

Movie Rating Predictor

By: Arwin Neil Baichoo – 1600463

Software Programming (CSE 1040Y)

BSc(Hons) Software Engineering, University of Mauritius

Table of Content

[Table of Figures 1](#_Toc480188499)

[1. Introduction 2](#_Toc480188500)

[1.1 Problem Statement 2](#_Toc480188501)

[1.2 Aims and Objectives 2](#_Toc480188502)

[1.3 Scope 2](#_Toc480188503)

[2. Background Study 3](#_Toc480188504)

[2.1 Predictive Modelling 3](#_Toc480188505)

[2.2 Existing Movie Success Prediction Model 3](#_Toc480188506)

[2.3 Analysis of Existing Prediction Model 3](#_Toc480188507)

[2.4 Machine Learning 4](#_Toc480188508)

[2.5 Summary of Findings 5](#_Toc480188509)

[3. Analysis 6](#_Toc480188510)

[3.1 Proposed System 6](#_Toc480188511)

[3.2 Functional Requirements 6](#_Toc480188512)

[3.3 Non-Functional Requirements 6](#_Toc480188513)

[4. Design 7](#_Toc480188514)

[4.1 Program Flowchart 7](#_Toc480188515)

[4.2 Prediction Model Algorithm 8](#_Toc480188516)

[4.2.1 Simple Linear Regression 8](#_Toc480188517)

[4.2.2 Averaging 11](#_Toc480188518)

[4.3 Prediction Model Flowchart 11](#_Toc480188519)

[4.4 Dataset File - Comma-separated Values 12](#_Toc480188520)

[5. Implementation 13](#_Toc480188521)

[5.1 Movie Dataset 13](#_Toc480188522)

[5.2 Code Organization: Headers and Functions. 14](#_Toc480188523)

[6. Testing 16](#_Toc480188524)

[6.1 Test Plan 16](#_Toc480188525)

[6.2 Sample Screenshots 17](#_Toc480188526)

[7. Conclusion 20](#_Toc480188527)

[7.1 Achievements 20](#_Toc480188528)

[7.2 Problems 20](#_Toc480188529)

[7.3 Future work 20](#_Toc480188530)

[8. User Manual 21](#_Toc480188531)

[8.1 Start Up 21](#_Toc480188532)

[8.2 Entering path to dataset 21](#_Toc480188533)

[8.3 The Main Menu 21](#_Toc480188534)

[8.3.1 Predict Output 22](#_Toc480188535)

[8.3.2 Update Dataset 22](#_Toc480188536)

[8.3.3 Use Another Dataset 22](#_Toc480188537)

[8.3.4 Exit 22](#_Toc480188538)

Table of Figures

[Figure 1 - Flowchart 7](file:///C:\Users\Admin\Assignment\MovieRatingPredictor\Movie%20Rating%20Predictor%20-%20Report.docx#_Toc480187471)

[Figure 2 - Example of a Regression line (red) with scattered points on a graph 8](file:///C:\Users\Admin\Assignment\MovieRatingPredictor\Movie%20Rating%20Predictor%20-%20Report.docx#_Toc480187472)

[Figure 3 Prediction Model Flowchart 11](file:///C:\Users\Admin\Assignment\MovieRatingPredictor\Movie%20Rating%20Predictor%20-%20Report.docx#_Toc480187473)

[Figure 4 Cleaned dataset visualized as table 13](file:///C:\Users\Admin\Assignment\MovieRatingPredictor\Movie%20Rating%20Predictor%20-%20Report.docx#_Toc480187474)

[Figure 5 - Software Start Screen 17](file:///C:\Users\Admin\Assignment\MovieRatingPredictor\Movie%20Rating%20Predictor%20-%20Report.docx#_Toc480187475)

[Figure 6 - Results for scenarios 1,2 and 3 17](file:///C:\Users\Admin\Assignment\MovieRatingPredictor\Movie%20Rating%20Predictor%20-%20Report.docx#_Toc480187476)

[Figure 7 -Scenario 11 18](#_Toc480187477)

[Figure 8 - Results for scenario 4 & 8 18](file:///C:\Users\Admin\Assignment\MovieRatingPredictor\Movie%20Rating%20Predictor%20-%20Report.docx#_Toc480187478)

[Figure 9 - Scenario 12 19](file:///C:\Users\Admin\Assignment\MovieRatingPredictor\Movie%20Rating%20Predictor%20-%20Report.docx#_Toc480187479)

# Introduction

With the advent of data science, the world has changed in ways previously unconceivable. A lot of things which used to be impossible to analyze and predict even a decade ago, have now become easier and more intuitive to predict, thanks to the numerous advantages of data science. The entertainment industry is no different. Over the years, data science and its immense power has been the root cause of a paradigm shift in the way operations are performed in the entertainment business.

## 1.1 Problem Statement

The motion picture industry is innately rich in data. Earlier, the movie industry used to employ the knowledge of certain industry trends, basic rule of thumb approaches and traditional wisdom and intuition to predict the success or failure of particular movies. This method was never very accurate or reliable, and has been found wanting in many areas over the years.

