THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF INDUSTRIAL AND MANUFACTURING ENGINEERING

MOVIE ANALYTICS – PREDICTIVE MODELING FOR MOVIE SUCCESS

PRESTON ADRIAN SOEPRANOTO
FALL 2017

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Industrial Engineering
with honors in Industrial Engineering

Reviewed and approved* by the following:

Soundar Kumara
Allen E. Pearce/Allen M. Pearce Professor of Industrial Engineering
Thesis Supervisor

Catherine Harmonosky
Associate Professor of Industrial Engineering
Honors Adviser

* Signatures are on file in the Schreyer Honors College.

ABSTRACT

Predicting a movie's profitability or return on investment (ROI) is a complex problem, especially for investors looking to make a sizable investment with the prospect of getting even greater dividends. In order to predict the ROI of a movie, movie studios would have to select the right mix of factors that would accrue to the success of the movie. Some factors involved in a movie production include and not limited to the production budget, actors, directors, genre and rating. This research integrates the classic factors tied to movies, and developed new factors that were deemed to be crucial to a movie's success to predict the ROI. Factors in this study are divided into four groups, "star feature", "integrated feature", "descriptive feature" and "time-based feature". This thesis aims to build a predictive model by utilizing historical film data from a paid resource, nashinformationservices.com, and to predict a film's success by looking at its ROI. Only movies released between 2007 and 2016 were used. An ROI of a film was more indicative of investors' goals as opposed to its box office revenue. The predictive model will be based upon machine learning techniques, specifically supervised learning. We approached this problem by classifying movies into different classes of ROI, ranging from "low ROI" to "very good ROI", and the goal is for our model to accurately place a film in the right class. Attempting to achieve this feat would require us to first complete the data preparation stage. Data then was used to train the model using various learning algorithms and the model was tested on how well it did using a 10-fold cross validation. Applying this methodology, we were able to see which algorithm would perform the best for our model. We further went one step ahead and tried to predict the numerical value of a movie's ROI through a regression analysis. The prediction accuracy (based on ROI) was used as a benchmarking measure. The model with the highest

overall accuracy, area under the receiving operating curve (AUC), precision and recall was chosen for the final application. Based on these metrics, we found that out of all of the multiple learning algorithms, the “Random Forest” performed the best for our model. This analysis also gave a clear indication on the impact of the features selected and how they correlate with the ROI of a film. The correlation between the features and the ROI of the movie was measured using the Gini Index. According to the Gini Index, the attributes “Total ROI for Supporting Actor”, “Total ROI for Director”, and “Total ROI for Actor” make up the top 3 on the list of features starting from the first to the third.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

This chapter presents the motivation for this thesis. We analyze the programs and methods used to solve the problem. To close, we provide an organization of this thesis.

1.1 Motivation

The film entertainment industry (in specific Hollywood) is one of the most sizable industries in the world. In the year 2016 alone it has gathered an overwhelming global box office revenue, from theatres, of $38.3 Billion and it grew to about $41.2 Billion in 2017 (Statista.com, 2017a). According to statista.com, one of the largest online statistics and market research portal and used by large websites such as Business Insider, the revenue for the film industry is expected to grow to about $44 Billion by 2018 (Statista.com, 2017b). Movies released in the United States and Canada have an average investment of $100 million per movie, another report from statista.com states. In spite of the millions invested, there is no guarantee that the investments made would bear dividends, with “blockbuster” and “bust” movies seem to keep showing up every year. Aware of the barriers that this industry presents, movie producers and investors must allocate their capital optimally to gain returns or profits.

Achieving success in the film industry is no easy feat. This legitimate statement is valid, as only 36% of movies produced between 2000 and 2010 in the United States had earned a profit, according to Lash & Zhao (2016). It is therefore critical to make strategical investment

decisions so that movie studios do not fall under the majority of those movie studios who had a run of misfortune. Now a question arises, how are studios able to predict the outcome for their success? How would they be able to predict the number of moviegoers who watch their movies? To address this overarching problem, studios must first consider the important factors that accrue to the production process before its release. The real mission is to find out which of these factors will make an impact to a movie's success.

1.2 Problem Definition

The significant problem we are undertaking for this thesis is the prediction of the success of a movie. Many have many interpretations of success, the most common success metric for a movie is a box office revenue, which is the total amount of sales generated after the release of a movie. While this may be a good indicator of success, a box office revenue has not incorporated the costs of producing a movie. As an investor, one would be more concerned about the profits generated by a movie, as this has taken cost out of the revenue to get a net income. Going one more step further would be how relative the value of the profits is to an investment that was made. A $10 million net income would be a massive amount for a single person, but if $1 billion is invested in the movie, then that net income of $10 Million, would not look like such a large sum anymore. Therefore, we have selected to predict the return on investment (ROI) of a movie, this is essentially the percentage of profits over the total cost or production budget (amount investment), to get a suitable metric for the success of a movie.

