

Mid-Semester Evaluation

Submitted by:

101503241 Vibhor Sharma

101503245 Vishal Sethi

101553006 Prateek Awasthi

401503030 Varun Jain

BE Third Year, CSE/SEM

CPG No: __137__

Under the Mentorship of

Mrs. Swati Kumari

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Computer Science and Engineering Department

TIET, Patiala

MAY 2018

ABSTRACT

Film studios in America, every year produces several hundred movies that make the United States the third most abundant producer of films in the world. The budget of these movies are of the order of hundreds of millions of dollars, making their box office success absolutely essential for the survival of the industry. Knowing which movies are likely to succeed and which are likely to fail before the release could benefit the production houses greatly as it will enable them to focus their advertising campaigns which itself cost millions of dollars, accordingly. And it could also help them to know when it is most appropriate to release a movie by looking at the overall market. So, the prediction of movie success is of great importance to the industry. Machine learning algorithms are widely used to make predictions such as growth in the stock market, demand for products, nature of tumors, etc. This paper presents a detailed study of Logistic Regression, SVM Regression, and Linear Regression on IMDb data to predict movie box office.

DECLARATION

We hereby declare that the design principles and working prototype model of the project entitled Movie Revenue Prediction is an authentic record of our own work carried out in the Computer Science and Engineering Department, TIET, Patiala, under the guidance of Mrs. Swati Kumari during 6th semester (2018).

Date: 30th May,2018

Roll No.	Name	Signature
101503241	Vibhor Sharma	
101503245	Vishal Sethi	
101553006	Prateek Awasthi	
401503030	Varun Jain	

Counter Signed By:

Faculty Mentor:

Mrs. Swati Kumari

Computer Science & Engineering Department,

TIET, Patiala

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Table 1.

LIST OF ABBREVIATIONS

DCU	Data Collection Unit
TF	TensorFlow
SVM	Support Vector Machine
DFD	Data flow Diagrams
MLP	Multi-level Perceptron
CSV	Comma Separated File
HTTP	Hypertext Transfer Protocol
SDK	SDK - Software Development Kit
XML	XML - Extensible Markup Language

1.1 Project Overview

More than 3000 films are released every year on an average in the world. Surprisingly, the country with the largest number of film releases has been India for a long time due to its multilingual and diverse character. A lot of people depend on the film industry for their livelihood and some have even made large fortunes by betting the right way. However, a yet larger number of people have lost money betting the wrong way.

For every film, the production team is faced with many choices such as who to cast, genre etc. Even actors at different points of time in their careers are at different levels of popularity and thus, price their services accordingly. It is crucial for producers to make decisions like these with a mathematical model predicting results based on such inputs.

Estimating with a reasonable accuracy, the revenue of films based on factors such as their cast, genre, language etc. can go a long way in helping producers, corporations and their investors in making smart choices and informed decisions about the way they allocate capital for different projects. This optimization in budgets of different avenues of a film's production such as marketing, advertising etc. will also be pivotal in maximizing returns for every dollar invested and would substantially reduce bad investments and wastage of capital. Thus, economizing the entire film industry and helping it achieve its full potential in terms of fiscal as well as social output.

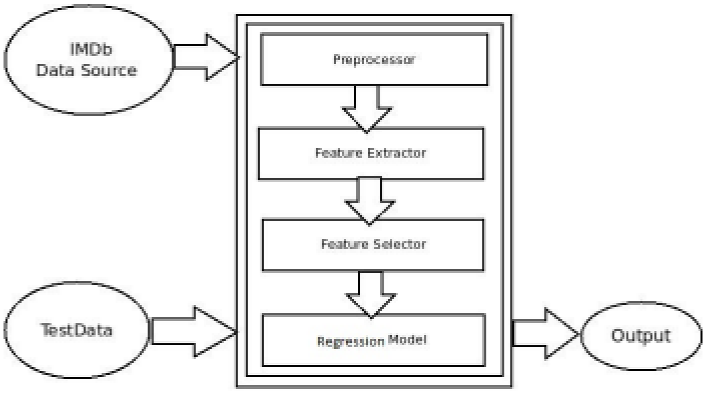


Figure 1: Brief Project Overview

1.1.1. The Data Collection Unit

We have used the famous IMDB 5000 Movie Dataset provided by kaggle. This dataset contains 5000+ movie metadata scraped from IMDB. It contains 28 variables for 5043 movies, spanning across 100 years in 66 countries. There are 2399 unique director names, and thousands of actors/actresses.

The variables are listed below:

Movie Title, Color, Number of Critics for Reviews, Movie’s Facebook Likes, Du-ration, Director’s Name, Director’s Facebook Likes, Actor 3’s Name, Actor 3’s Facebook Likes, Actor 2’s Name, Actor 2’s Facebook Likes, Actor 1’s Name, Actor 1’s Facebook Likes, Gross, Genre, Number of voted users, Cast’s total Facebook likes, Number of faces on poster, Plot Keywords, Movie’s IMDB Link, Number of user for reviews, Language, Country, Content Rating, Budget, Title Year, IMDB Score, Aspect Ratio.

We actually didn’t use these many features to train and predict using our models. These features have continuous values so they are all normalized so that, they have values within range 0 to 1. The features we have used to train and test are listed below:

Director’s Facebook Likes, Number of Critics for Reviews, Actor 3’s Facebook Likes, Actor 2’s Facebook Likes, Actor 1’s Facebook Likes, Movie’s Facebook Likes, Duration, Number of voted users, Number of user for reviews, Budget, Year, IMDB Score, Total Facebook Likes.

Then the Genres are kept in the following categories and they are considered as binary features meaning if a movie has the following genre Action, Adventure, Mystery the feature value will be like this, [1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0]: Action, Adventure, Fantasy, Sci-fi, Thriller, Comedy, Family, Horror, War, Animation, Western, Romance, Musical, Documentary, Drama, History, Biography, Mystery, Crime.

The content rating is also considered as binary feature. But it only refers to a certain category. So the feature will be like this for Parental Guidance Suggested, [0, 1, 0, 0, 0]. The categories of content ratings are listed below:

General Audiences, Parental Guidance Suggested, Parents Strongly Cautioned, Restricted, Adults Only.

For the features Language & Country, we have kept 1 if the language is English and 0 for all other languages & for USA, 1 and 0 for others respectively.

