

Predicting Movie Success with Data Mining

Interim Report

DT228

BSc in Computer Science

**Povilas Kubilius**

**C16370803**

**Leo Tilson**

School of Computer Science

Technological University, Dublin

**04/12/2019**

Abstract

Declaration

I hereby declare that the work described in this dissertation is, except where otherwise stated, entirely my own work and has not been submitted as an exercise for a degree at this or any other university.

Signed:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Povilas Kubilius

Date

Acknowledgements

Body text

Table of Contents

[1. Introduction 7](#_Toc21978017)

[1.1. Project Background 7](#_Toc21978018)

[1.2. Project Description 8](#_Toc21978019)

[1.3. Project Aims and Objectives 8](#_Toc21978020)

[1.4. Project Scope 8](#_Toc21978021)

[1.5. Thesis Roadmap 8](#_Toc21978022)

[2. Literature Review 9](#_Toc21978023)

[2.1. Introduction 9](#_Toc21978024)

[2.2. Alternative Existing Solutions to Your Problem 9](#_Toc21978025)

[2.3. Technologies you’ve researched 9](#_Toc21978026)

[2.4. Other Research you’ve done 9](#_Toc21978027)

[2.5. Existing Final Year Projects 9](#_Toc21978028)

[2.6. Conclusions 9](#_Toc21978029)

[3. Prototype Design 10](#_Toc21978030)

[3.1 Introduction 10](#_Toc21978031)

[3.2. Software Methodology 10](#_Toc21978032)

[3.3. Overview of System 10](#_Toc21978033)

[3.4. Front-End 10](#_Toc21978034)

[3.5. Middle-Tier 10](#_Toc21978035)

[3.6. Back-End 10](#_Toc21978036)

[3.7. Conclusions 10](#_Toc21978037)

[4. Prototype Development 11](#_Toc21978038)

[4.1. Introduction 11](#_Toc21978039)

[4.2. Prototype Development 11](#_Toc21978040)

[4.3. Front-End 11](#_Toc21978041)

[4.4. Middle-Tier 11](#_Toc21978042)

[4.5. Back-End 11](#_Toc21978043)

[4.6. Conclusions 11](#_Toc21978044)

[5. Testing and Evaluation 12](#_Toc21978045)

[5.1. Introduction 12](#_Toc21978046)

[5.2. Plan for Testing 12](#_Toc21978047)

[5.3. Plan for Evaluation 12](#_Toc21978048)

[5.4. Conclusions 12](#_Toc21978049)

[6. Issues and Future Work 13](#_Toc21978050)

[6.1. Introduction 13](#_Toc21978051)

[6.2. Issues and Risks 13](#_Toc21978052)

[6.3. Plans and Future Work 13](#_Toc21978053)

[6.3.1. GANTT Chart 13](#_Toc21978054)

[Bibliography 14](#_Toc21978055)

# 1. Introduction

**As least 2 pages, but as many as you like**

## Project Background

What makes a movie successful? We have all watched movies that evoked strong emotions out of us. Perhaps it was the relatable characters, or maybe a resonant ideology was presented, or maybe simply the visuals and art in the movie were impressive which left its mark on you. Story telling a core aspect of our cultures. We have told stories for as long as we can remember, and we don’t even know what the first story was ever told [1]. Even back into prehistory, humans drew pictures on cave walls to tell stories. As humans evolve and develop more advanced tools and civilisations, so do the methods of storytelling. Before any civilisation we told stories only by word of mouth, then with the discovery of fire we also discovered that charcoal is great for making marks on stone walls, Chauvet Cave is probably one of the most famous examples of prehistoric cave paintings. Later with the discovery of writing, it transformed how we tell stories again, taking on the form of symbols on clay tablets. One of most ancient surviving literary work we have today is the Ancient Sumerian “Epic of Gilgamesh” [2], composed in Mesopotamia around 1800 BC, but only second after the Pyramid Texts in Egypt which have been dated to about 2400–2300 BC [3]. Surely these stories didn’t survive thousands of years only because they were written on persistent material, because if that were the case then we would a lot more clay tablets of ancient literature. It is more likely that the Epic of Gilgamesh was such a profound tale to the ancient people of Mesopotamia that they made effort to make copies of the clay tablet and put in effort into keeping them safe for preservation. Today the most advanced form story telling comes in the form of digital media, such as movies, tv series and even video games.