Predicting the precise ROI of a movie may pose as a challenging problem, as no stakeholder would actually expect a precise ROI value for his or her movie. Typically, these

investors would have a range in mind of how much returns they are expecting to make. Given this consideration, we would then create ranges or classes and classify movies into a number of categories based on their ROI. This range can vary from for example, a “flop” movie to a “hit” movie. Having a continuous variable and then converting it to a finite number of classes is called discretization. Discretization takes advantage as discrete values are more relatable than continuous values, the data can be simplified using this method, and it allows multiple learning algorithms to operate more quickly and accurately.

We perform two kinds of classification, binary and multiclass. Binary classification will split the output variable (return on investment) into two classes. We have set a threshold value to split the data based on the information on the data distribution of our output variable. The multiclass classification will discretize the output variable into 4 different classes with four different ranges, based on different levels of the output variable, after observing our data distribution of our output variable. To solve the problem and get the model to accurately predict the movies to their respective classes, we will attempt to build a model based on machine learning techniques. These techniques would let the model train using some training data, and learn to correctly classify movies based on historical data. Next, we will test the model on some testing data, which will help determine the model’s performance using certain metrics of accuracy from the field of data analytics. In the next section, we will discuss a summary of the methodology adopted.

1.3 Methodology

Before we solve the problem we must first understand its objective and the items that would affect the solution of the problem. We consider as stated earlier, two variable types: input or independent (Predictors), and output or dependent (Response). Independent variables will affect the outcome of the dependent variables. From the series of independent variables, we would have to narrow them down. Our dataset provides a whole list of information including actor names, director names, genre, opening weekend theatres etc. Selecting the independent variables from the dataset to create the predictive model is the first step. Once this has been done, we looked at other external factors, outside of the dataset, that may provide a better insight to the outcome of the dependent variable and ultimately if it has a significant impact on the outcome. After finalizing the list of all the independent variables, we can then start to begin our feature engineering process. The predictive model will then run these features and learn to predict the outcomes based on the historical data we have. Finally, movies which are in contention to be produced can be incorporated in the model for prediction.

Using the data mining software called RapidMiner (2013), we are able to conduct the data preparation and then the modelling process. The software will allow us to read and write from an Excel file, clean and consolidate the data using filtering functions inscribed in the program, apply machine learning techniques for training and testing the data, and finally visualize performance vectors for predictive models. RapidMiner allows a GUI to craft and run analytical workflows or flowcharts, so there is no need for explicit programming to conduct our study. The program has a selection of learning algorithms to select from to solve classification or regression problems. From the list of algorithms, we will perform multiple analyses and ascertain which of them performs the best. We can determine the best performance based on accuracy

metrics that will be discussed in the succeeding chapters. After running the model, we will be able to find out the independent variables that have high correlation with the dependent variables.

1.4 Thesis Organization

Chapter 2 provides a summary of the movie success model we are exploring. We discuss the outlook of the problem and what sort of success metrics are we interested in. This chapter will present early research studies that have attempted to solve this problem, and interprets the approach they are taking to do so.

Chapter 3 will provide an in-depth analysis of the dataset. Split into two parts, the data description and the summary descriptive statistics. The first part will describe the various file types, listing the features in each file and its information, or metadata, while the second part shows us the final dataset that would be used for our model and some statistics that come with it. Our problem definition is outlined and in Chapter 4, and an exhaustive methodology to tackle the problem is described in Chapter 5. The framework of the model and the list of the features used and how we calculated them are defined in Chapter 4. Chapter 5 will exemplify all the quantitative performance metrics used for the results and the list of processes to return these metrics from the model.

The final two chapters 6 and 7 scopes the analysis and discussions of our dataset. Chapter 6 will assess the efficiency of our features with respect to the predictive model. Seeing how our features have fared to impact the output variable will allow us to evaluate the strengths and limitations of our model. Chapter 7 synthesizes the whole thesis with a conclusion and

summary of the qualitative results. At the end of the discussion, a set of potential ideas for future work are raised.

