

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

(not sure how to summarize this project without being very repetitive)

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

How should write the acknowledgements?

Table of Contents

[1. Introduction 7](#_Toc26629990)

[1.1. Project Background 7](#_Toc26629991)

[1.2. Project Description 7](#_Toc26629992)

[1.3. Project Aims and Objectives 8](#_Toc26629993)

[1.4. Project Scope 9](#_Toc26629994)

[1.5. Thesis Roadmap 9](#_Toc26629995)

[2. Literature Review 10](#_Toc26629996)

[2.1. Introduction 10](#_Toc26629997)

[Artificial Neural Networks 10](#_Toc26629998)

[Data Mining 11](#_Toc26629999)

[2.2. Alternative Existing Solutions to Your Problem 11](#_Toc26630000)

[Cinelytic 11](#_Toc26630001)

[Vault AI 12](#_Toc26630002)

[ScriptBook 13](#_Toc26630003)

[2.3. Technologies you’ve researched 13](#_Toc26630004)

[Machine Learning Technologies 13](#_Toc26630005)

[Web Application Technologies 14](#_Toc26630006)

[2.4. Other Research you’ve done 14](#_Toc26630007)

[2.5. Existing Final Year Projects 15](#_Toc26630008)

[2.6. Conclusions 16](#_Toc26630009)

[3. Prototype Design 17](#_Toc26630010)

[3.1 Introduction 17](#_Toc26630011)

[3.2. Software Methodology 17](#_Toc26630012)

[Agile Methodology 17](#_Toc26630013)

[Feature-driven development 18](#_Toc26630014)

[3.3. Overview of System 18](#_Toc26630015)

[3.4. Front-End 19](#_Toc26630016)

[3.5. Middle-Tier 20](#_Toc26630017)

[3.6. Back-End 21](#_Toc26630018)

[3.7. Conclusions 22](#_Toc26630019)

[4. Prototype Development 23](#_Toc26630020)

[4.1. Introduction 23](#_Toc26630021)

[4.2. Prototype Development 23](#_Toc26630022)

[4.3. Front-End 23](#_Toc26630023)

[4.4. Middle-Tier 23](#_Toc26630024)

[4.5. Back-End 23](#_Toc26630025)

[4.6. Conclusions 23](#_Toc26630026)

[5. Testing and Evaluation 24](#_Toc26630027)

[5.1. Introduction 24](#_Toc26630028)

[5.2. Plan for Testing 24](#_Toc26630029)

[5.3. Plan for Evaluation 24](#_Toc26630030)

[5.4. Conclusions 24](#_Toc26630031)

[6. Issues and Future Work 25](#_Toc26630032)

[6.1. Introduction 25](#_Toc26630033)

[6.2. Issues and Risks 25](#_Toc26630034)

[6.3. Plans and Future Work 25](#_Toc26630035)

[6.3.1. GANTT Chart 25](#_Toc26630036)

[Bibliography 26](#_Toc26630037)

# 1. Introduction

## 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.

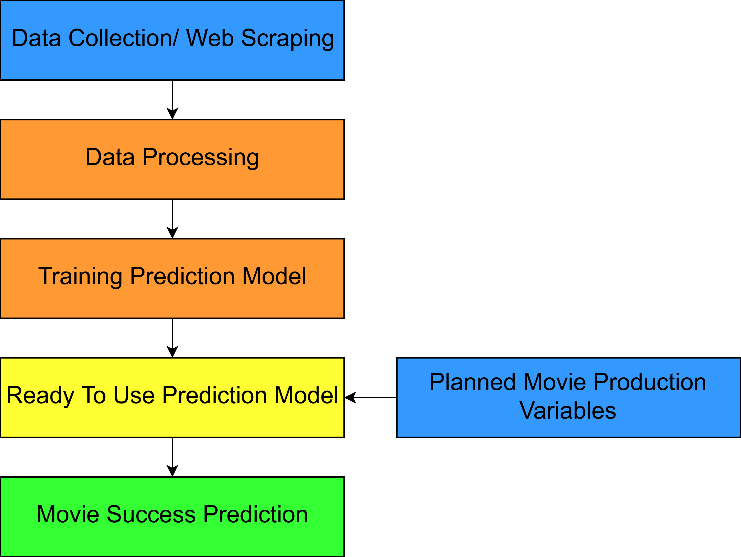
## Project Description

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.

(not too sure how to write the objectives, I would have thought it would be clear from project description my objective is to make movie success prediction model as a web application)

## 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.

(I plan to write this after I finish writing the chapters)