In the present day, with the advent of data science and the exciting opportunities that are promised through data mining and analysis, the industry is actively in the process of formulating a new, improved and reliable method of accurately predicting the success and failure of particular movies. Stakeholders in most major move industries of the world are turning towards data science and analytics with the intention of increasing their success rate.

## 1.2 Aims and Objectives

**Aim**:

The aim of this is to create a data analysis & prediction software that will give user an insight on what the rating of a movie might be

**Objectives:**  
1. Provide a software which will help user in decision-making prior to the release of the movie.

2. Design a software which will be able to perform data mining from a big amount of movie data and generating a prediction from it.

## 1.3 Scope

The scope of this project is developing a software which will be able to process relatively large amount of dataset to produce a prediction. The dataset will be drawn from online movie database websites.

# Background Study

## Predictive Modelling

Predictive modeling is a process that uses data to forecast outcomes. Each model is made up of a number of variables, which are variables that are likely to influence future outcome. Once data has been collected for relevant variables, a model is formulated. The model may employ a simple linear equation or it may be a complex neural network, and will process the data, identifying patterns and relations that will help to produce new values, using knowledge extracted from that data.

## 2.2 Existing Movie Success Prediction Model

Currently, many major movie studios of the world are implementing data analytics as a leading means to measure the possibility of success of many of their projects. The primary factor that helps decide the possibility of success for a particular movie is the knowledge about what draws interest of people. This knowledge can be achieved with a degree of accuracy through the analysis of various different online sources including search engine results, video views and comments, ratings on expert websites and social media content. Analyses of past success records of other movies of the same genre or those starring the same people can also be brought into the fold to deliver accurate results. The primary goal is to be able to accurately forecast the possible box office returns, feedbacks and rating for a particular movie using the relevant kinds of data.

## 2.3 Analysis of Existing Prediction Model

In order to conduct an analysis, it is expected to have a large store of information, including the past records of movies, Production Company, other movies of the same genre, similar casts, types of story and different avenues of marketing and promotion. Apart from these important factors, there are other influential factors like engagement related to please and trailer launches, social media buzz and public forum comments.

## 2.4 Machine Learning

Machine learning can be described as computing systems that improve with experience. It can also be described as a method of turning data into software. It is a method of creating software models that are trained from huge volumes of data and then used to predict certain patterns, trends, and outcomes, based on the supplied input data. The primary benefit of this approach is that the resulting program that is developed has been trained via massive quantities of learning data and finely tuned capable of predicting the likelihood of an output based on the provided data.

**Machine Learning Algorithms**

**Supervised learning** is a type of learning algorithm that uses known datasets to create a model that can then make predictions. The known data sets are called training data sets and include input data elements along with known response values. From these training datasets, supervised learning algorithms attempt to build a new model that can make predictions based on new input values along with known outcomes. Some algorithms:

* **Classification**, used to predict one or more discrete variables, based on the other attributes in the dataset.
* **Anomaly** detection is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset. It is what usually powers spam detention algorithms. It is also used in banks to detect possible frauds based on your transaction patterns. Detecting buying things at an unusual place or purchasing something unusually pricy.
* **Regression** used to predict one or more continuous variables, such as profit or loss, based on other attributes in the dataset. Its used explicitly when there is the need to predict a numerical value, like future stock market price, or what will be the peak temperature tomorrow.

**Unsupervised learning** is another type of learning algorithm. The success of the new predictive model depends entirely on the ability of the machine to correctly infer and identify patterns, structures, and relationships in the incoming data set. An example:

* **Clustering** deals with finding a structure in a collection of unlabeled data, situations where we want the machine to conduct its own analysis on the dataset, determine relationships, infer logical groupings, and generally attempt to make sense of chaos.

## 2.5 Summary of Findings

Basically, the process of data analytics when it comes to the movies is concerned with processing large volumes of past data from various possible sources, identifying certain set patterns through learning algorithms, comparing them with existing data, available data points and using the knowledge thus gained to make better decisions for enhanced performance.

The predictive model that will be built is likely to use regression learning algorithms as various variables that will be dealt with are likely to be numerical, such as budget of a movie, or the rating itself that will be predicted.

# Analysis

## Proposed System

The system will be a software that will take in multiple variables of a movie, use a dataset containing a large amount of records of past movies and their respective variables, process those with a predictive model to finally generate a prediction of the rating of the input movie.