Chapter 2

Background Literature

2.1 Early models

Barry R. Litman (Litman 1983) pioneered one of the first box office prediction models. Litman identified the paramount problem of unpredictability in creating a profitable film within the motion picture industry. In a pursuit to work out this problem, Litman developed a multiple regression model based off film data from 1972 to 1978 and factors that he perceived were influential to a movie's financial success. According to Litman, three decision-making criteria contribute to a theatrical success of a film. They are the creative sphere, the scheduling and release pattern and the marketing effort. The first criteria, the creative sphere, consists of variables that pertain to the director, actors, film rating and production budget. Litman believed that the role of a director is important, the effect of a highly rated movie actor was declining, a larger production budget is correlated with a higher quality film and the film rating is a restrictive variable. Next, the scheduling and release pattern included the selection of the distributor, release pattern and release date. Having a larger distributor and releasing films at peak dates will usually return higher profits. Finally, the marketing effort scopes the amount of money put in advertising, number of critical reviews and award wins or nominations. Litman's model had ascertained that most genres, ratings, the role of a movie star and all peak periods, with the exception of Christmas, do not play a big role to a film's success. He deemed these parameters important in his model, starting with the highest: production cost, critical rating,

science fiction-horror genres, distributor, Christmas release and Academy awards. His model justified almost half of the variance in the dependent variable, which is the revenue, with a statistical fit of 0.485 (Litman 1983).

Five years later, Litman collaborated with Linda Kohl (Litman and Kohl, 1989) to conduct a similar research and develop on it, but this time it focused on movies that were released in the 1980's. This reciprocated research was conducted as there were trends such as technology and consumer options that had to be incorporated to the model. Consumer options or simply revenue streams included premium cable, home video cassettes and pay-per-view cable. These options meant the movie industry had to deal with more competition. In addition, parameters such as number of theatres, director power, market concentration and if a film was a sequel were used in the model. They found that horror films did not make a huge impact, as it was oversaturated. The effect of academy awards was trivial as film studios were pivoting towards the VCR market. This new model also made aware that the actor "power" is weighted more heavily (Litman and Kohl, 1989). The discrepancy between these two research efforts was evident due to the dynamic nature of the film industry, which poses the challenges in creating a box office prediction model. During that time, changes included the presence of cable television and VCRs. Nowadays; movies can be streamed online with products such as Netflix and other open source websites.

2.2 Contemporary Models

Since Litman's first research published in 1983, other studies have attempted to develop their own predictive models to forecast the financial success of movies. A study similar

to Litman, by Terry et al, done in 2005 leveraged multiple-regression models to predict the box office revenue generated from a sample of 505 movies. Parameters included the percentage of positive critical reviews, restrictive ratings, sequels and children and action films. Also incorporated were number of award nominations, theatres movies are released in and production budget. What made their study different was the inclusion of all films between 2001 and 2003 that were released to more than 25 theatres. The new model yielded a statistical fit of about 0.700. One key aspect in their research was the effect critical reviews had on box office values (Terry, Butler, & De'Armond, 2005)

2.3 Star Power

There is no question that a movie studio's biggest assets are its actors, but how much weight is actually given to them to determine their financial success? That was when Wallace, Seigerman & Holbrook (1993) raised the question, "How much is a movie star worth?" They took data from 1956 to 1988 that covers 1,687 entries of movies to conduct various stepwise regressions. Wallace et al found that indeed notable actors have significant influence in the success of a film. They also found that an actor's career progression would play a part in a film's income.

To tackle the underlying conundrum of how movie stars tie into a movie's success, we must first ascertain the definition of a star's power. Hiring a movie star is similar to making an investment on an NBA superstar, is it worth the money while probing for that success? Should the number of Oscars they have won define these actors? This inevitably may lead to the exclusion of stars like Brad Pitt and Harrison Ford. On the other hand, we could

potentially use an actor's gross salary, but that would also imply that any actor playing banal roles in blockbuster movies could become big stars. A new idea that had recently come up was to take advantage of rankings made by the Internet Movie Database (IMDB) to calculate an actor's "power". Nelson & Glotfelty (2012) brought up this idea. The average amount of user visits on the IMDB website is about 58 million per month, and the company is able to record user patterns and based off these patterns they can formulate a ranking for people and movies from the volume of pages visited. Nelson et al (2012) firmly believe that the rankings made are correlated with the worth of an actor. The researchers found that conducting a regression analysis with the variable, "actor worth", they noticed that substituting an actor that has an average "worth" with that of a high "worth" increases the box office by $5,225,365 when the budget and theatres were predetermined and by $28,011,775 when they were not. A second analysis was done with a substitution of three actors, this time the revenue was incremented by $49,318,858 when the budget and theatres were predetermined and by $79,501,904 when they were not.