# 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.

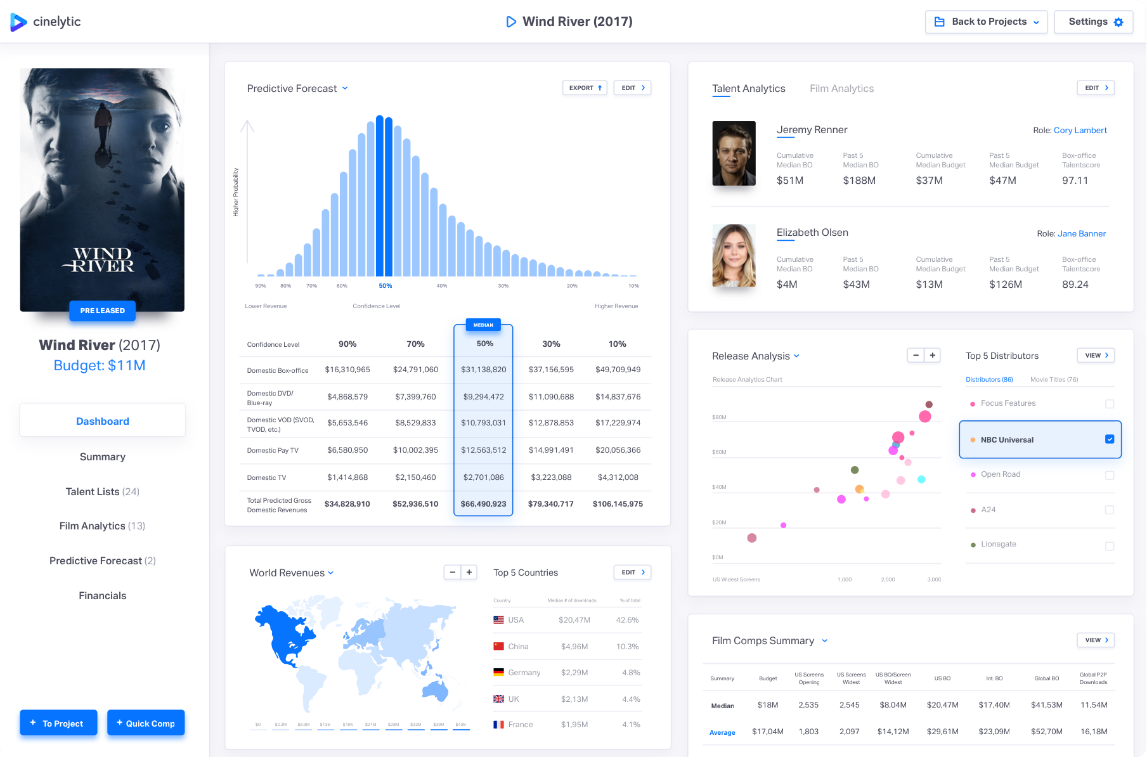
## 2.2. Alternative Existing Solutions to Your Problem

The use of artificial intelligence in Hollywood’s film industry is become more prevalent in the last decade. Making movies can be very expensive, so the production companies want to know if their idea for a movie will be worth while making. There are several companies that have started up in the last decade to provide a software solution for this. Companies like Cinelytic, Vault and ScriptBook use data mining and artificial intelligence to make estimate predictions on what will be the movie’s box office based on budget, planned actors for the cast and other relevant factors. [4]

### Cinelytic

Cinelytic is a software platform which uses artificial intelligence to provide analytic insight for movie studio and independent film makers on possible success and profit for the movie they plan to produce. Cinelytic uses data that is both proprietary and licensed by third parties to create prediction models that allow the users to evaluate the project value and minimised risks. [5]

Cinelytic makes movie success predictions from inputs such as the movie’s budget and the main actors cast in the movie. It provides comprehensive visualisation of data and graphs of the results, each after easily understood and business decisions can be made quickly. Cinelytic has been proven to be useful to film makers for “unparalleled accuracy”. Cinelytic software doesn’t touch on the creative parts of the movie projects. CEO of Cineltyic, Tobias Queisser, said “we want to bring the ‘gut’ part of the decision down to 60%. The creative part should probably still override, but in order to create a better product, execute and market it better, and find a more financially satisfying outcome, it helps to use a more methodical approach to project evaluation and risk assessment” [6].



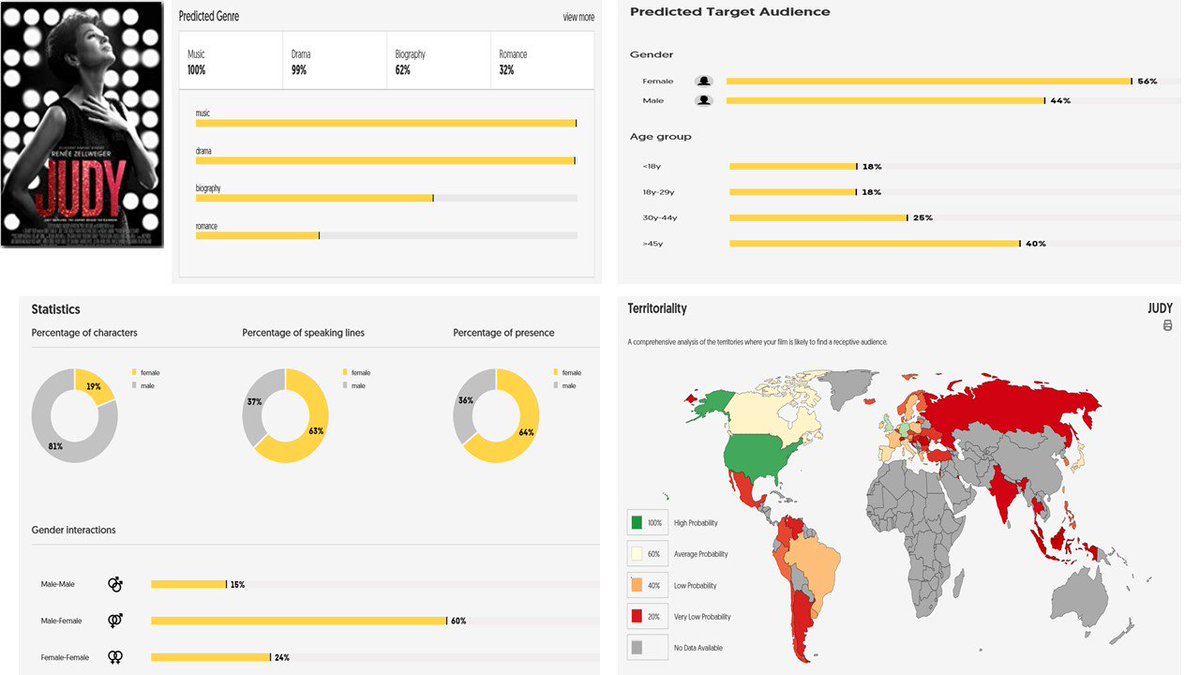
### Vault AI

Vault AI is an Israeli company that uses artificial intelligence that analyses the actual content of movies. Vault “reads” the screenplay and analyses the script, things that happen in the scene, character development and even plot structure. It then is able to make a prediction of the box office income of the movie [7]. Vault AI have several products that offer different type of analysis of movies. One of them is “RealAudience” which uses artificial intelligence to predict who will be the main global demographic who will watch the movie, and how much tractions the movie will get from these audiences, even suggest improvements for the movie to attract even more people of a specific demographic. [8]