1.1.2 Backend

The backend is the central unit that controls all the other parts of the application. Backend is built using nodeJS and the database used is mongo dB. Backend serves all the assets, when a user accesses the website, it also runs the authentication process (the sign in / sign up). Backend also maintains a mail server which sends important mails like password-reset mail in case user forgets his/her password.

1.1.3 The Frontend

The frontend is the unit is composed of all the files that are sent to the user, when he/she visits the website. These are the static files that contain all the html files, CSS files, JavaScript files and all the images used on the website. The frontend is built using ReactJS as the main framework.

The component is basically divided into two parts:

1. Search Part: In this the user can search past movies, their details and their revenues and can also search for the predicted revenue for upcoming movies predicted by others.

2. Prediction Part: In this part, the user can predict movie revenues based on the features and thus, the result of the prediction will be saved for the future use for others to view.

1.2 Need Analysis

There are multiple applications of a movie revenue prediction system. As the name suggests, it can be a powerful tool for movie producers, production houses and their investors in predicting what can be a fair estimate of the revenue the movie generates for a set of features such as an actor, a genre, a duration, release month and numerous other such variables. This will enable them to make better decisions about who to hire, when to release the film, who to market to, how much to spend on advertising etc. Additionally, it will also aid in the capital allocation for various projects ensuring that the most returns are generated for every last penny invested.

In a country like India, which has a colossal amount of single-screen theatres, the managers of these theatres have to often choose between which movie they screen and which they don’t. This tool can be boon to them in making an informed decision at best and an educated guess at worse about the films they offer at their establishments. Even actors and casting directors can benefit from this application by knowing what is trending and popular at any given point of time, thus improvising to maximize monetary gains. Lastly, the common people or the general population of a country can also be a big beneficiary of this product, in deciding which film will do better and which might not be the best choice for placing a wager on.

1.3 Problem Definition and Scope of the Project

The focus of this thesis is to formulate a method to preprocess the data set and evaluate which attributes are the most useful, by evaluating the correlation between the attributes and the success rate of the machine learning. Using this method the model can achieve a viable success rate when –trying to predict the rating and box office revenue.

The data set is going to be obtained on IMDB using a web scraper of our own creation. The data set will be limited to 10,000 unique movies by reason of making the workflow more manageable when processing the data. The machine learning algorithms that are going to be used are algorithms available in the classification learner module in MATLAB, mainly, SVM, Decision tree, KNN. [2]

1.4 Approved Objectives

- Predicting the revenue of movies and feature films with a reasonable accuracy.
- Economizing the project, cutting costs and producing maximum returns on investments.
- Helping theatres decide which movie to screen with what frequency to maximize profits.
- Providing patterns of popularity and profitability of particular genres, actors etc. for their Personal evaluation and betterment.
- Helping investors and commoners alike in deciding which will be a better project to bet Their stakes on.

1.5 Methodology Used

1.5.1 Access

IMDb makes its data publicly available for research purposes from which a local, offline database can be populated. FTP servers sponsored and maintained by IMDb contain stores of flat-format .list files that contain the same information found online through IMDb's web interface. For this project data access and preparation was facilitated by the help of two existing software systems: SQLite and imdbpy. SQLite is a widely-used SQL implementation supporting all standard SQL constructs and can be used to query all information found in the database in a high-level, declarative manner. imdbpy is a freely-available Python package designed for use with the IMDb database that implements various functions to search through and obtain data. Python scripts were developed to automatically pull the required feature data from the local SQLite database.

1.5.2 Pruning

The full database contains nearly 3 million titles, of which roughly 700,000 are feature films. Many of the titles found in the database contain incomplete information or are inappropriate for the scope of this investigation. Thus, in an attempt to both decrease training time and increase the accuracy of the prediction, the full title list was pruned using a series of SQL queries. The criteria by which IMDb titles were omitted are as follows: • Titles which are not movies (e.g. TV, videogames, etc.). • Adult films • Films missing budget data in US dollars • Films missing gross earnings data in US dollars • Films missing user rating data • Films not released in the United States After pruning the entire database of nearly 3 million entries, only 4260 titles remain (less than 0.002% of the original database). While this quantity is a tiny fraction of the overall database, the pruning constraints enforced are justifiable for the purposes of this prediction system; the pruned titles include films which are not released in major theater circuits, films we cannot generate labels for, and films not released in the US. Note that gross earnings reported in US dollars (our focus here) correspond to earnings from US theaters only, and therefore the financial-based metrics for film success are strictly an indicator domestic performance.

1.5.3 Features

Currently, the following features are drawn from each training film: • cast list • director(s) • producer(s) • composer(s) • cinematographer(s) • distributor(s) • genre(s) • MPAA rating • budget • release month • runtime

Several prediction models were implemented to learn from these features and make predictions on the success of films drawn from a test set. The prediction comes in the form of two primary success metrics.

Many of the movies listed on IMDb contain an average user-rating on a scale of 0 to 10 which corresponds to public opinion of that movie. The rating values exposed by the IMDb are rounded to a single decimal point; in this analysis we have rounded rating values to the nearest integer. Thus, the system predicts rating rounded to the nearest integer, turning a regression problem into a classification problem with 11 effective classes representing the integers 0-10 inclusive.

Gross earnings prediction.

Since the magnitude of gross earnings is not necessarily representative of movie success, especially if the movie had a large budget, the system predicts gross earnings of movies as a multiple of their budget, or bmult. We label the bmult of each of movie into the following bins: $[0 - 0.5)$, $[0.5 - 1)$, $[1 - 2)$, $[2 - \infty)$. As with rating predictions, this significantly simplifies the models by reducing the space of output classes. Note that bmult is a rough approximation of a film's of return on investment, where bmult greater than 1 corresponds to a profitable movie

1.6 Assumptions

1. We have assumed that we will keep getting data from IMDB and Rotten Tomatoes in the future also.
2. We are assuming that users will find premium features useful.
3. We are assuming that our site will not be hacked.
4. We are assuming that most producers predict the movie revenue for their upcoming films beforehand.
5. We are collecting data from scrapping which can only be done after some fixed intervals of time only.
6. We are fully dependent on internet.
7. Web browsers should be JavaScript enabled, otherwise website could not be accessed.

1.7 Summary of Project Outcomes

The final outcome of the project would be a website which can let its users to find the details about the movies as well as they can predict the revenue of the movies with attributes defined by them.