2.4 Initial Predictions

Previous models that were described in this literature were able to clearly identify dependent variables, or the outputs, such as the opening week box office revenue or the total domestic box office revenue by leveraging various independent variables, or inputs. The models mentioned before however, require input parameters that have not been predetermined until after the movie is released or produced. Thus, producers and investors will have to hedge the risks of investing in a movie before confirming whether or not it will be a success. That was when research studies were initiated to predict the box office revenue and profitability in the initial

phases of a film's production prior to big investments being locked in. One of the first predictions to estimate the box office revenue of movies, prior to their release that was done by Sharda and Delen (2006). They suggested the utilization of neural networks. An artificial neural network operates in a similar fashion to our brain's neural network, which contains a set of algorithms to pick out relationships and trends in a set of complex data that are unnoticeable by the human brain or simple computer algorithms. Sharda and Delen went about this problem using classification techniques as opposed to defining it with forecasting. According to them, movies can be split into nine categories beginning with "flop" movies and reaching heights that are typically called "blockbusters" based on the monetary values of the box offices. Some of the input parameters incorporated were the star value, genre, film rating, competition, special effects, sequel and number of theatres. The neural network performed fairly, it being able to predict the precise film category 36.9% of the time and 75.2% of the time it was able to predict one category away from its exact one.

Sharda and Delen as well as other scholars did further development of the research. Ghiassi, Lio & Moon (2014) modified the first prediction model by annexing more factors and utilizing a dynamic artificial neural network that produced improved outcomes. However, they also took away the genre, special effects and star power out of the list of variables. The supplemental factors were production budget, runtime, pre-release advertising expenses and seasonality variables. The modifications yielded dividends as the prediction accuracy went up to 94.1%.

2.5 Late Prediction by Schreyer Scholar

David Wagura, a Pennsylvania State University Schreyer Honors College Scholar, previously did a study leveraging the same source of the dataset this thesis has adopted (2016). In his research he developed a recommendation model that would suggest the best actors to hire for a movie based on existing features such as current leading actor. David built his model using a clustering technique that separates different groups of movies from the dataset. The clusters were differentiated by the financial success of movies. David made a variable called *earnings per theatrical engagements* to determine his metric of success for movies and actors. His model would determine the rank of best actors using this metric. An initial actor had to be initialized for his model to run. If at least one actor has been selected, then a network can be built to find the best combination of actors. Using data from the clusters, "a search method could be used where the database is queried to find the best actor" (2016). The paper visualizes how a network looks for actors that would work best with the initial actor.

2.5 Assessment

Achieving a satisfactory standard for success in predicting a box office revenue has always been a barrier for most researchers and scholars. This paradox is often associated with models formulated with variables that are not retrieved post-release of the movie. One way to mitigate this barrier would be to design neural networks that will make predictions that are very accurate with the data only accessible in the initial phases of production. Most researches associated with movie success predictions have the potential to become breakthroughs but the lack of data applicable has obstructed it to achieve that status. Most datasets required to develop

the models hold ownerships to film studios, which seemingly never want to release it to the public. To account for the missing data, most studies size the data based on their theories or hypotheses. For that reason, studios should start collaborating with the researchers in this field to create models with more concrete data to make better investment decisions for their films.

Chapter 3

Data Description

3.1 Data Description

The main source of data came from Nash Information Services who happen to be the pioneers of two websites: www.opusdata.com, a source that provides users to retrieve movie information for the purpose of this thesis, and the second is www.the-numbers.com, where information about movie finances could be found. The data was recorded starting from October 17, 1997 and it measured economic information on films. The company has been actively collecting film data and kept it in their database with the intent to provide services for consumers and institutions. Inherently one form of service administered by OpusData is an academic extract for academic research functions. The data used from this extract is used for this work and was downloaded on November 2015 and updated on February 2016.

The extract comprises of exactly 10,161 movies that were released or re-released since 1997 that Nash Information services has domestic revenue numbers. Movies that were released after 2007 have exhaustive classification data, as for movies released before 2007, 75% of them have the complete classification data. The scope of the classification data includes factors such as genre, creative type, production budget, box office and a plethora of other attributes. Only the films produced post 2011 have all actors and technical staff including an above the line roles recorded. Films produced pre 2011 mostly have actor and technical staff

recorded. The daily box office data revenue was traced starting 2005 for all movies that reported it. That information was only available for the top 10 films from 1997 and 2005. The inclusion of the international box office totals started after 2000 for most films with the exception of some independent films. Additionally, tracking of DVD sales and Blu-ray sales began in 2006 and 2009 respectively.