### ScriptBook

ScriptBook is a company that uses artificial intelligence with natural language processing to analyse the script and story of the movie. It analyses the characters and story of the movie script and makes predictions of what the audience ratings will be like, which characters are likeable, the genre of the script and even to which audience does the script appeals to most. This doesn’t make box office predictions like the previous software do but provides deeper insights into the story itself [9]. ScriptBook have another very interesting product based on artificial intelligence called DeepStory. Having analysed thousands of movie scripts, DeepStory can generate compelling stories and scripts by itself. DeepStory has advanced character awareness, creating characters with their own personalities and traits. [10]



## 2.3. Technologies you’ve researched

### Machine Learning Technologies

There are several machine learning technologies that implement artificial neural networks. It’s is possible to implement an artificial neural network with most object-oriented programming languages such Java and C++. Instead of trying to build an artificial neural network from scratch, it was a better to option to use programming language libraries that allow easy set up and utilisation of artificial neural networks. Python is very popular today due to its simplicity, flexibility and available libraries to be used. This project will use python with third-party python libraries to create, train and use artificial neural networks.

TensorFlow is a python library that provides all machine learning capabilities. TensorFlow is a free and open source library. TensorFlow can be used to implement any type of machine learning technique such as deep learning. Due to this, TensorFlow has a very flexible architecture making this library applicable to most artificial intelligence projects. TensorFlow was chosen for this project due to its wide application and its popularity online means there is a lot of support and documentation for this library.[11]

Keras is another open source python library that implements artificial neural networks. The main advantage of Keras is the ease of use. Keras has a very simple and straight API functions to allow for fast creation of neural networks and testing. Even thought Keras is a standalone package, TensorFlow uses Keras as a submodule in its architecture. Keras is the front end of the TensorFlow library. This combination provides a powerful library, with user-friendly API functions to create and train neural networks with the impressive capabilities of the TensorFlow backend giving the library wide application and flexibility. This high level of abstraction allows the focus to remain on the high-level design and implementation of the neural network without having to worry about the low-level technicalities of neural networks. [12]

I have chosen TensorFlow and Keras as technologies to be used in creating and training movie success prediction models.

### Web Application Technologies

My goal is creating a web application as the front end for my project. Everyone has access to the internet and hosting my prediction model online gives it wide availability and ease of use. I have chosen python as the programming language of choice for this project, due it the simplicity of use and fast development, this project will also use python-based web framework to host my web application.

Django is a python based free open source web framework. The main advantage of using Django is due to its simplicity of use, plug-and-play architecture and flexibility. This allows for faster design and creation of website without having to deal with the technical low-level details associated with web frame servers. It implements website secure as core of its design allow the developer to avoid the most common security flaws in website design. Django is lightweight and highly configurable. It is compatible with most common sever technologies such as Apache and Nginx. It supports a database backend for most common database technologies as well, such MySQL and MongoDB. Django is highly extensible and allowed for configuration of third-party application and packages. This project will use Django as the web frame to create the web application due to the many benefits stated above, especially ease of use and rapid development.

As part of the website, I will use JavaScript to create an interactive and dynamic website. JavaScript is most widely used and well documented technology, there is no reason to use anything unorthodox.

## 2.4. Other Research you’ve done

Domain specific research (how many papers should I review? I just reviewed 1 so far, I will come back and add more research reviews later)

There are several research papers online that have experimented with using data mining to predict movie success. They are very similar in approach to this project. This project uses some of the ideas presented in these research papers.

The most recent paper talking about movie success prediction is by Saurabh Kumar from VIT University. In that project, a dataset was downloaded from Kaggle.com that had 651 rows, meaning 651 movies. Additional data about the movie such as it’s run time was acquired by web scraping IMDb pages with the relevant movie. Movie’s success was measured by IMDb ratings, number of votes, critic and audience ratings, critic and audience scores for the movies. After cleaning the dataset, the data was used to train a naïve Bayes classifier model with supervised learning method like stochastic gradient decent. Using input variables like budget, number of Oscars won by the cast playing in the movie, and using audience score of the movie, made predictions of the IMDb ratings with almost 80% accuracy. [13]

## 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

(What should I include in the conclusions section?)

With the current research available, it provides a good guideline of what is a good approach to the problem of trying to predict movie success, and which variables are highly correlated to the movie ratings and which are correlated to movie box office. The research shows that it is indeed possible to have a good estimate of around 70% accuracy what the movie ratings will be just by taking into account production variables such as budget, runtime, genre and the awards and success of the actors in the movie. Based on the successful use of different machine learning techniques, it seems that a neural network would work well and provide similar level of accuracy as the prediction models that did not use artificial neural networks.

# 3. Prototype Design

## 3.1 Introduction

Following on from the previous chapter, this chapter will discuss how those technologies will be implemented in this project and what the over all design of the project is going to look like. This chapter will discuss software development methodology which will be used during the development of this project and the architecture design of the web application. This will include diagrams of system design, class diagrams of code, diagrams of web application architecture and use case diagrams. The final section in the chapter will discuss the type of testing that will be used and how this projected will be evaluated.