## Functional Requirements

3.2.1 Dataset

3.2.1.1 The system shall allow the user to select a movie dataset file.

3.2.1.3 The system shall allow the user to append new data to the dataset file.

3.2.2 Menu

3.2.2.1 The system shall have a menu displaying all functionalities of the software.

3.2.2.2 The menu shall allow user to exit the software.

3.2.3 Prediction

3.2.3.1 The system shall allow user to input information about a new movie.

3.2.3.2 The system shall predict movie rating for the movie information input.

3.2.3.3 The system shall predict movie rating on a scale of 10.

3.2.3.4 The system shall predict movie rating from movie’s budget, genre, age rating and duration.

## 3.3 Non-Functional Requirements

3.3.1 Dataset

3.2.1.2 The dataset should be of a format that can also be processed by text editors.

3.3.2 Prediction

3.3.2.1 The system shall give a detailed report of how rating was obtained.

3.3.3 Error Handling

3.3.3.1 The system shall give error messages if exception or errors in input arise.

# Design

## C:\Users\Admin\Downloads\Movie Prediction - Page 1.pngProgram Flowchart

Figure 1 - Flowchart

## 4.2 Prediction Model Algorithm

The movie variables from which rating will be predicted are:

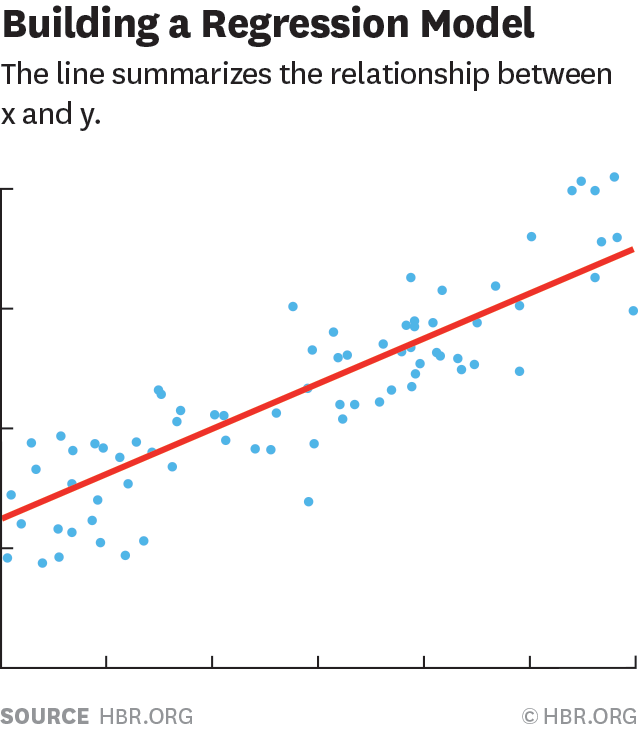
* Duration (numeric)
* Budget (numeric)
* Age Rating(alphanumeric)
* Genre(alphanumeric)

Since program will be dealing with different data types, the relation between each of the variables and rating will be found.

Four different ratings based on each one of the above variables, using the relation obtained. Those rating will then we averaged to obtain the final predicted rating.

### 4.2.1 Simple Linear Regression

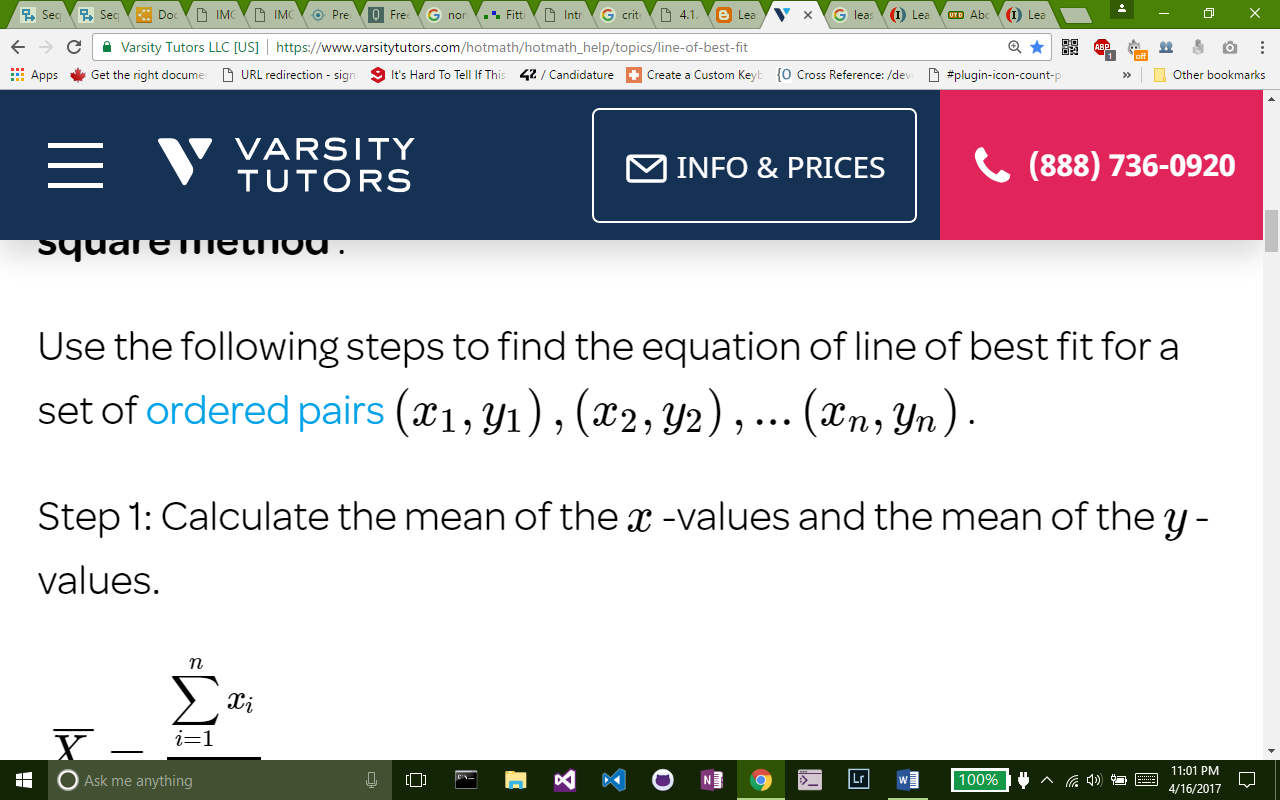
Linear regression consists of finding the best-fitting straight line through scattered points on a graph. The best-fitting line is called a regression line. This regression line in the case of the software will represent the relationship between our criterion variable, the rating, and another predictor variable. This relationship can then be used to determine any rating value for a given predictor value along the regression line.