Table 3.1. Content of Raw Data and Number of Entries

File Number	Contents	Number of Entries
1	Summary of each movie	10,161
2	Actors Roles	97,619
3	Movie Keywords	23,698
4	Language movie is spoken in	5,291
5	Lists of production companies	10,600
6	Lists of production countries	7,751
7	MPAA Ratings	9,481
8	Theatrical release information	16,259
9	Technical credits	91,323
10	Video release information	4,178
11	Daily box office	173,846
12	Weekend box office	105,067
13	Weekly box office	101,998
14	Weekend International box office	22,557

Collectively 9,924 films had information on the domestic box office total in this data extract. The films had a unique identifier labelled as “odid”, where its name is tagged under “display_name”. The whole data set is categorized into 16 different .csv files that had the primary key “odid” or “display_name” as mentioned earlier. While movies in this dataset dates back to 1925, we are narrowing our research to movies released during the 10-year period of 2007 to 2016. Movies in this period are more up-to-date and mirror the modern state of the industry, and the amount of time that transpired since the release of movies were noticeable for revenue data to be accurately updated. Given these considerations, the first working dataset for

our experiments consisted of 1,578 movies. We will also not be using the information on DVD sales and Blu-ray sales, rather focusing on money generated from solely the theatres, which will lead us to only, contain 14 different .csv files. The breakdown of the contents and number of entries are listed in Table 3.1.

The first file lists all the films and various attributes that represent these films. They include quantitative and qualitative values such as; production year, running time, sequel, opening weekend revenue, opening weekend theatres, maximum theatres and theatrical engagements. Attributes describing the movies' creative type, source, production method and genre are added. Lastly, the data scopes information on the production budget, domestic box office, international box office and inflation adjusted domestic box office. The attributes aforementioned are not complete however. Table 3.2 below shows the number of entries possessing data for the numerous attributes.

Table 3.2. Number of Entries with specific attributes

Movies that have:	**Number of Entries**	**Percentage of Overall**
Domestic Box Office Total	10,161	100%
Comprehensive movie summary (genre, production method, source, creative type, etc.)	7,828	77%
Partially complete movie summary (genre, production method, source, creative type, etc.)	8,755	86%
Production Budgets	3,690	36%
Running Time	4,018	40%
International box office total	4,320	43%
International releases	721	7%

File 2 would encompass data on acting credits. The films are recorded on multiple accounts, one per actor. The actors in each film are arranged by billing, which is exactly how they are shown in the movie credits. For every actor, the character name, as well as the type of role he or she held, leading or supporting, are listed here. File 3 contains keywords connected to each movie. Keywords include depictions such as Romance, War, Robots, and other descriptions. File 4 has all the languages that were spoken in the films if the data is present. File 5 provides an agenda of the production companies that are in relation with the movie. The sixth file gives data on which countries the film was made in. File number 7 contains the film ratings, for example R, PG-13 and G. The eighth file presents the movies' release dates, territory released in and the distributor name. There are cases where certain films have more than one release date because of rereleases. File 9 would reveal technical credits that highlight names of the crewmembers who worked on the movies. These names have roles associated with them, ranging from director to stunt coordinator. The tenth file has the release dates and names of the video.

File 11 provides the daily box office values. There is also data on the total domestic box office and ranking at that time. This file contains the previous ranking, number of theatres, number of tickets sold per day, total tickets sold and the days in release as well. File 12 displays all the attributes in File 11 during the entire weekend only this time. File 13 again shows all the attributes in File 11, but in this case, it was for an entire week. Finally, File 14 has the weekend international box office for newer movies. This file will involve attributes such as the territory of release, rankings, date listed, currency used and local revenue. It will also include the number of theatres, tickets sold, days in release and lastly total revenues internationally and domestically (in USD).

3.2 Summary Descriptive Statistics

Before we begin our analysis on the data, it is imperative that we know the basic summary statistics of the dataset. Table 3.3 below visualizes the descriptive summary for some key attributes.

Table 3.3. Dataset and Basic Statistics

Attribute	Average	Standard Deviation	Minimum	Maximum	Number of Entries
Production Budget ($)	40,892,181	51,037,788	587,000	425,000,000	536
Domestic box office total ($)	50,551,160	75,151,761	388,000	760,507,625	536
Inflation adjusted domestic box office ($)	54,442,250	80,676,895	0	826,198,130	536
International box office ($)	71,165,582	140,589,976	0	2,023,411,357	536
Opening weekend revenue ($)	15,210,280	24,132,653	0	208,806,270	536
Opening weekend theatres	1,800	1,494	0	4,468	536
Maximum Theatres	1,961	1,410	0	4,468	536
Production Year	2007	2.3437	2007	2016	536
Sequels	0.1112	0.3144	0	1	536

Looking at the table, one thing to first point out is the standard deviation of the production budget. At a glance, the standard deviation of the budget is higher than that of the average. However, we cannot have a below zero budget, so this nominal value can be justified by the minimum and maximum values of the production budget. If we observe these two values, and look at the largely scaled difference, it is no surprise that the standard deviation is quite significantly large. High budget movies would carry high expenses paid out to actors, crewmembers, advertising, equipment, etc. The same trend is also seen for the domestic box

office where the difference between their average and standard deviation is even greater. Again, box office will never be below zero, and the wide range between their maximum and minimum values justify the great standard deviation. Monetary variables such as the inflation adjusted domestic box office, international box office and opening weekend revenue will all follow the same reasoning. These variables carry a vast range that yield standard deviations to more than the average, even if these parameters cannot be negative.