1.8 Project Schedule

Table 1 describe the project schedule

Table 1

15 Jan-19 Feb	Planning and Division of work
20 Feb-27 Feb	Analyzing key factors of the project
28 Feb-6 March	Learned about Beautiful Soup , Robobrowser and React JS
7 March-24March	Building a Scrapper and extracting data
25 March-5 April	Learned about regression and Support Vector Machine
6 April-13 April	Initiating construction of the project
14 April-20 April	Launching the first prototype
21 April -11 May	Building the models
12 May-18 May	Conceptualizing the front-end
28 May-29 May	Making a presentation and a report.

2.1 Background

This tool can be helpful for the new producers who wants to invest in film making and can help single screen theatre.

2.2 Existing System(s)/Related Works

There are multiple works related to this project, but none of which predict movie success before production begins. In 2002, a number of researchers evaluated whether or not a film's box office performance could be foreseen by estimating the probability of a film's revenue passing a certain threshold [6]. Through analysis of activity on Wikipedia pages, another algorithm was created to predict whether a film flops or becomes a blockbuster [7]. There are also multiple resources estimating lifetime gross of a film based on their success during opening weekend [8]. Google; however, has created an application that predicts box office revenue previous to opening weekend based on the search volume of the movie's trailer [9]. The main difference between these related works and ours, is that none of them predict movie success before the production begins. For example, many of these works incorporate movie reviews, from sources such as Rotten Tomatoes, as well as trailer views and movie hype - the extent to which the movie is publicized, promoted and discussed in the public. Although some of these attributes have been proven to influence a movie's success [10], research also indicates that movie reviews do not have a significant relationship with box office success [11]. Due to this research, we chose to develop our application without the use of reviews and were therefore able to focus on predicting a movie's success before it is even created. Additionally, many of these aforementioned works did not discuss how actors influence a movies success; therefore we decided to include actors in our project.

2.3 Problem Identified

Is it possible to classify the rating and box office revenue of a movie using metadata freely available on the web?

2.4 Methods and Tools Used

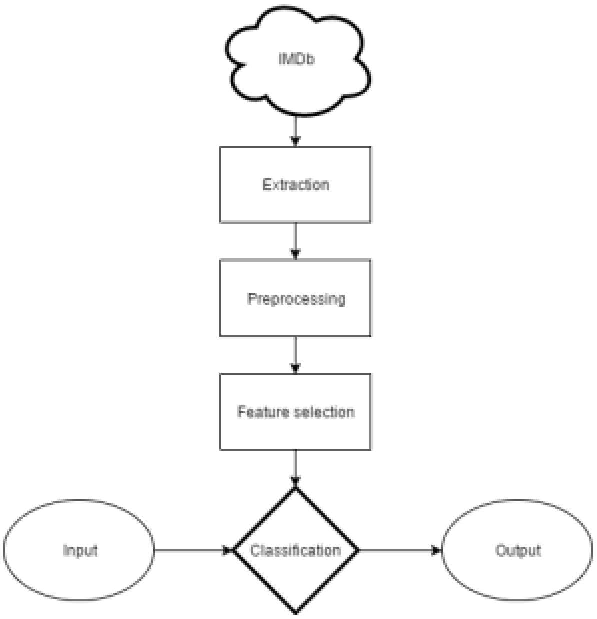
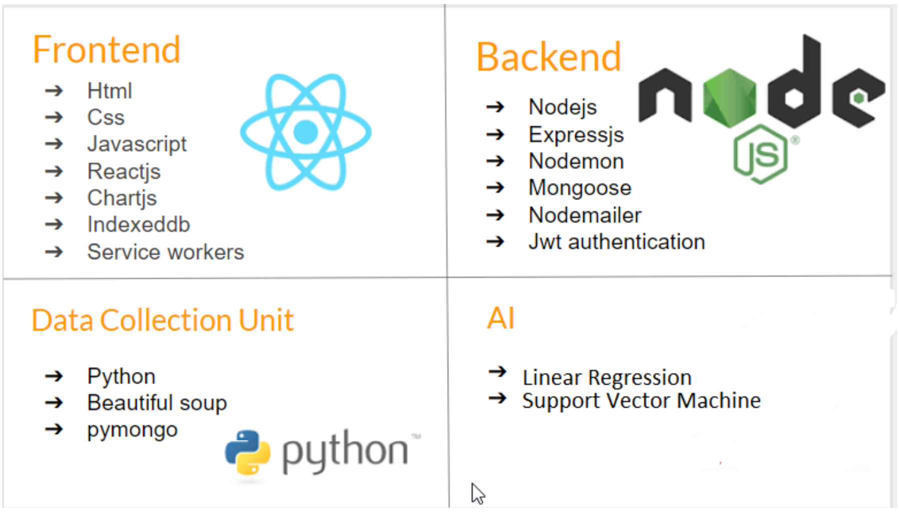


Figure 2 : Model Learning



Frontend

- Html
- Css
- Javascript
- Reactjs
- Chartjs
- Indexeddb
- Service workers

Backend

- Nodejs
- Expressjs
- Nodemon
- Mongoose
- Nodemailer
- Jwt authentication

Data Collection Unit

- Python
- Beautiful soup
- pymongo

AI

- Linear Regression
- Support Vector Machine

Figure 3: Tools used

REQUIREMENT ANALYSIS

3.1 Software Requirement Specifications

Table of Contents

3.1.1 INTRODUCTION

3.1.1.1 Document Purpose

The document is the one that describes the requirements along with interfaces for the system. It is meant for use by the developers and will be the basis for validating the final delivered system.

3.1.1.2 Movie Revenue Predictor SCOPE

The purpose of this project is to specify the requirements of the computer based software application, which is a font making system. This Software Requirements Specification provides a Complete description of all the functions and specifications of modules.

3.1.1.3 Indented Audience and Document Overview

The intended users are producer and normal people seeking for the prediction of movie. This document describes the font making system requirements in terms of systems requirements, Use Case, Data Flow and Activity diagrams.

3.1.1.4 Definitions and Abbreviations

- **UML (Unified Modelling Language)**

  UML is a language for specifying, constructing, visualizing and documenting the software system and its components with a set of rules and semantics.

- **Use-Case Diagrams**

  Use case diagram is a graph of actors, set of use cases enclosed by a system boundary, communication (participation) association between the actors and the use cases and a generalization among the use cases.

- **Actor**

  An actor represent a set of roles that user of a use case play when interacting with the use cases. Actor identified here is administrator and staff.

- **Use case**

  A use case is a description of a set of sequence of actions that a system performs to yield result of value to an actor.