But what makes a story more popular, successful and with more chance to survive into the future for future generation to hear or watch these stories? Is it solely dependent the story itself? All good stories will become successful based on their merit alone or maybe there are more factors involved. Perhaps the way the story is told is more important than the story itself? Two different people can retell the exact same story and one can bore us to death, the other can grip our attention with such intensity that we sit on the edge of our seats, totally immersed in the plot of the story being told. The same is probably also true of movies. The production, cast and delivery of the movie can have a major impact on the movie’s ability to grip us and leave a strong impression on us or leave us bored, forgetting about the movie after a few days. This project aims to explore the relationships between production variables in making a movie and how successful the movie was after its release. How does one measure the success of a movie in the first place? Everyone has their own subjective view on which movie was amazing and which ones disliked, but yet there is a general consensus of the majority on which movies were generally better than other. Main gauge used in determining this are movie ratings and reviews. Ratings by the regular audience and ratings by movie critics. This places a numerical value on movie success. One can even consider the revenue the movie produces as a measure of success. By placing numerical values on production variables, such as the budget to produce the movie, the amount of awards the main actors have, the success of previous movies directed by the directed, we can use computational data analytic techniques to spot patterns and correlations between production variables and the variables which dictate the success of the movie.

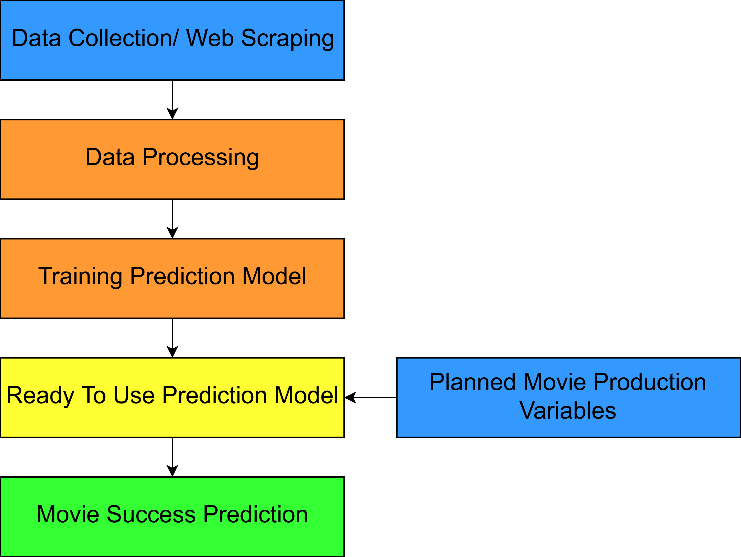
## Project Description

This project will use artificial neural networks and machine learning to learn about the production variables and the success of thousands of movies from the past, and using these models predict the success outcome of planned movies once given the planned production variables, such as the budget of the movie, its runtime, the main actors, writers, directors, genre, release date and other relevant factors.

To be able to use this system, it will be made in the form of an online web application. The end users will have an interactive and friendly user interface to fill out the details of the planned production variables and then get an estimate on what the success of the movie would be based on the user’s input.

To make the model in the first place, the key aspect of this project will be data acquisition, which then we can mine to acquire valuable information from. Basic datasets will be acquired online from sites like kaggle.com and IMDb. But these datasets will not be enough, to gain more relevant data it will use data scrapping techniques to gather relevant data from movie websites like IMDb and Rotten Tomatoes.