Our maximum theatres and opening weekend theatres are close in all of their values. This trend is because most theatre distributors would tend to maximize the number of theatres during the opening week of the movie to build momentum on the "hype" surrounding the movie through the advertising, reviews and just word of mouth to entice the audience's curiosity. In our dataset, we have a higher average for maximum theatres than that of the opening weekend theatres. It can be pointed out that studios would actually increase the number of theatres showing a film after the first week of its release. When a movie yields low opening weekend revenues, the number of theatres would usually be reduced in the succeeding weeks. In contrary, when a movie yields a higher than expected revenue, film studios will increase the number of theatres to raise more revenue and utility. The average of the production years will return a value between the earliest and latest movies to be released and it will show if it leans towards the present or the past. The average production year is 2010, and that is about right in between our range of 2007 and 2016.

After filtering, the data set and we decided that our thesis would only work with entries that have all the complete full features. There are thousands of entries on the primary data set, however only a limited amount of entries would comprise of all the features representing the core of the model. Features such as actor and director names would have an impact on the

success of a film, if we were to run our model based off the raw data set we would miss a handful of key features that are detrimental to the model. Thus, we have justified the current amount of entries we are working with for the model, which is 536.

Chapter 4

Framework and Feature Engineering

4.1 Framework

We will start by first introducing the framework for our movie prediction model, which is shown in Figure 1. The initial stage would be the data collection; this is the basis for our model as we are working with historical movie data. As previously mentioned in Chapter 3, our primary source of data comes from nashinformationservices.com. The full data extract provides metadata on films (genre, production method, cast and crew, production companies, production countries, etc.) and aggregate financials (total domestic and international box office, production budget, opening weekend revenue, etc.). In addition, it offers time-series data for daily, weekend and weekly box office. Our data is stored in the form of .csv files. There are multiple sheets in this extract as one movie can have multiple theatrical releases, MPAA ratings, production companies, etc.

Stage 2 involves the data preparation; this entails cleaning, consolidation, modification of the raw data set into a new primary dataset. This process is critical to the whole framework, as the data preparation normally takes a majority of the time taken to conduct the whole research. We need to ensure that the obtained features and texts from the 14 .csv files are put in a one single format, and that duplicate entries will not exist in the database. Consistency in the format means taking away characters such as “-”. Doing so will allow the matching of the same movies that may have been written differently in their titles coming from different data

sources. In this stage, we have utilized a data analytics software platform called RapidMiner. This tool was well suited to fit our data preparation and eventual machine learning, and predictive modelling needs.

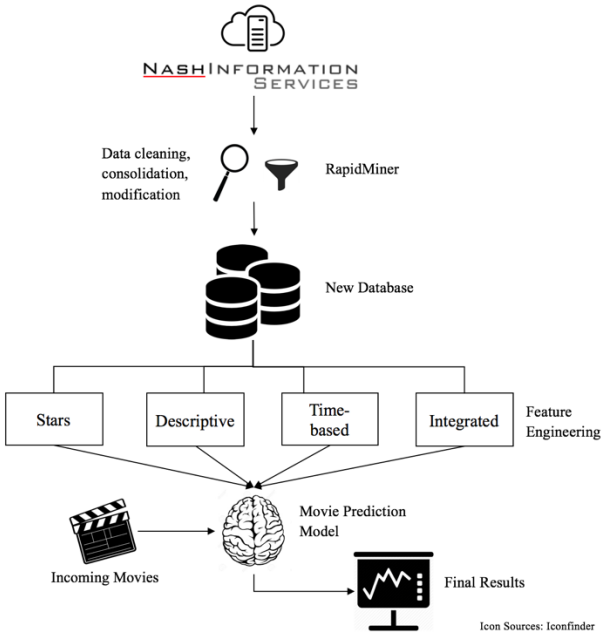


Figure 1: Movie Prediction Framework

The third stage, which will go into what we like to call “feature engineering”. Feature Engineering will introduce new parameters or features from the data we collected and will play a critical role in training our model. The features invented in this experiment are divided into four main categories: stars, descriptive, time-based and integrated features. We will discuss in detail what these features are and the logic behind their selection in the upcoming paragraphs.

Our final step is then putting the predictive model in action. Incorporating our newfound set of features, the predictive model will be trained in this stage. Those who will be using our model will be able to input their choice of features within the scope of the model and

the prediction method that would yield the best results in terms of the accuracy of our framework. In the next chapter, we will give a comprehensive walkthrough on our methodology and the experimental results obtained.