- **Data Flow Diagram**

  A data flow diagram (DFD) is a graphical representation of the "flow" of data through an information system, modelling its process aspects. ADFD is often used as a preliminary step to create an overview of the system without going into great detail, which can later be elaborated.

- **Sequence Diagrams**

  Sequence diagrams are easy and intuitive way of describing the behavior of a system by viewing interaction between the system and its environment. A sequence diagram shows an interaction arranged in a time sequence.

- **Collaboration diagram**

  A collaboration diagram represents a collaboration, which is a set of objects related in a particular context and interaction, which is a set of messages exchanged among the objects within the collaboration to achieve a desired outcome.

- **Activity diagram**

  The activity diagram is used to describe the various activities taking place in an application.

3.1.1.4 DOCUMENT CONVENTIONS

In general this document follows the IEEE formatting requirements. Use Times New Roman font size 16, throughout the document for text. Use italics for comments. Document text should be single spaced.

3.1.1.5 REFERENCES

- www.stanford.edu
- www.google.com
- www.wikipedia.com

3.1.2 OVERALL DESCRIPTION

3.1.2.1 PRODUCT PERSPECTIVE

The website helps the user to search for the upcoming movies and details about the movie and can predict the movie revenue based on the attributes provided by the user.

3.1.2.2 PRODUCT FUNCTIONALITY

- System must able to get the information for the previously released movies.
- System must be able store all data received through scrapping in database.
- In case of any error, System must signify the client to refill and re-upload the template.

3.1.2.3 USER CHARACTERISTICS

The user does not require any high level knowledge for the use of the software.

3.1.2.4 OPERATING ENVIRONMENT

Software Requirements:

- Operating system: Ubuntu 16.04 / windows.
- Application software: Movie Revenue Prediction
- Hardware Requirements: A PC with proper speed and memory
- Design Tool: Star UML, Smart Draw, Lucid Chart.

3.1.2.5 DESIGN AND IMPLEMENTATION CONSTRAINT

3.1.2.5.1 Standard Development Tools

The following list presents the constraints, assumptions, dependencies or guidelines that are imposed upon implementation of the software:

- Response time for loading the Software should not exceed a minute.
- User should have the knowledge about the actors, directors, crew members etc. which are very easy to learn.

3.1.3. SPECIFIC REQUIREMENTS

3.1.3.1 External interface requirement

The following list presents the external interface requirements:

- The product require web-browser.
- Operating system requires a high quality screen resolution.

3.1.3.2 Detailed Description of Functional Requirements

Table shows a template that I’ll be using to describe functional requirements for users.

Purpose	This helps user to predict movie revenue.
Inputs	The user need to provide details of the movie for which you want.
Outputs	The generated results would be in form of tables and graphs .

USERS DOMAIN

1. Log-in: The User may login into the website after signing up for the website and hence later on store the data and prediction done by him earlier.

2. Home Page: The home-page will display upcoming movies and user can watch the trailer for the movies and can see the details about the movie.

3. Prediction: Users can predict the new movie revenue and can also predict the revenue which a movie having hypothetical attributes would generate.

3.1.4. NON FUNCTIONAL REQUIREMENTS

3.1.4.1 Performance Requirement

It should work precisely and quickly process the desired input filters

in order to enhance the working environment of the software.

3.1. 4.2 Data Storage

All the data received through web-scrapping should be stored and get updated constantly in database.