## 3.2. Software Methodology

### Agile Methodology

Agile methodology is a type of software development. Agile uses an incremental approach in developing software. It is a fast approach to writing software whilst remaining flexible to any changes in the requirements of the software. With Agile, the work is split up into time periods called “Sprints”, usually about 2 weeks, where the development and testing of the software is done. At the start of each sprint, the team, called a “scrum team” have a meeting to plan out what tasks and features of the software should be developed over the course of sprint. The team discusses the plans with a product manger who relays any requests from the product customers, takes into account any needed changes in the software requirements and plans out the work to be carried out. The team then works to achieves the set-up goals and complete the software development tasks laid out in the meeting. This makes Agile mythology into a cyclical development method, where the work done is reviewed and taking into consideration when planning the next steps in development.

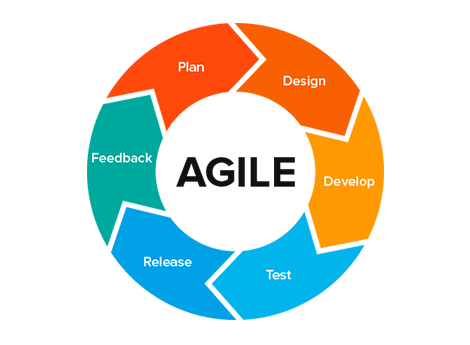


Diagram of Agile development cycle

### Feature-driven development

Feature-driven development (FDD) is a type of software development methodology similar to Agile methodology. FDD is also an iterative and incremental approach to software development. FDD main focus is creating feature for the software, focusing one feature at a time before moving on to the next feature. This is where it differs from Agile methodology, Agile do not focus on individual feature as much as they focus on breaking up the project development plan into small to-d0 tasks and plan to implement a certain number of tasks in a Sprint. FDD instead focus on developing fully working individual features in accordance to the principles outlined in the “Agile Manifesto”. With this, FDD combined the best industry practices of software development into one cohesive whole. The main advantage of using FDD is that it’s simple 5 step process makes it easy to fast develop and deliver tangible results of the software, present working software features to the costumer. Just like agile, FDD also remains very flexible to any change in software requirements and rapidly adapts does changes.

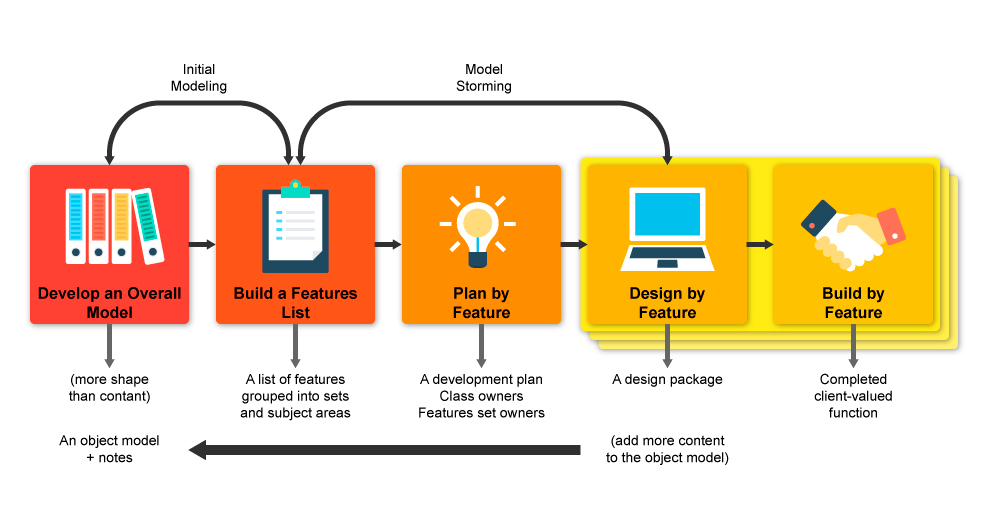
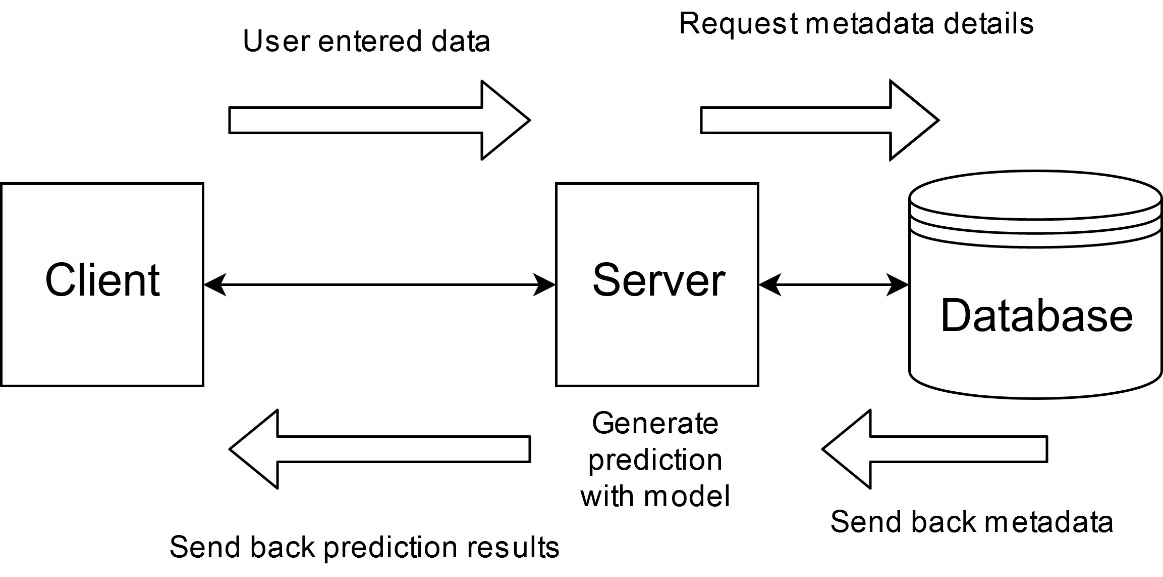