A stakeholder will be able to employ our movie prediction model by loading the data available on the movie being produced onto the model. Our model would then conclude the amount of return yielded because of the investment and thus determining if it would be profitable enough for the stakeholders.

4.2 Data Preparation

The data preparation is a big step of any predictive modelling to begin with, this usually takes a majority of the labor in research. The data needs to be cleaned, modified and consolidated before we can begin any sort of modelling, and for these purposes, we have heavily utilized a data science software platform called RapidMiner. To reiterate, we would be only using the movies spanning the years 2007 to 2016. Our first step would be to select the features we deem are going to affect our response variable, return on investment (ROI), from the .csv files provided to us from nashinformationservices.com. Once we have chosen the features, we were able to remove the columns or features that were unnecessary and come up with a new data set that has covered features that were available to us. Since certain features existed in various .csv files, we would have to consolidate files all these into one working data set. The following step would be to clean the data and ensure that we have a consistent format across all the different entries, this means removing some characters such as “-”, to help with the filtering process. Next, the data sets of interest represented in the diversified .csv files will be

consolidated into one major data set. Finally, a few adjustments were made to optimize our data set for the predictive model. Data preparation is shown in Figure 2.

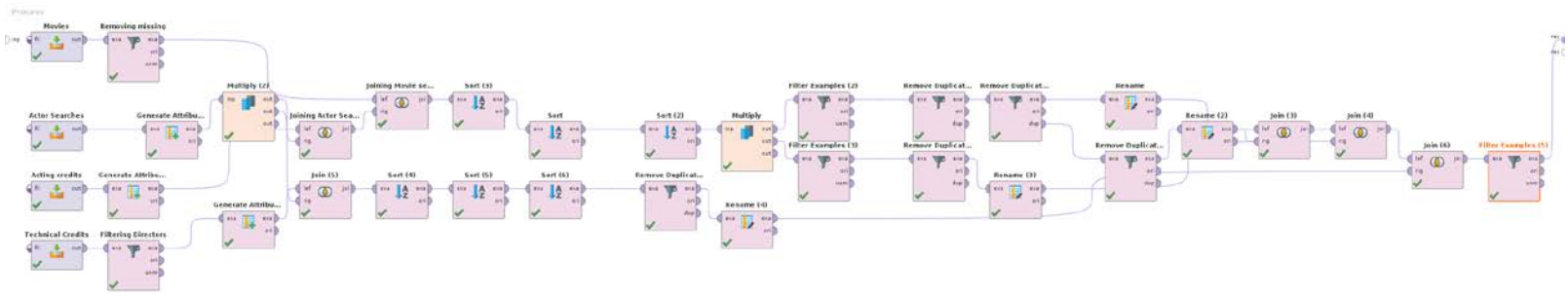


Figure 2. Data Preparation Flowchart

In addition to the raw data set prepared for us, we had supplementary data that we believe would tremendously influence the movie prediction model. This new data would be the amount of google searches for a single actor and director. The reason behind this extraction of annex information is discussed in a later section in the paper. To obtain the amount of google searches for a single actor or director, we explored GoogleTrends to aid us with this information. However, GoogleTrends was not able to provide information beyond a certain amount of entries. After further research, a tool that we were able to leverage to help us get actor and director search results was in fact still a google product, which in this case was GoogleAdwords. A component of GoogleAdwords, named Keyword Planner, allowed us to input a list of actor and director names in a .csv format and upload it on to the site, see Figure 3. Upon completion, an output of the number of search results per actor or director was generated, as shown in Figure 4. The number of search results we are interested here is the monthly average search results spanning from September 2013 and September 2017. The monthly average search for a person is a good indicator to comprehend how popular an actor is now.