APPENDIX A - USE CASE DIAGRAM

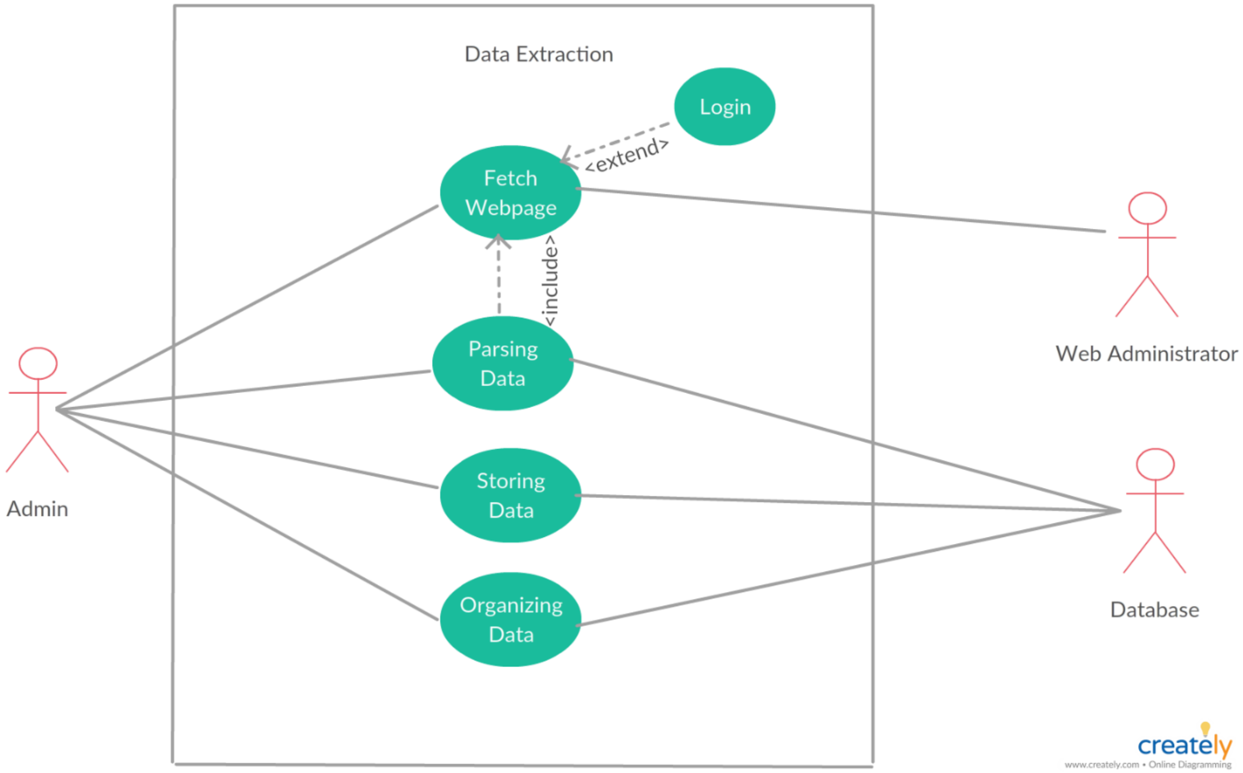


Figure 4: Use case diagram 1

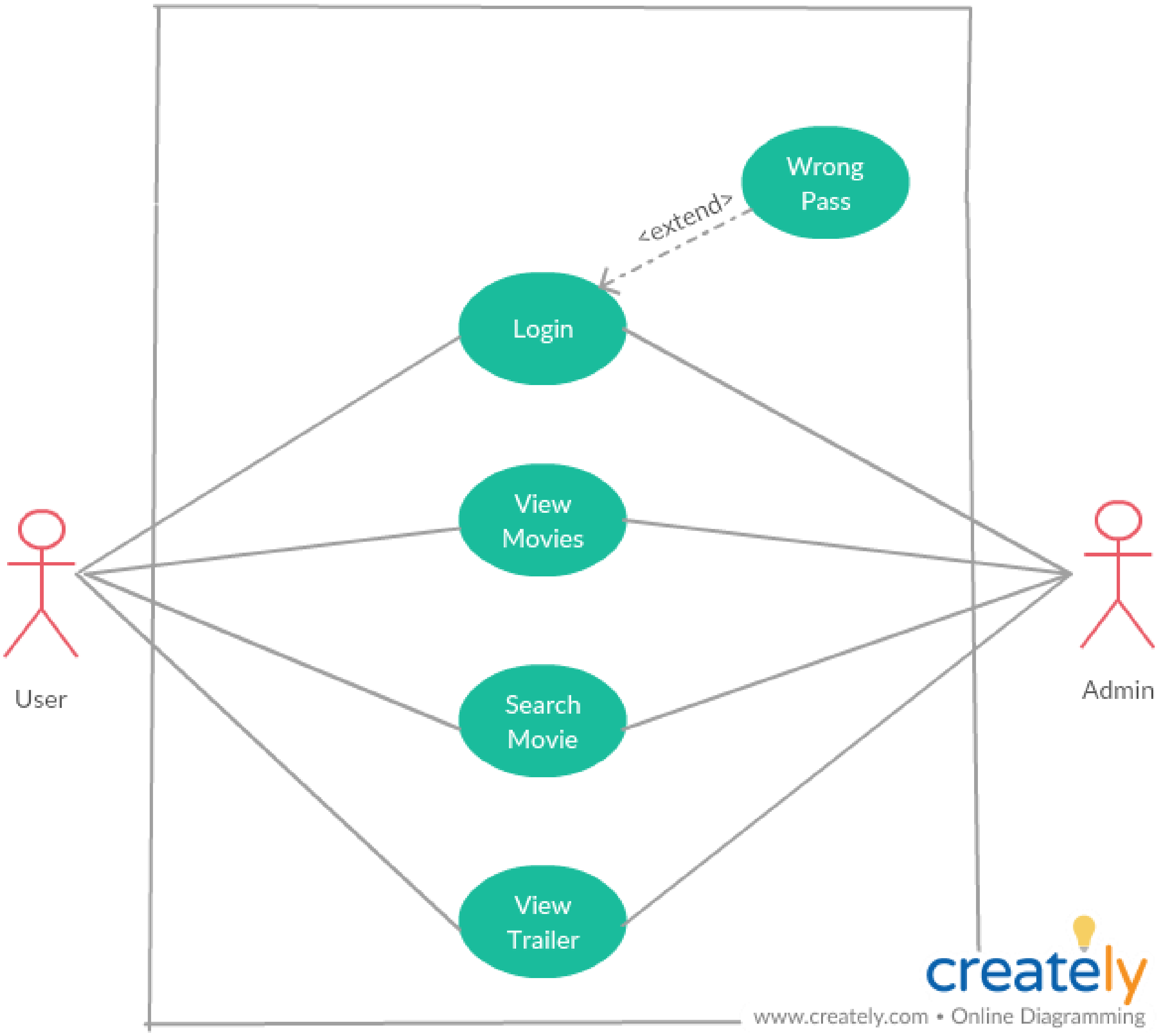


Figure 5: Use case diagram 2

.