Diagram of FFD development

## 3.3. Overview of System

The main system of this project will be in a form of a web application. It will be a 3-teir architecture, client-server with a database in the backend. Python Django will be the web frame used to host and run the web application. The front-end will be a light client web page, made from HTML\CSS and using JavaScript to make it interactive and responsive. The front end will look like a form with fields to fill out with details about the movie they want a prediction on. The data will be sent to the server for processing. The database will store all the metadata, such as details on how many awards an actor has won and how many movies with their ratings has a director directed. The server will get the needed extra metadata and processes into a format ready to be used by the model. The model will analyse the given data and give a prediction of the movie ratings and box office, the server will post process the results from the model into a comprehensible form, make visual representations of the data and send it back to the client to be viewed by the end user.

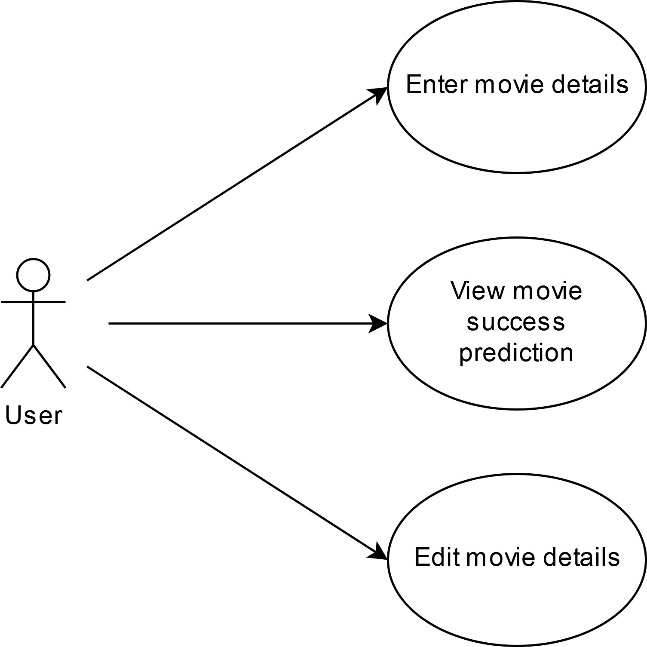


Overview diagram of the web application and data flow

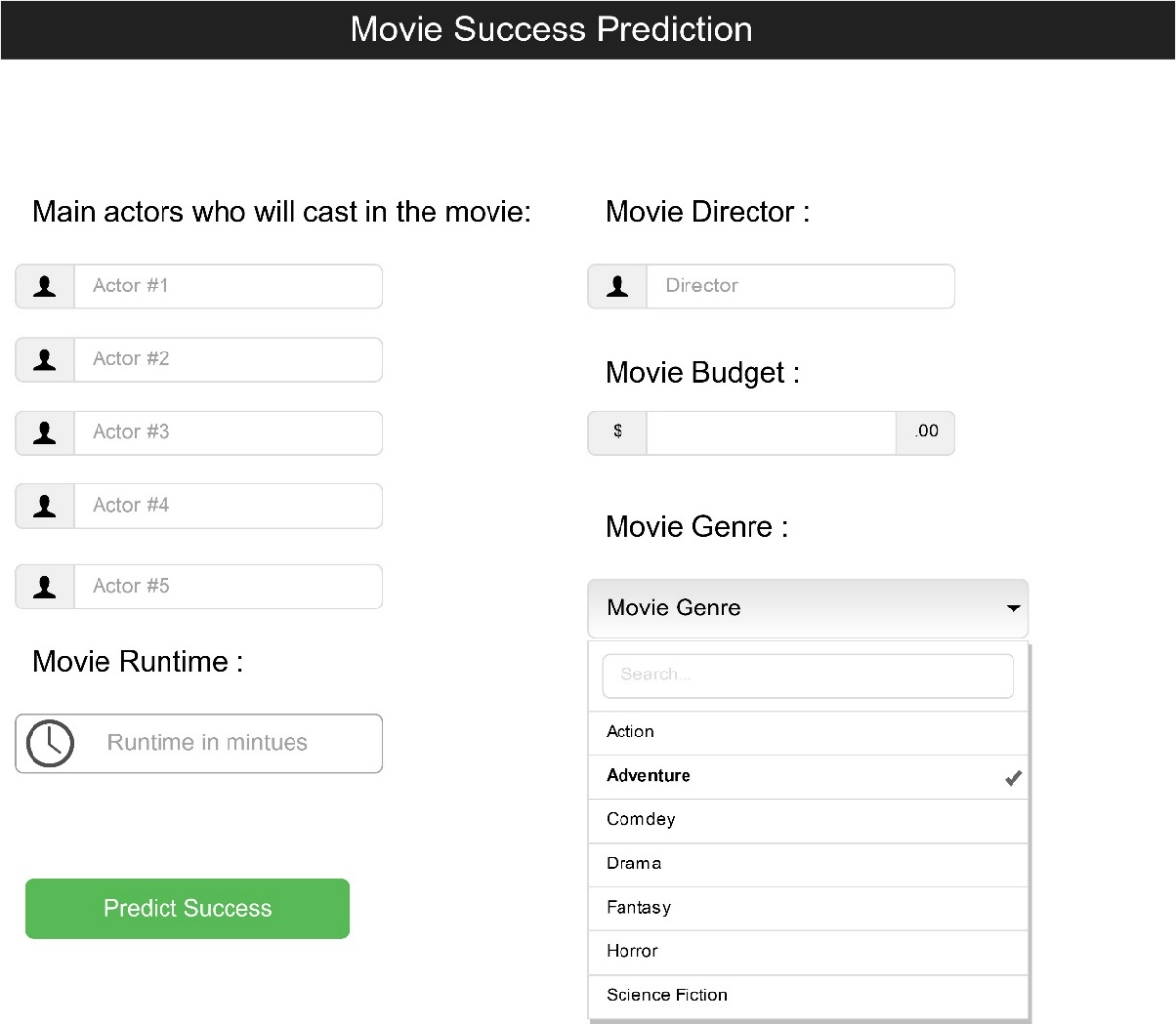
## 3.4. Front-End

The front-end, the client, will be a simple web page. Written in HTML for basic web elements, and CSS for styling and making the website look good. There will be several web pages, first one to be the form where the users fill out the details about the movie and a results page where the prediction and visual data will be shown to the user. None of the processing of data will be done client side, the web pages will just take user input and send that data to the sever for processing.