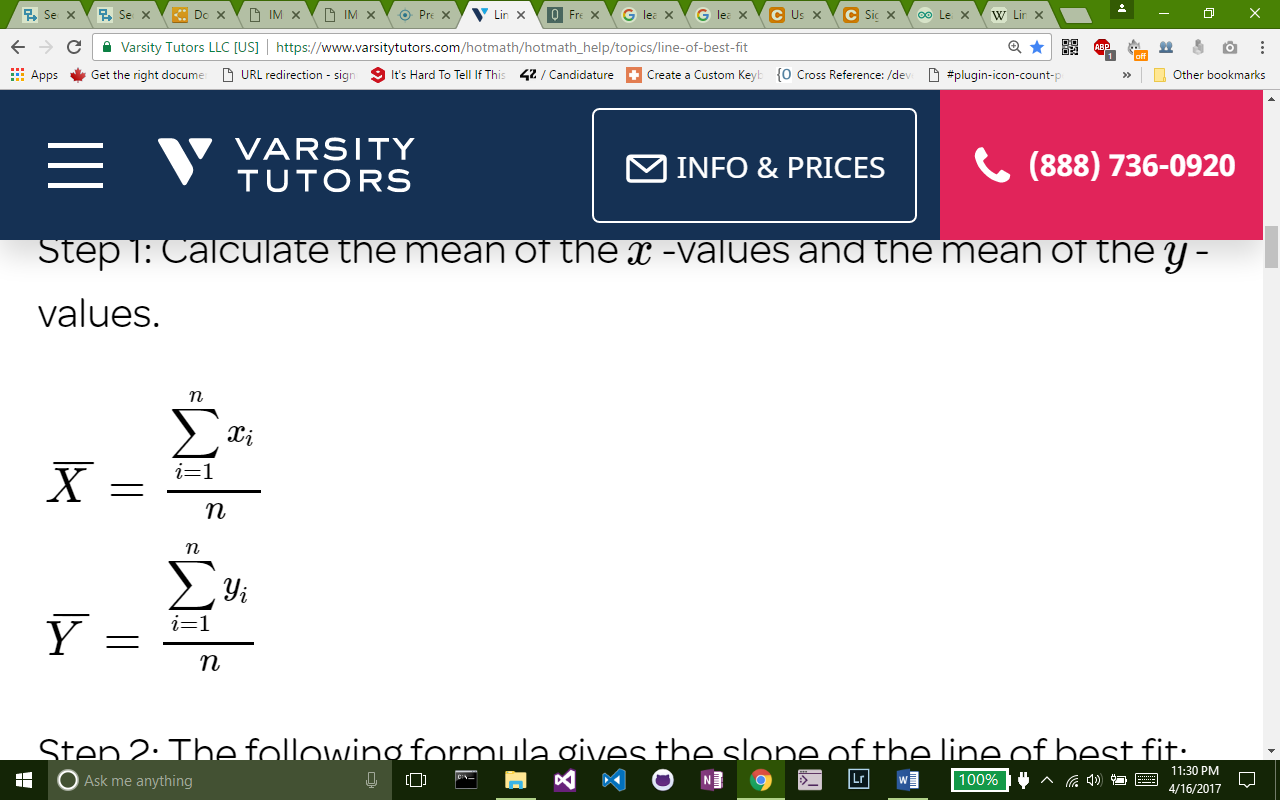
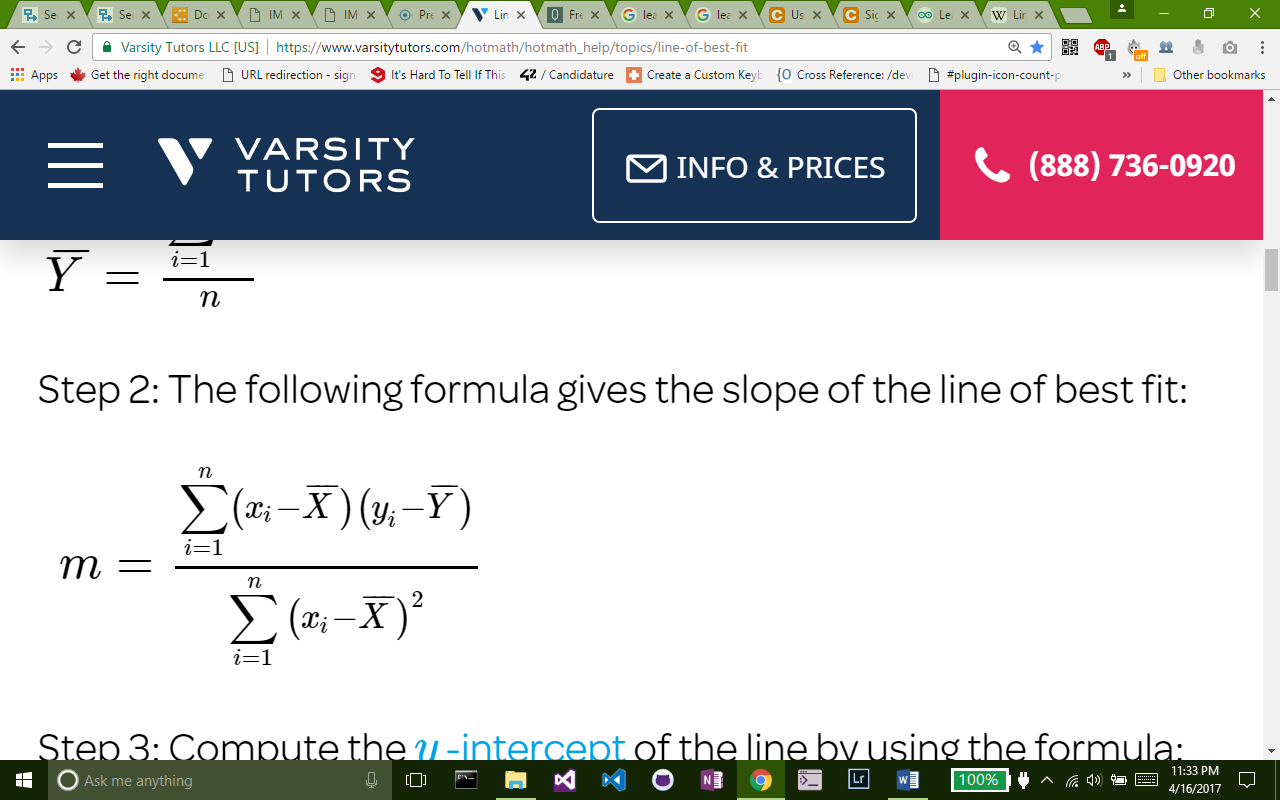
