FYP Project Proposal Form 2019/2020

Student Name: Povilas Kubilius	**Student Number: C16370803**
Mobile Number: 087 337 8177	**Supervisor: Leo Tilson**
Programme Code: DT228	
Project Title: Movie Success Prediction with Datamining	

Summary (approximately 200 words)

The goal of this project is to predict the success of a planned movie, what ratings it will get, how much revenue it will make in the box office, based on planned input variables such as movie budget, genera and cast.

This will be a web application, where users will be able to fill out these planned movie details and my models will make a prediction and show the user what type of rating and revenue the planned could get.

I will gather metadata about movies from available to download datasets and use web scraping techniques to fill in any gaps in the data and acquire any additional needed information. After cleaning and processing my datasets, I will use it train an artificial neural network model to make predictions on the success of movies.

I will host the model online on the interactive web application. Using the entries from the end user, I will web scrape any necessary needed metadata about the entries, such as how many awards does the actor which was inputted by the user have, or how many movies has a director directed and what ratings those movies had, and then use my pre-trained models to make the predictions.

Background (and References)

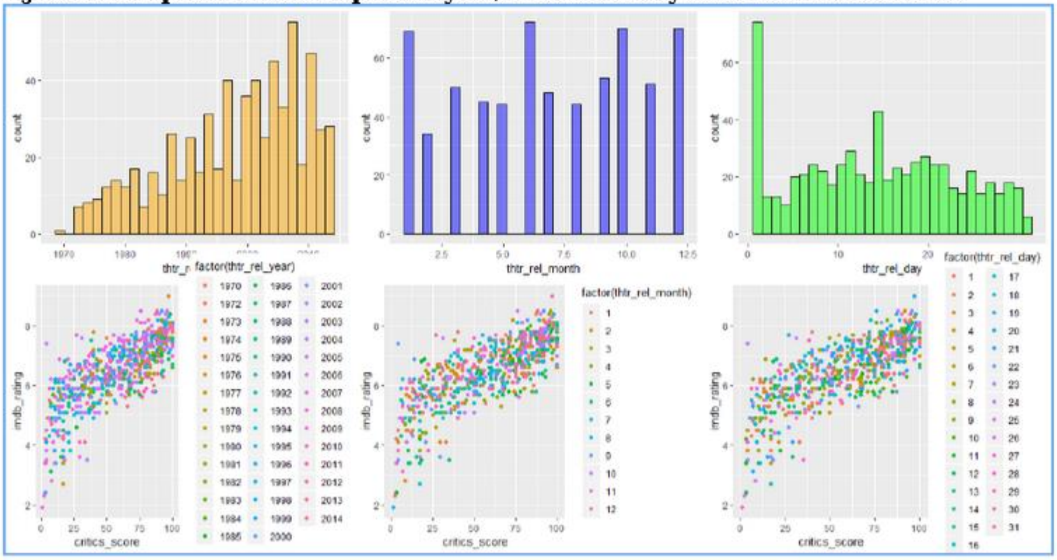
I love watching movies and TV series. As any person who enjoys watching movies, I have also dreamed about my own movie or TV series, imagining the cast that would play my characters, then I also wondered if my idea would be successful, what ratings would my movie get. For this project I want to explore if there is any correlation between things like the movie genre, studio, cast, directors, writers, music composers, even the movie budget and if the movie becomes a big hit with high ratings. If there is a correlation, this would be an invaluable tool to the movie industry, as can have a rough estimate if someone movie idea would turn them a profit.

I assume there must be a correlation, because even if a movie is a flop plot wise but it still makes a big profit, Like Transformers: Revenge of the Fallen [1]. So, what variables do make a movie successful in terms of box office and ratings?

I looked online if anyone has already tried this and how accurate the predictions were. I found there are several papers who were able to use datasets on movie metadata to make predictions on movies ratings. But I found no software or online app that I could use to try it out. I found no source code or implemented algorithms on this either. Hence, I want to create a web application version of this.

I found a paper from Stanford University that would correlate the movie's budget, number of awards won by the cast, critic sores and ratings and could make a prediction accuracy of around 60% [2]. Using linear regression models with machine learning seemed to one way to predict movie success but not as accurate as other models, such as "Support Vector Machine" type of machine learning. Another paper I found, took into account more variables such as date of release of the movie and used multiple ratings from critics and audience as a measure of success, multiple awards won by actor and even how many times they were nominated. Also achieving similar results as the previous paper [3].