The use case for the web app from the user perspective is simple. The user goes on the web page, they will enter the details for the movie and click the button to predict success. The button will take the user to the results page to view the movie predictions. The use than will have the option to edit the movie details, which will bring them back to the original form page, but the old data is still in the fields, so the user only needs to change some fields as they wish instead of filling out the whole again.



User use case diagram



Prototype of from web page

The form will have 5 fields for actors. It will have interactive feature whereas the user types in the names of actors, a small drop down will appear of actors in the database, by searching the database for the actors that match the current user input, like half names and surnames. This is a way to autofill in the field and make the form filling faster, making the website feel responsive and interactive. The same goes for fields for movie director. Have a numerical field for movie budget and movie runtime. A selection for movie genres will have a drop-down list to genres the available movie genres. A button to with “Predict Success” will send all inputted field data to server and then the button will redirect the user to the prediction results page.

## 3.5. Middle-Tier

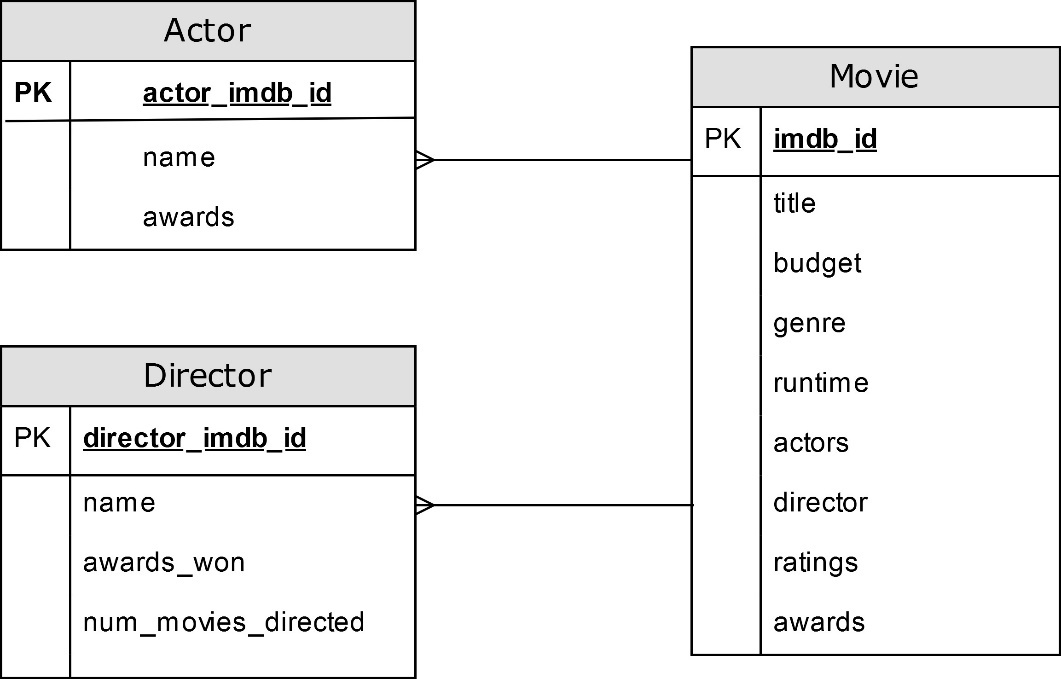
Middle-tier will be run on a python Django server. The server will have all the logic behind the web application. The movie success prediction model will be stored on the server. The sever will receive input from the client, movie data to make a prediction on. The server will query the back-end database to get metadata on the data, such as number of awards won by the actors which the user entered. Python scripts on the Django servers will process the data and normalise the data to be used by the prediction model. Run the data through the prediction model to request an output. The project will make several types of predictions, ratings of the movie and box office income. For this there needs to several prediction models, one for each type of prediction. The sever will have to run the processed data through all the prediction models. The results are going to be post processed into a readable comprehensive format. The server will merge all the outputs together into one dataset. Using the dataset with results, the data will be sent back to the client to be displayed. Generate graphs and visual representation of the prediction results.

To keep the data on the database up to date, and to keep the prediction models accurate, once a week, the server will do web scraping to check if any of the data on the database is different on the source websites, like IMDb and Rotten Tomatoes. The database will be updated with the new data and then using the database, re-train the artificial neural networks to create new and up to date prediction models, ready to be used by end users.

## 3.6. Back-End

For the back end, there will be a MySQL server holding the all data about the movies, actors and directors. There will be 3 tables in the database, one for movies, actors and directors. The main advantage of using a database is faster performance on processing data. If there was no database, the server would have to web scrape movie websites to get detail metadata on the actors and movies. This can take time, depends on the internet connection and depends on the movie websites to running. If there a lot of users using my web application, this will slow down the application significantly. Using a database solves those issues.

The database will be made with MySQL. It is free and open-source relational database. It has also has a driver, like a library or API, for python making it easily compatible with Django and the all the python scripts running on the server.