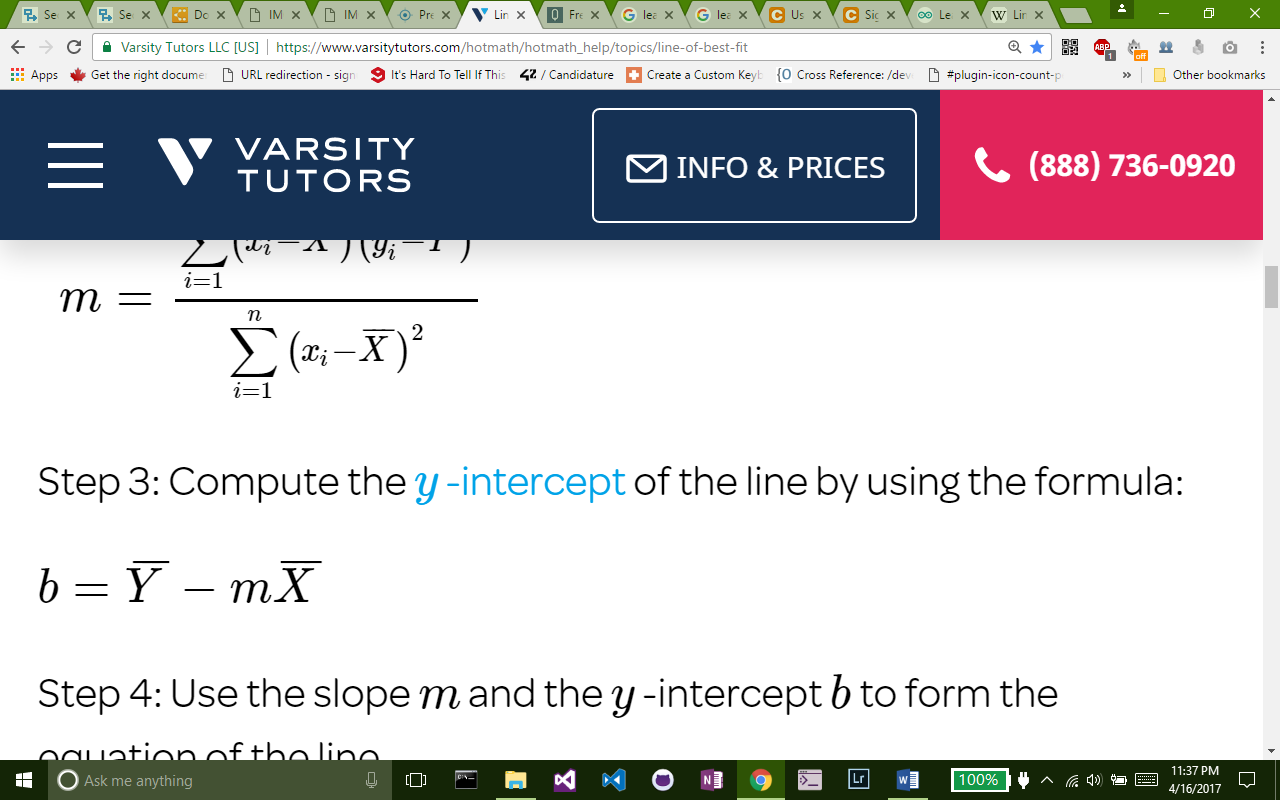