Figure 9:- Barplot and Scatter plot for year, month and day basis movie release data



I think it is possible to have a rough estimate of how successful based on planned variables. I would like to explore how well I could predict movie success using an artificial neural network and taking into account as many variables as possible. Gather data from multiple sources like Rotten Tomatoes and IMDb. IMDb even has a small section outlining the plot of the movie, maybe with some thematic analysis (also machine learning model) the plot of the movie could be taking into account when predicting success, no one has done plot line thematic analysis of movies and correlate with their success along all previously mentioned variables.

As a movie buff, I feel very passionate about this project and feel keen to explore the powerful abilities of deep machine learning to predict real life results.

References

[1] Transformers: Revenge of the Fallen earned gross of $836,303,693 with budget of 200 million, but 20% on Rotten Tomatoes - https://www.rottentomatoes.com/m/transformers_revenge_of_the_fallen
[2] A Predictor for Movie Success - http://cs229.stanford.edu/proj2013/EricsonGrodman-APredictorForMovieSuccess.pdf
[3] Movie Success Prediction using Data Mining - 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

Proposed Approach

I will use python for most things, from web scraping to hosting the web app and training the artificial neural network

To make prediction on movie success, I need to first acquire a lot of data about movies. I will download publicly available datasets from sites like kaggle.com to start off with a dataset. IMDb website also have some datasets on the movies. This will be my initial dataset to start off. I will use web scraping techniques to get any data I'm missing. I will use python library "Requests" to automate website request and "lxml" and "BeautifulSoul" python library to parse through the received HTML to find the data I need, i.e the budget of a movie shown on IMDb sites.
I will use python scripts to parse my data sets and clean up the data and format it into a .csv file I can then use to train the neural artificial network

To create, train and use an artificial neural network to make perfection I will use python library "TensorFlow" and "Keras". I will load in the .cvs file with the movie metadata, using python libraries like "Pandas" and "Numpy" to format the data to standardize/normalize the input date so it will be ready to train the artificial neural network model. I will use the TensorFlow to set up, compile and train the neural network model. I will try to predict the ratings for a movie given the input variables. I want to also predict the possible revenue the movie would make. I might need to make multiple artificial neural network to predict multiple things.

I will use python web frame called "Django" to make and host my web application. I plan to have a client-server architecture for this. The web app will be client side, getting the input from user, then on server side I will do the necessary processing with web scraping and making movie prediction with my machine learning model. I will have a user-friendly interface and will have form where the user can enter the details about the planned movie. To make the user interface I want to use JavaScript to make a responsive and interactive website. Django will then process the input, it will web scrape IMDB to get metadata needed, like check if the inputted actor has any new awards since the last time, I updated the database. I will have a database of the movies I collected during the dataset acquisition and check the database for static information like how successful where the movies that the inputted actors have acted in to determine how good the actor might be. After all the data needed is found and processed, it will get put into the neural network and after it receives the output, it will display the results to the user.

I want to use Feature Driven Development. Since it's a project that is managed only by myself, I think it is most appropriate as opposed to Agile or Test-Driven Development. Although Feature Driven Development is still very similar to Agile, in fact it takes all coding best praises and put them together into a cohesive whole.

This a fully software project, so there is no need for any additional hardware.

To evaluate the project, I will get people who are tech savvy and others who are not, to see how easy and comprehensible is to use the web application and user interface.
I will test the model is accurate using a test dataset I will make of previously made movies. Knowing the ratings and revenue of a previously made movie, I can use the actors and budget of that movie and see if my model predicts the success accurately. If not, I will fine tune the model to make better prediction. I will also write automated tests to test my software systems, like the web scraping systems, for example does the program find the number for the budget, if I have known budgets for movies, and the numbers I get from automated web scraping match, then I know my program works as should be.

FYP Project Proposal Form 2019/2020

Deliverables

Interim Report

A project dissertation

Front end: A web application that will take in user input to predict movie success with

Back end: Web hosted on the Django web framework, database of movies, an artificial neural network to predict movie success.

Python Scrips to clean datasets, web scrape for needed data online, creation and training of the neural network.

Scripts to test the model is accurate and automated tests to ensure the system works as intended.

Technical Requirements

Laptop

Website hosting site

Uses of programming language Python and its libraries: TensorFlow, Keras, Numpy, Pandas, Requests, BeautifulSoup, lmxl.

Web framework Django (python web framework)

FYP Project Proposal Form 2019/2020

Project Reviews – Please include reviews of two of LAST 2 years projects from either DT228, DT282 or DT211C.

Project 1
Title: Secure Document Sharing

Student: Owen Kane

Description (brief):

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.

FYP Project Proposal Form 2019/2020

Project 2
Title: Education Tool for Web-Based Vulnerabilities

Student: Cormac Kelly

Description (brief):

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.

Proposal Sign off:

Student Signature:	**Date:**
Lecturer Signature:	**Date:**