ERD of the movie database

## 3.7. Conclusions

This chapter discussed the overall design of the web application and the technologies used. Discussed the software development methodologies, with their advantages. Showed the dataflow between architecture tiers, from the front end all the way to the back end.

The next chapter will discuss the progress made so far on the project, what has been implemented and what challenges have been encountered so far, and possible solutions to them.

# 4. Prototype Development

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

## 4.1. Introduction

(I’m not sure what else to add here, I’m just presenting my progress)

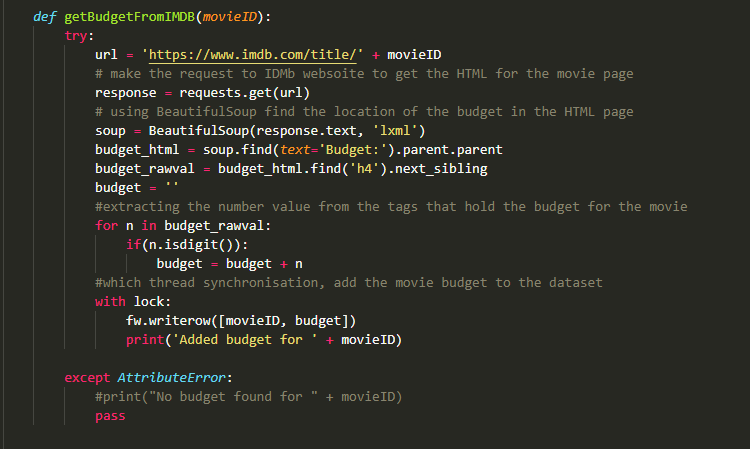
This chapter discussing the progress made so far in this project.

## 4.2. Prototype Development

The first part in developing this project was setting up a version control system. This project uses Git with the online GitHub integration. Git was the version control of choice as it’s the most popular version control software that’s effective and east to use. GitHub allows the project to be stored in online allowing synchronisation of my project across multiple deceives.

## 4.3. Gathering Data

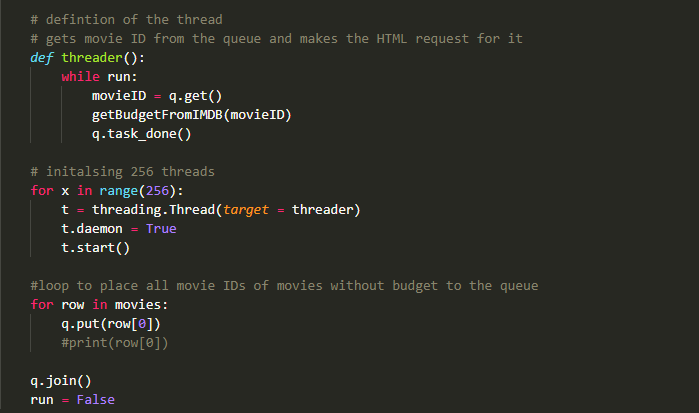
In order to train an artificial neural network a large amount of data is required. Kaggle.com is an online website that publishes datasets, free to download and use. I found a dataset with 350’000 movies with details like budget, genre and runtime [20]. That dataset also had tables with main casting in the movie. IMDb website also have a dataset available to download [21]. This dataset contains all the movies and tv series on the IMDb website, the movie ratings and titles with their IMDb IDs. The Kaggle dataset was incomplete, there were movies that were missing the budget for example. To get this additional information, I web scraped IMDb website to find movie budget. Using python libraries “Requests”, the web scraping script made HTTP requests to the IMDb websites, using IMDb movie IDs as part of the request URL. Using python library “BeautifulSoup”, the received HTML DOC was parsed to the budget listed for that movie on the web page, extracted the number value and added the budget to the in the movie dataset row where the movie budget was missing.



Python code used in the requesting the movie web page and parsing for the budget

From the multiple dataset and filled in I wrote a python script to parse through the datasets and merge the data into a single CSV file. This file will then be read by python script to train the neural network

One challenge in the web scraping was the speed of getting and processing the HTML docs. There were about 175’000 movies which didn’t have budget metadata. Making a request one at a time to check for budget on all the movies would have taken a few days to complete. To speed up the process, I used 256 threads to make 256 requests at a time and parse the received HTML doc. This greatly improved web scraping performance, instead of taking few days to complete, it took a few hours instead.

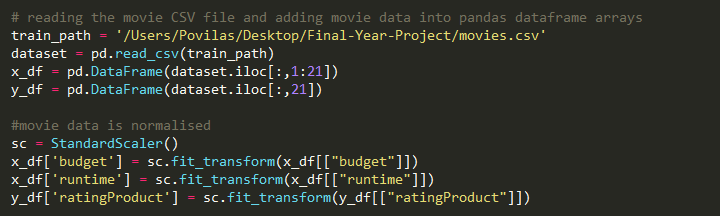


Python code using threads to make HTML requests

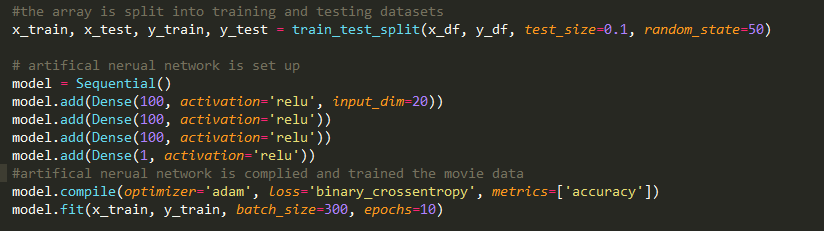
## 4.4. Artificial Neural Network

After the necessary data has been acquired, from datasets and web scraping. It is time to train the artificial neural network.