**Least-Square Linear Regression**

Figure 2 - Example of a Regression line (red) with scattered points on a graph

The least squares method is one of the most popular techniques to determine a line of best fit, by minimizing the sum of squares created by a mathematical function. A "square" is determined by squaring the distance between a data point and the regression line.

**Mathematical steps** to find the equation of line of best fit for a set of ordered pairs

1. Calculate the mean of the 𝒙 -values and the mean of the y -values.
2. Use following formula to determine slope m
3. Calculate the y –intercept b of the line by using
4. Use the slope m and the y -intercept b to form the equation of the line.

There exists a variant of the least squares method for coding.

**Pseudocode** to find the equation of line of best fit for a set of data points:

Start

for each data point (x, y)

sum\_x += x //sum of x values

sum\_y += y //sum of y values

sum\_x2 += x \* x //sum of x2

sum\_xy = sum\_xy + x \* y //sum of y \* x

n++ //number of points

end for

xm=sum\_x/n // mean x

ym=sum\_y/n // mean y

m = (n \* sum\_xy - sum\_x \* sum\_y) / (n \* sum\_x2 - sum\_x \* sum\_x)

b = ym-m\*xm

End

**Application**

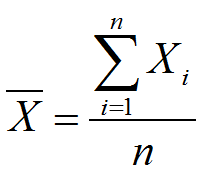
The relationship will be obtained in terms of a line equation y=mx+b, where x is the predictor variable.

The rating based on that predictor variable is given by replacing the value of the predictor variable into the equation.

This method will be used to predict ratings based on *budget* and *duration* of movie.

### Averaging

An average or arithmetic mean, is a number expressing the central or typical value in a set of data, which is calculated by dividing the sum of the values in the set by their number.



**Application**

Averaging will be used to determine a representative rating based on movies of a certain *genre* or particular movie *age-rating*, as well to determine final rating.

## C:\Users\Admin\Downloads\Blank Diagram - Page 1 (1).png4.3 Prediction Model Flowchart

Figure 3 Prediction Model Flowchart

## 4.4 Dataset File - Comma-separated Values

The comma-separated values (CSV) format is a widely used text file format often used to exchange data between applications and to store datasets as record consists of one or more fields, separated by commas.

CSV record structure sample: movie\_title, genres,duration, age\_rating,budget,rating

Interpretation:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| movie\_title | genres | duration | age\_rating | budget | rating |
|  |  |  |  |  |  |

# Implementation

## Movie Dataset

The movie dataset that will be used throughout this project was obtained GitHub repository **sundeepblue/movie\_rating\_prediction,** and contains about 5000 movies with 28 variables, all scraped from The Internet Movie Database (abbreviated **IMDb**).

In order to work with the data, this dataset was cleared:

* Redundant fields were removed.
* Rows with missing data were removed.

After cleaning, the amount of workable data within the dataset summed up to about 4300 records, with 6 fields.

Figure 4 Cleaned dataset visualized as table

**Notes on fields:**

* movie\_title : *The Movie title, alphanumeric*
* genres : *Multiple genres for a movie, separated by a “|”, alphanumeric*
* duration – *Movie duration in minutes, integer*
* content\_rating – *Content rating of the movie, alphanumeric*
* budget – *budget for the movie in dollars- integer*
* imdb\_score – *rating for the movie, decimal*

## Code Organization: Headers and Functions.

The code is organized into 5 main files:

**display.h**

This header contains functions related to displaying general information or massages in the console.

|  |  |
| --- | --- |
| Void Function | Purpose |
| displaySplash () | Displays a splash screen when the program launches |
| displayMenu() | Displays the software’s main menu |
| displayMovie(movie new\_movie) | Takes in an argument of type movie and displays it. Mainly for debugging purposes. |

**structures.h**

This header contains all data structure definitions.

|  |  |
| --- | --- |
| Structure | Purpose |
| movie | Stores title, duration, genres, age rating, budget and rating of a movie. |
| point | Represents a point with an x and a y coordinate. |
| equation | Represents mathematical equation of a line and stores a gradient and a y intercept value. |

**operation.h**

This header is the core of the software and contains all function that perfom some kind processing.