APPENDIX B - ACTIVITY DIAGRAM

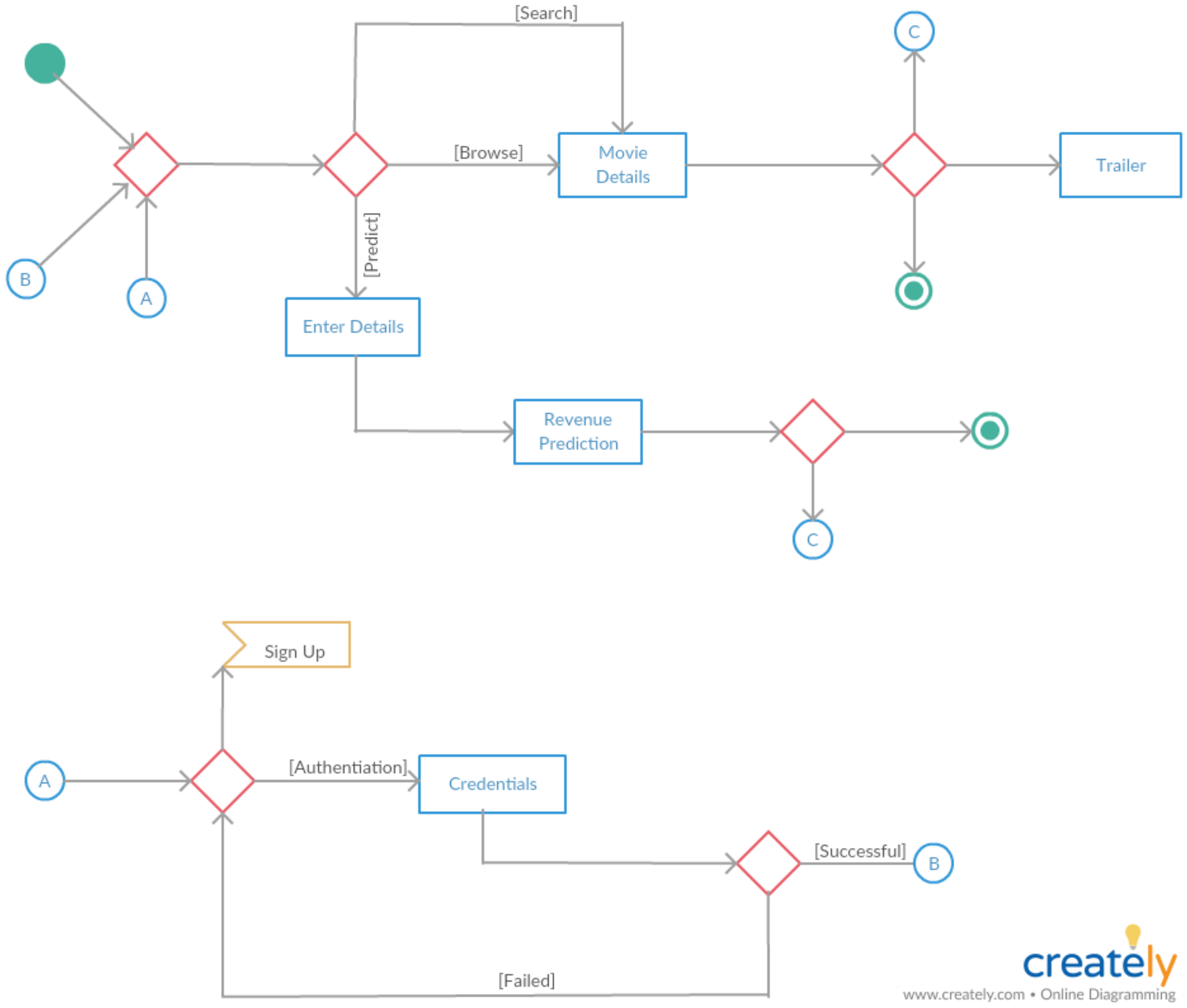


Figure 6: Activity Diagram

APPENDIX C - DATA FLOW DIAGRAM

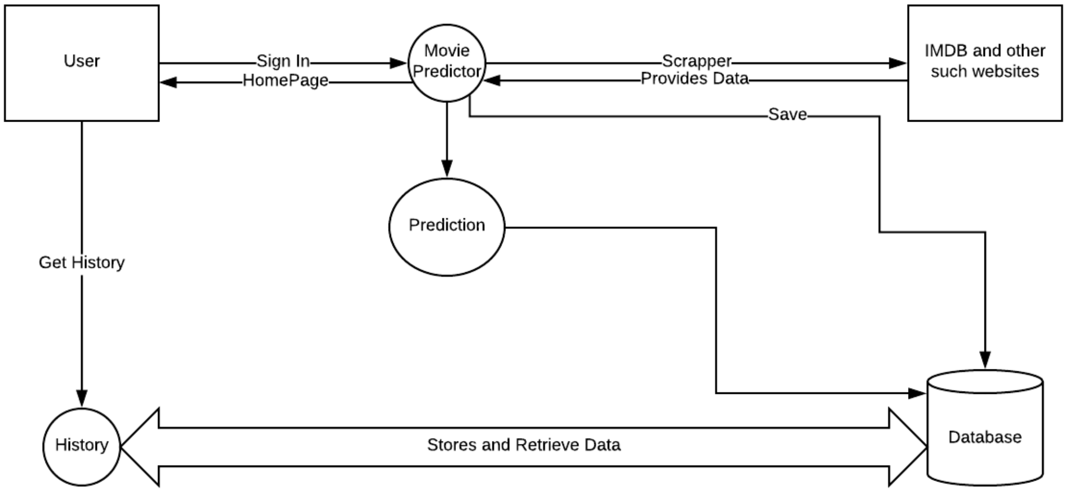


Figure 6: Data Flow Diagram 1

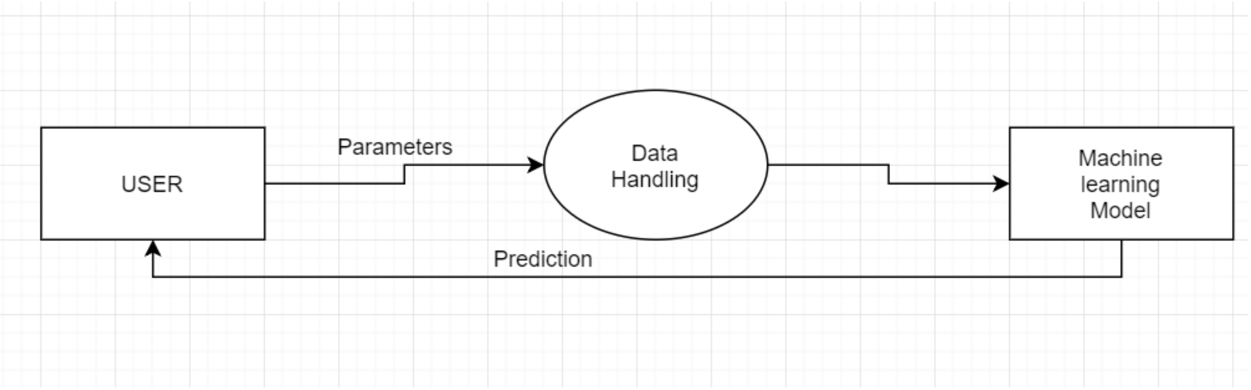


Figure 6: Data Flow Diagram 2

3.1.3.2 Cost Analysis

1. Domain+hosting : (1000+12*500)=INR 7000
2. Total Cost: INR 7000

3.1.3.3 Work Breakdown Structure

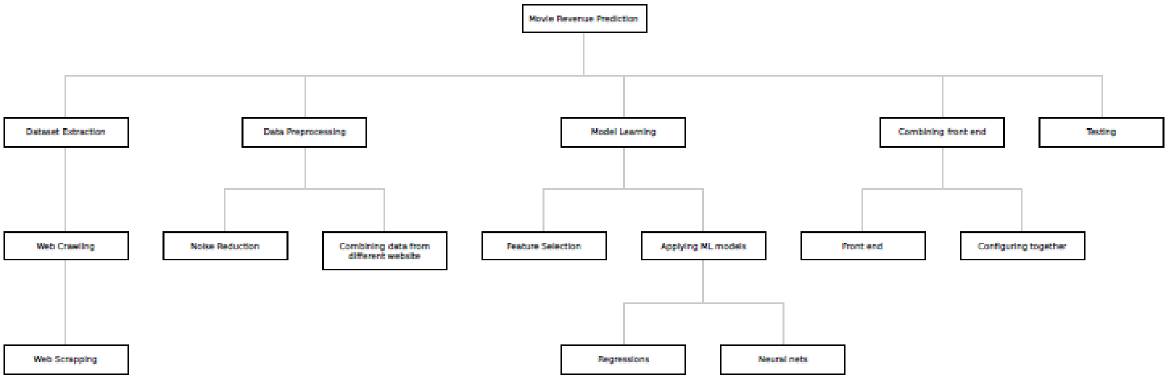


Figure 7: Work Flow

DESIGN SPECIFICATIONS

4.1 Flowchart of Proposed System

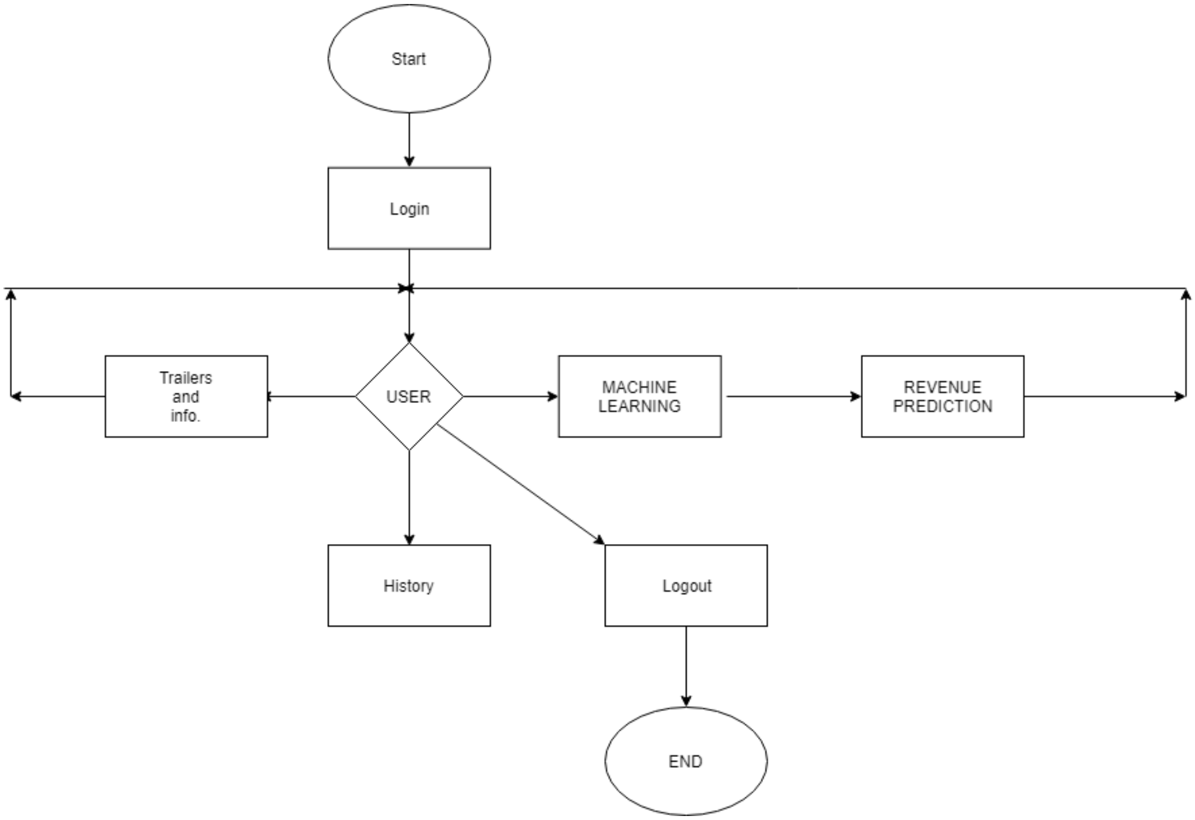


Figure 8: Proposed System Flowchart

4.2 System Architecture

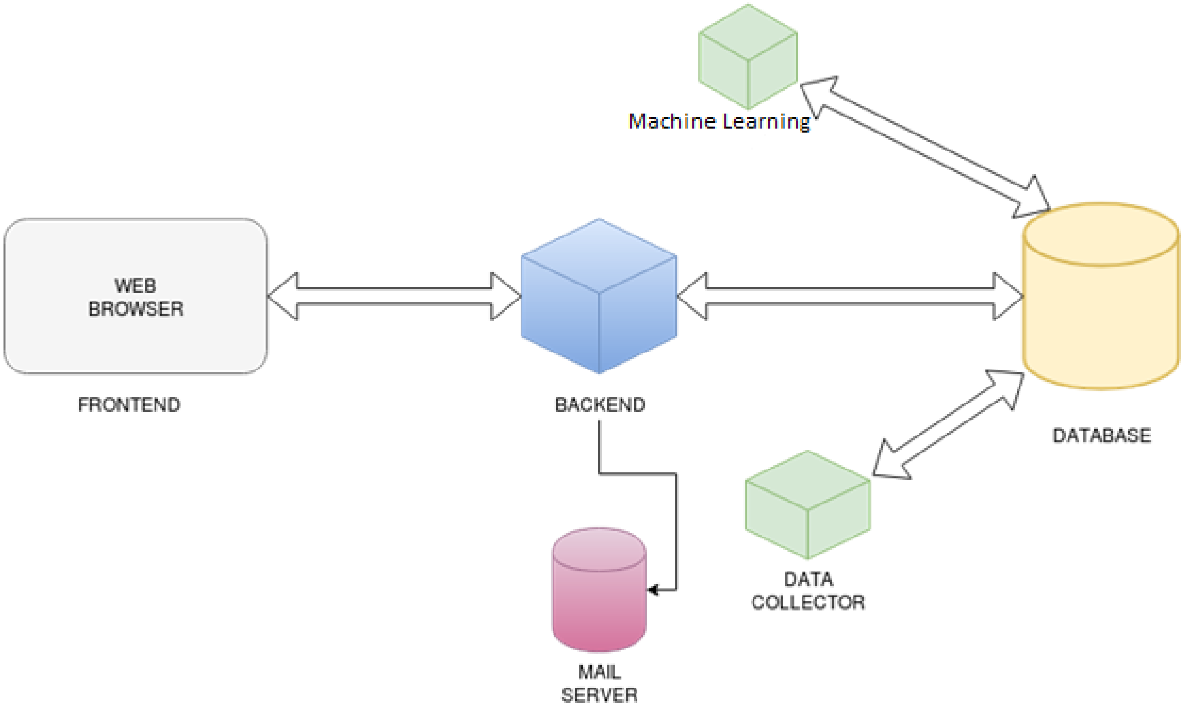


Figure 9: System Architecture

4.3 User Interface Diagram

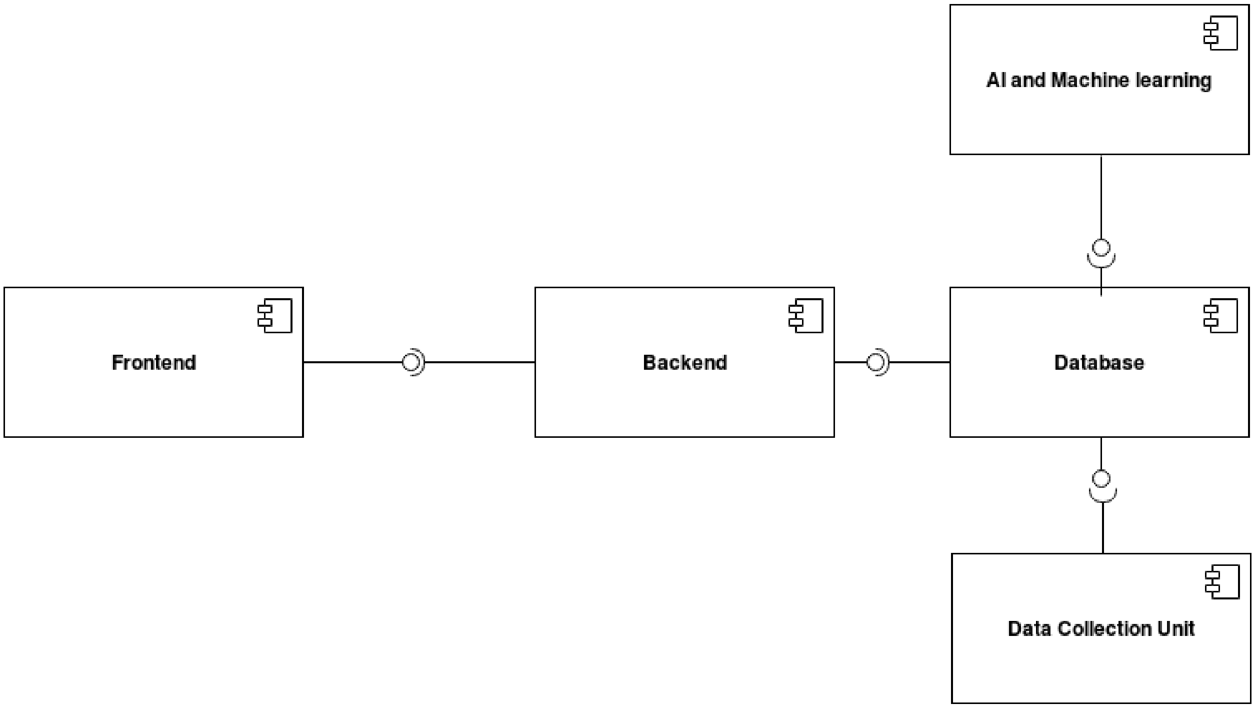


Figure 10: User Interface Diagram

4.4 System Components

- Frontend
- Backend
- Database
- Data collection Unit
- AI and Machine Learning

CONCLUSIONS AND FUTURE DIRECTIONS

5.1 Work Accomplished

All relevant data extracted from various sources such IMBD, Koi-Moi etc.

Data cleansed and pre-processed to take care of impurities, redundancies.

Model trained by using Support Vector Machines, linear regression.

5.2 Conclusions

This project when completed will work as follows – The user which will be an independent agency, producer or an investment firm will enter all related data about an upcoming movie, data will include things like the star cast, genre etc. After that, our algorithm will produce a number which will be a reasonably accurate of estimate of the earnings or revenue of the movie.