The datasets will then be processed into a single table or file in a format that will be readily acceptable to train the artificial neural network. This will include hot encoding categorical data into binary representations, normalizing highly variant data like budget and runtime and placing the data into formatted arrays. The neural network model will train on the provided dataset to predict movie success and will also be tested for accuracy on a subset of the dataset which the model has not been trained with.



The end system will have a client-server architecture. The client side will be what the end users will interface will, the processing of input data, the prediction model and the database will be located on the server. The server will also use web scraping to get additional needed date. Even though the server will have a database of movie data, some variables like how many awards won by an actors or date as possible, the server will web scrape for that data and use it as input for the prediction model to get an estimate on the movie success and return to the result to be visually viewed by the end users.

## Project Aims and Objectives

Overall aim and some milestones along the way to achieve the aim

* 5-9 objectives

The overall aim of the project is to make an estimate prediction of movie ratings greater than 50% accuracy. To provide a web-based user interface for the system that anyone can use.

In order to have the most accurate possible predictions, a large amount of data is required.

## Project Scope

This project is a probabilistic estimate of the movie’s success. It cannot predict with certainty. The project also does not take into account the actual quality of the content of the movie. It doesn’t analyse the plot, complexity of the story, character development and quality of acting by the actors. It doesn’t analyse the style of the movie or it’s visuals. It doesn’t analyse the content of the movie’s scripts, or the title of the movie or it’s summary. To predict movie success based on its content would require the movie to have already been filmed and edited for analysis, which at that point is impractical because one would want to know a rough estimate of the movie success before money is invested into its production. Using sentimental or thematic analysis for things like movie title, plot summary or script doesn’t return meaningful numerical data that be used to train an artificial neural network. This project doesn’t predict success for TV series. The success of a TV series is highly variable and can depend on season and even episode. First season of a TV series can be highly successful then dwindle over time, so having only initial production data cannot predict the success of a TV series over the long term.

## Thesis Roadmap

One sentence explaining what each of the following chapters is about.

# 2. Literature Review

## 2.1. Introduction

This chapter explores the background research related to artificial neural networks, data mining and role of artificial intelligence in movie success prediction. There are a lot of problems in the world and for most of human history it was the humans themselves who could solve their problems. This has worked very efficiently for thousands of years as we can how we developed and progressed our civilisation from the stone age to the modern age of information and technology. As our inventions and civilisation have been more complex, so have our problems. We now have access a tremendous amount of raw data that we collect with our technology but trying to extrapolate meaningful information about the data is becoming increasing more difficult. Dealing with Big Data is becoming a new problem. Manually analysing Big Data and hoping to find patterns and extracting useful information to make prediction is impractical by humans alone. For example, given 30’000 rows of movies with many columns of metadata about the movie, how would it be possible to predict a planned movie’s ratings or box office income? This where the use of artificial intelligence comes into play. The most powerful aspect of artificial intelligence comes from machine learning techniques such as the use of artificial neural networks. Using the data acquired, we can train the neural networks to find and recognize patterns in the data, creating a model which can be used to make predictions from new observations as inputs. The whole processes of acquiring large data sets, processing them and using artificial intelligence to extract useful information and make future predictions is called data mining.

## Artificial Neural Networks

Artificial neural networks are statistical/mathematical models that try to imitate real biological neural networks like the human brain. Artificial neural networks are made up of interconnected nodes, or “neurons”, separated out into layers. The connections between the nodes are called synapses, which send signals from one neuron to another. The synapses also have an assigned “weight” to it. Each neuron has an activation function, it calculates the total sum of the weights it received from signals from other neurons, and if the sum is grater than the threshold, it actives and send a single to each neuron it’s connected to in the next layer.



What makes Artificial Neural Networks very powerful is backpropagation. When the output of the model does not match the expected output during training, it calculates the degree of error from inputs, expected output and actual output, then iterates backwards through the networks, adjusting the weights and biases of the synapses, in hopes that this will create an output to more closely match the expected output.

Once an Artificial Neural Network has been sufficiently trained, we can input new data into the network to see what the predicted output is. This project will use an artificial neural network, trained on datasets of metadata of previously made movies to then make predictions with new given input.