First the data is imported from the final csv files that has all the necessary movies with complete metadata. Using python library Pandas, the data is split into 2 “DataFrames”, which are like arrays. First array holds the movie production variables such as budget, genre and run time. The second array holds the ratings for the movie. The first array will be used as inputs for the neural network and the second array is the expected output, so the network can calculate its loss and configure it’s weights. Before the data is inputted into the neural network, the input data needs to be normalised, the genre is represented as binary input into the networking while the budget are disproportionally very large values. Normalising the budget puts it at same level as all other variables making it efficient to be used by the neural network.



The neural network is created with the TensorFlow and Keras python libraries. A densely connected sequential neural network is made with 2 hidden layers and 100 neurons in each layer. The nuerual network is compiled and the processed datasets from above are fitted and the neural network is trained to make a movie prediction model to predict movie ratings.



## 4.5. Conclusions

(I am confused on writing conclusions, what am I to conclude from developing my prototype, that’s not yet predicting anything?

# 5. Testing and Evaluation

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

## 5.1. Introduction

Last chapter we discussed the project development progress so far. This chapter will explore how the code will be tested to ensure proper functionally and how the projects web application will be evaluated.

## 5.2. Plan for Testing

The project application will be tested all the time. Firstly, manually testing by running the code and checking it works as expected. The application will be run on a server, which can be hosted in the local machine. This allows extensive test done immediately with every implementation of a feature or change in code. It is possible to set up automated tested using python libraries like Selenium that automate browser activity.

Whitebox testing will be used to test the source code running the server and data processing code. Unit test will be used to ensure the code internally functions the way it’s meant to. Creating test cases will allow for automated testing to be executed.

Web scraping will be tested using Blackbox testing. Testing python web scraping scripts is important as to ensure the HTML pages being returned from online are the right pages containing the right movie data. To test this, a small dataset will be manually complied with reference to the online sites. This is a small dataset with known and expected outputs. Then running the web scraping scripts and comparing the results from the script is the dataset of expected results will ensure web scraping is working as intended and returning accurate data.

The neural network is always tested after training. The dataset is split up into two, one for training the model with, the other for testing. After the model is trained, the test data is run through the model to see if the predictions made are accurate, and to what degree. Instead of configuring the network, it just runs all the movies not used in the training through the model to test the average accuracy in prediction of movie ratings.

This is a form of inbuilt Blackbox testing. Since the model is always tested immediately, this allows for quick development and configuring the neural network to make increase it’s accuracy.

## 5.3. Plan for Evaluation

( what should be included in the evaluation section? I think it’s something about letting the users test it and use the web application and listen for their feedback.

## 5.4. Conclusions

# 6. Issues and Future Work

**As least 2 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

[1] Storytellingday.net. (2019). History Of Storytelling – How Did Storytelling Begin?. [online] Available at: <http://www.storytellingday.net/history-of-storytelling-how-did-storytelling.html>.

[2] Miller, L. (2019). The oldest story ever written. [online] Salon. Available at: <https://www.salon.com/2007/04/24/gilgamesh/>.

[3] Mark, J. (2019). The Pyramid Texts: Guide to the Afterlife. [online] Ancient History Encyclopaedia. Available at: <https://www.ancient.eu/article/148/the-pyramid-texts-guide-to-the-afterlife/>.

[4] Vincent, J. (2019). Hollywood is quietly using AI to help decide which movies to make. [online] The Verge. Available at: [https://www.theverge.com/2019/5/28/18637135/hollywood-ai-film-decision-script-analysis-data-machine-learning](https://www.theverge.com/2019/5/28/18637135/hollywood-ai-film-decision-script-analysis-data-machine-learning%20).

[5] Cinelytic.com. (2019). Cinelytic | Built for a Better Film Business. [online] Available at: <https://www.cinelytic.com>.

[6] Kay, J. (2019). How data company Cinelytic aims to reduce risk in the film business. [online] Screen. Available at: <https://www.screendaily.com/features/how-data-company-cinelytic-aims-to-reduce-risk-in-the-film-business/5136245.article>.

[7] En.wikipedia.org. (2019). VaultML. [online] Available at: <https://en.wikipedia.org/wiki/VaultML>.

[8] Vault-ai.com. (2019). RealDemand™ Market Intelligence. [online] Available at: [https://www.vault-ai.com/RealDemand-market-intelligence.html](https://www.vault-ai.com/RealDemand-market-intelligence.html%20).

[9] ScriptBook. (2019). ScriptBook. [online] Available at: <https://www.scriptbook.io>.

[10] ScriptBook. (2019). ScriptBook. [online] Available at: [https://www.scriptbook.io/#!/deepstory](https://www.scriptbook.io/%23!/deepstory).

[11] TensorFlow. (2019). TensorFlow. [online] Available at: [https://www.tensorflow.org/](https://www.tensorflow.org/%20).

[12] Keras.io. (2019). Home - Keras Documentation. [online] Available at: [https://keras.io/](https://keras.io/%20).

[13] Kumar, Saurabh. (2019). Movie Success Prediction using Data Mining For Data Mining and Business Intelligence (ITA5007) of Master of Computer Application School Of Information Technology and Engineering. [online] Available at: <https://www.researchgate.net/publication/332396741_Movie_Success_Prediction_using_Data_Mining_For_Data_Mining_and_Business_IntelligenceITA5007_of_Master_of_Computer_Application_School_Of_Information_Technology_and_Engineering>.

[20] Stephanerappeneau. (2017). 350 000+ movies from themoviedb.org. [online] Available at: [https://www.kaggle.com/stephanerappeneau/350-000-movies-from-themoviedborg](https://www.kaggle.com/stephanerappeneau/350-000-movies-from-themoviedborg%20) .

[21] IMDb. (2019). IMDb Datasets. [online] Available at: [https://www.imdb.com/interfaces/](https://www.imdb.com/interfaces/%20).