|  |  |
| --- | --- |
| Function | Purpose |
| string getDataPath() | Gets path of dataset from user ,performs some validation and returns it. |
| string getMenuInput() | Gets the user’s choice in the main menu, and returns it. |
| movie getMovieDetails() | Gets user input for a new movie, and returns the movie information in a variable of type movie. |
| movie toMovie(string line) | Converts a CSV movie record of type string to an object of type movie and returns it. |
| equation computeL2(string dataset,string independent\_var) | Computes the equation of line of best fit of an independent variable against rating using the Least Square Regression Method and returns it as an object of type equation. |
| float computeAvearge(string dataset,string var,string val) | Computes average of a given variable and returns it.  var specifies which of age\_rating or genre we are predicting  val specifies which genre or age rating should be considered while going through the dataset |
| void predictRating(string dataset) | Computes final rating prediction using other functions ,and outputs a report on the console. |
| void addNewData(string path) | Uses getMovieDetails()appends new data to dataset. |

**main.cpp**

The main() function calls the other function, as well as pass the path to dataset to them as needed, and terminates the program when user decides to do so.

**splash.txt**

Stores splash screen used by displaySplash ()

# Testing

## Test Plan

## Sample Screenshots

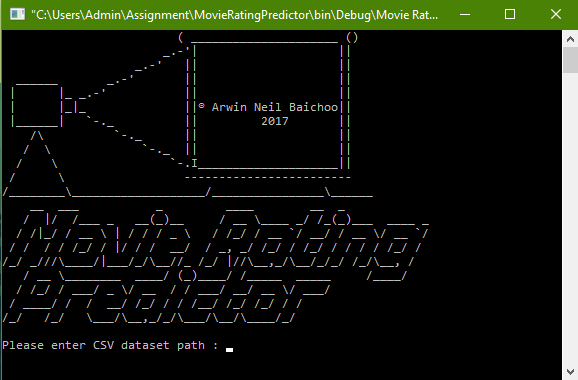


Figure 5 - Software Start Screen

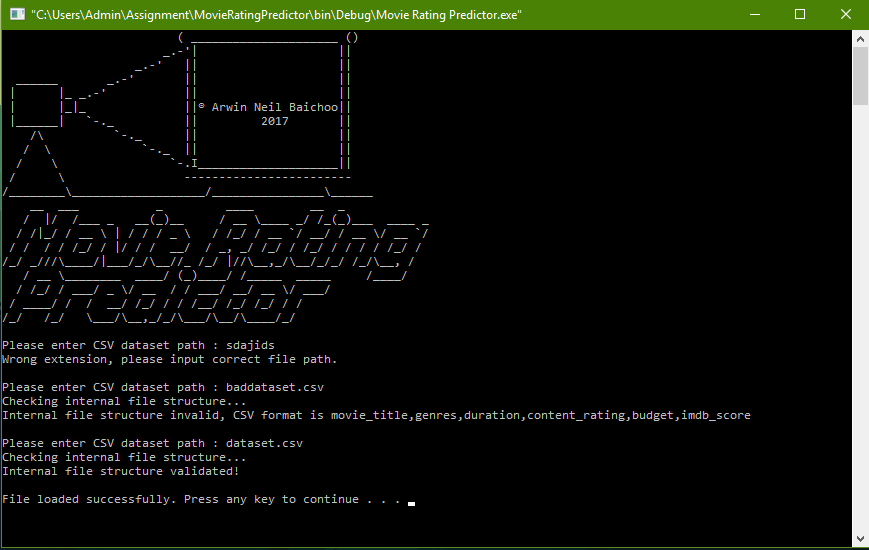
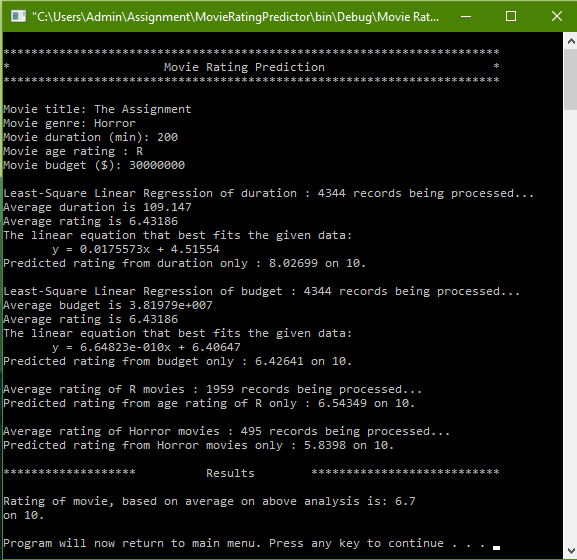