## Data Mining

Data mining is the whole process of finding patterns in large dataset and extracting useful new information. Data mining involves using artificial intelligence, machine learning and statically analysis. Data mining also involves databases and data management, working with pre-existing datasets, pre-processing the data, using machine learning techniques to analyse the data to find new patterns, create statistical prediction models, post-process the raw data from the models then use visualisations to display the data in a meaningful and comprehensive way as to be understood by humans. Data mining can produce models with powerful predictive abilities which can solve business problems and predict trends.

## 2.2. Alternative Existing Solutions to Your Problem

Software you’ve looked into

The use of artificial intelligence in Hollywood’s film industry is become more prevalent in the last decade.

Cinelytic, vault-ai, scriptbook

## 2.3. Technologies you’ve researched

Programming languages, operating systems, etc.

## 2.4. Other Research you’ve done

Domain specific research

## 2.5. Existing Final Year Projects

Secure Document Sharing - Owen Kane

This project creates a secure online system to create, edit and share documents over the internet. It uses client-side AES encryption algorithm to encrypt the files before they are sent over the internet. This way the data will never be sent in plain text format for any man-in-the-middle to see the contents of the data in case where they are sniffing and capturing passing packets online.

This is a good approach to file sharing. This increases the privacy and security of data from being access by unauthorized users. The technologies used are also like what I want use, like Python and JavaScript, in a client-server architecture. Any transition of data between the tiers in the architecture use a secure encrypted transfer protocol, SSL/TLS. SSL is used when data is retrieved from the database to the server, and again when data is sent from server to client and vice versa. This a good approach, with I’ll have do the same in my own project.

The project was very well tested. Used multiple types of tests, such as ad-hoc testing, unit testing and integration testing. Testing is vital to any coding project, but more so to project with computer security as possible bugs in the guys can expose vulnerabilities and opportunities for hackers to steal confidential or sensitive data.

Education Tool for Web-Based Vulnerabilities - Cormac Kelly

Interesting project scans your Java files for possible SQ L Injection vulnerabilities. It is designed as an education tool. I like the way it is a web application, making it accessible and easy by the user. It encourages to design code with security in mind and using this tool as quick test for any obvious security flaws pertaining to SQL Injection. I like the idea behind the project, to raise awareness about computer security and encouraging to write secure code.

The project used many technologies and languages. For the code base, Python, Java and JavaScript were used. These are well suited and straightforward languages to use to make a web application and the server back end. These languages also have graphical user interface libraries to make the program easily accessible.

I like this project due to its emphasis on the user interface. It’s perhaps the most important aspect of any software because that’s all the user is going to see. It’s important that is comprehensive and easy to use. As I will also need a user interface for my web application that doesn’t look confusing or bland.

## 2.6. Conclusions

# 3. Prototype Design

**As least 6 pages, but as many as you like (but lots of diagrams, which count towards the page total).**

## 3.1 Introduction

## 3.2. Software Methodology

## 3.3. Overview of System

Include a diagram

## 3.4. Front-End

Including screen prototypes and Use Cases

## 3.5. Middle-Tier

## 3.6. Back-End

Including ERDs, and maybe ISDs

## 3.7. Conclusions

# 4. Prototype Development

**As least 2 pages, but as many as you like (but lots of code samples).**

## 4.1. Introduction

## 4.2. Prototype Development

## 4.3. Front-End

## 4.4. Middle-Tier

## 4.5. Back-End

## 4.6. Conclusions

# 5. Testing and Evaluation

**As least 2 pages, but as many as you like**

## 5.1. Introduction

## 5.2. Plan for Testing

## 5.3. Plan for Evaluation

## 5.4. Conclusions

# 6. Issues and Future Work

**As least 5 pages, but as many as you like**

## 6.1. Introduction

## 6.2. Issues and Risks

## 6.3. Plans and Future Work

### 6.3.1. GANTT Chart

# Bibliography