We expect that using our product, production houses will be able to optimize their budget for various things such as marketing, distribution pertaining to a profit margin which falls within their estimated revenue. This margin of safety will save them from any significant downside of their investments and will thus save money which can be put to other good use.

At the same time, single-screen theatres who want to maximize their profit will get an informed guess on screening which film will yield them the most profit. Thus, in the long run producing better returns and drawing more institutional as well as private investors and in turn, helping the industry grow.

5.3 Social and Economic Benefits

In 2017, more than 30 big budget films were declared flops and caused economic loss at the box office in India. Such flops not only cause great personal losses to the producers but also cause a negative feedback loop for investors in the industry. Our project, will greatly help in producing accurate results, churning profits for the producers in form of accurate revenue and profit for any given investment. Thus, delighting investors and helping the Indian film industry grow prosper and flourish.

5.4 Reflections

Contemporarily, there are many independent entities such as critics, channels and magazines who predict movie revenues without any intelligent algorithms or machine learning. Such actions, intended to garner better circulation for tabloids and magazines greatly hurt producers, distributors and investors who put their own money on the line and take on the risk. Our project is an attempt in the direction of improving the mechanism for movie revenue prediction with machine learning to yield dependable and credible results to boost investors' confidence in particular, and the film industry in general.

5.5 Future Work Plan

Extracting additional data from sources such as RottenTomatoes etc.

Finalizing the machine learning model to obtain the highest possible accuracy.

Designing the front-end and aesthetics.

Combining the model and the front-end.

REFERENCES

[1] Darin Im, Minh Thao, Dang Nguyen, Predicting Movie Success in the U.S. market, Dept.Elect.Eng, Stanford Univ., California, December,(2011)

[2] Jiawei Han, Micheline Kamber, Jian Pei, Data Mining Concepts and Techniques, 3rd ed.MA:Elsevier, (2011)

[3] Richard O. Duda, Peter E. Hart, David G. Stork, Pattern Classification, 2nd ed.NewYork: Wiley, (1973)

[4] Cohen, J., Cohen P., West, S.G., & Aiken, L.S. Applied multiple regression correlation analysis for the behavioral sciences. (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum Associates(2003).

[5] Christopher M. Bishop Pattern Recognition and Machine Learning, Springer, (2006).

[6] Cristianini, Nello; and Shawe-Taylor, John; An Introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, 2000. ISBN 0-521-78019-5.(2010)

[7] W. Zhang and S. Skiena, Improving movie gross prediction through news analysis, IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Milan, (2009)

[8] Sagar V. Mehta, Rose Marie Philip, Aju Talappillil Scaria, Predicting Movie Rating based on Text Reviews, Dept.Elect.Eng, Stanford Univ., California, December, (2011)