Figure 6 - Results for scenarios 1,2 and 3



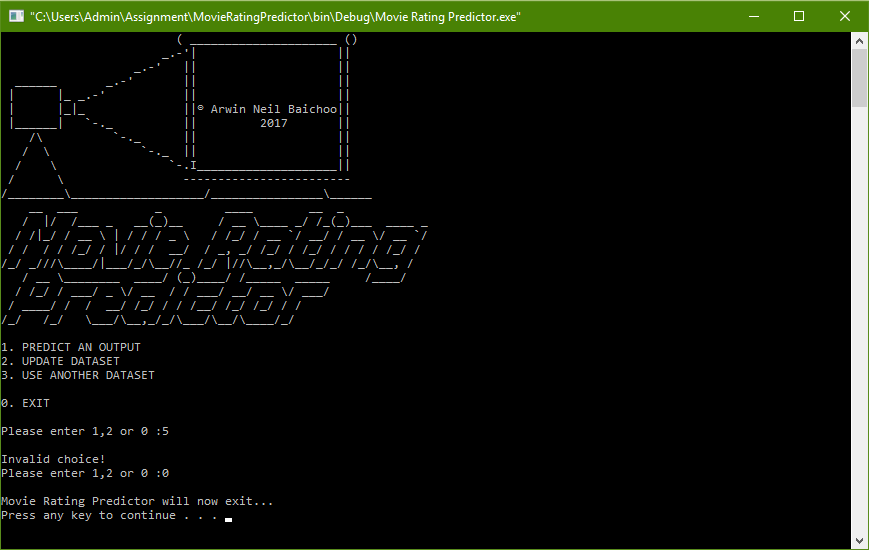
Figure 7 -Scenario 11

Figure 8 - Results for scenario 4 & 8

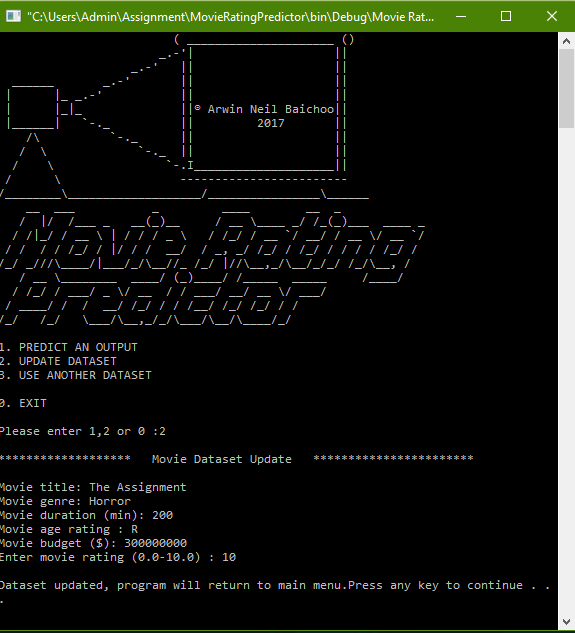


Figure 9 - Scenario 12

# Conclusion

## Achievements

Through this assignment I;

* Gained a lot of knowledge in the field of data science
* Learned the Least Square Regression algorithm.
* Learned how knowledge can be easily extracted from information available on the internet
* Learned C style printf() number formatting when dealing with decimals.
* Got me some experience with time management with regards to this project
* Got a good experience on working with a white card by myself on a project.
* Realized the importance of working in a team.

## Problems

* Research was time consuming, especially for finding the appropriate regression algorithms.
* CodeBlocks, the IDE used for this projects have a couple of inbuilt C++ functions missing, notably stoi (), which should have been an essential part for this project. Luckily I was able to use some alternatives in the cstdlib in the form of strtof().

## Future work

In the future, I would like to improve the program by:

* getting a larger dataset to work with.
* learning more about alternative algorithms to average that might produce results with more accuracy
* improving the predictive model by using more variables that also have an impact on the rating of a movie.

# User Manual

## Start Up

When the software starts, a splash screen will appear followed by a prompt to enter path to the dataset.

## Entering path to dataset

When entering the path to the dataset, the following guidelines should be followed:

* The path to dataset should be a valid one
* The dataset in the Comma-separated values format
* The CSV header for dataset to validated is as follows:

movie\_title,genres,duration,content\_rating,budget,imdb\_score

Once the dataset is validated, the software will display the main menu.

## The Main Menu

The main menu will present the following options:

* 1. Predict Output
* 2. Update Dataset
* 3. Use another Dataset
* 0. Exit

Keying in one of the numbers will open the corresponding functionality. Else an error will be displayed.

## Predict Output

If at **8.3** you keyed in 1, the Movie Prediction screen will be displayed. To predict a movie rating, type in information about the movie as follows:

* Movie Title
* Genre
* Duration of movies in minutes
* Age rating of the movie
* Budget of the movie in dollars

Once all information entered, you will be presented with a detailed log of the processing and a report of the ratings predicted.

The software then returns to the main menu.

## Update Dataset

If at **8.3** you keyed in 2, the Update Dataset screen will be displayed.to ad a movie record to the dataset, type in information about the movie as follows:

* Movie Title
* Genre
* Duration of movies in minutes
* Age rating of the movie
* Budget of the movie in dollars
* Rating of the movie on a scale of 10.0

Once all information entered, a confirmation message will be displayed.

The software then returns to the main menu.

## Use Another Dataset

If at **8.3** you keyed in 2, the a prompt to enter path to dataset will be displayed. Refer to section **8.2** for guidelines on entering path to your dataset.

## Exit

If at **8.3** you keyed in 0, an exit message will be displayed, after which software terminates.