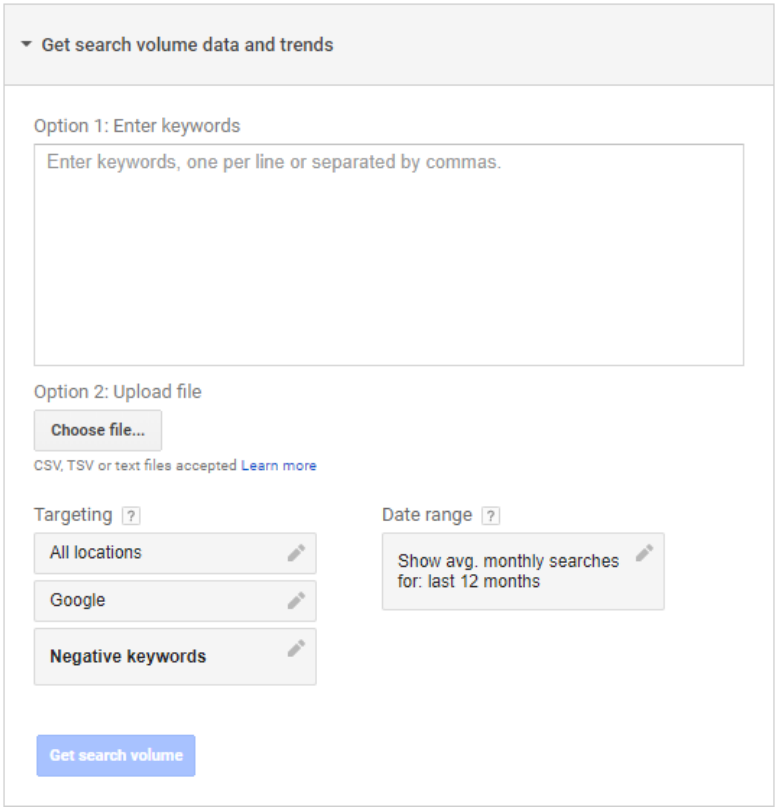


Figure 3. Google AdWords Keyword Input

4.3 Feature Engineering

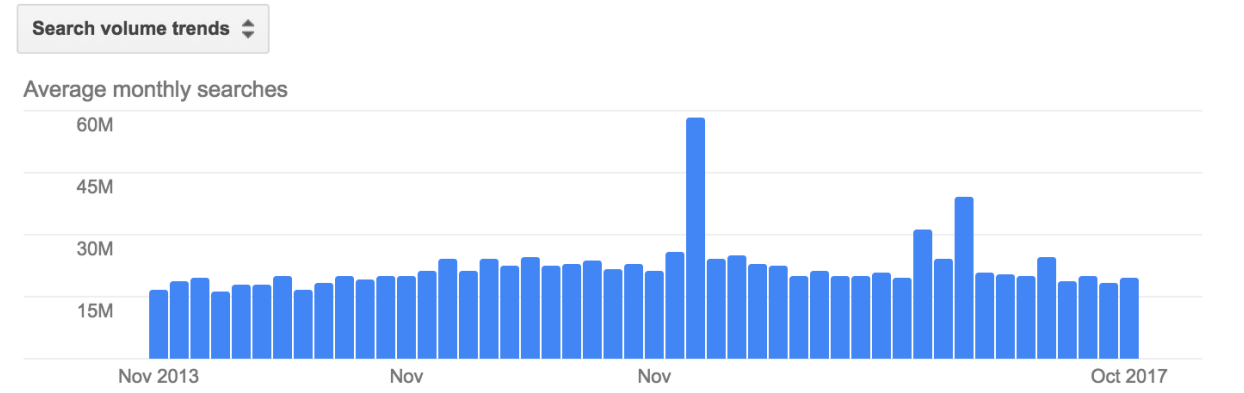
Based on the data set that was accessible for us fromnashinformationservices.com, we drew four different categories of features. These four features were the “stars” features, “descriptive” features, “time-based” features and “integrated” features that included extra modifications to our “star” features.

4.3.1 Star Features

What is the movie industry’s biggest assets? It is their people. People who create movies, the successful and noteworthy actors and directors such as Leonardo DiCaprio, Emma

Watson and Quentin Tarantino who are so ever-present in the entertainment world we are living in today. Actors and directors are able to have that “pull effect” on the audience as some of them may prove to be more popular with the masses. Every producer wishes to have the largest amount of highly rated actors and directors involved in their movie, but these producers will always take the cost of acquisition into account and the truth is highly rated actors don’t come at a bargain. Our aim is to predict the return on investment, and that is why our star features for a movie is heavily based on the actors’ and directors’ average return on investment for each movie. It is crucial that we find cast members who fit into the bill of the movie and can reap the potential profit.

We believe that by creating a feature, determining the so-called “star power” of an actor or director, that combines several factors into one consolidated variable, we can easily visualize and rank the cast members based on a universal metric.



Ad group ideas | Keyword ideas

Keyword (by relevance)	Avg. monthly searches
abdul sattar edhi	1,830,000
aishwarya rai	1,220,000
adam sandler	1,220,000
alan thicke	1,000,000

Figure 4. GoogleAdwords Keywords Volume Trends

I. **Actor Star Power:**

This feature has been the foundation of many different research papers; however, each paper has their own definition for a star power. Since we have a cast of leading and supporting actors, this measure will be incorporated for both types of actors. An individual actor's potential to produce the highest profits or his/her "star power", is going to be based on three different factors:

1. **Average Return on Investment** of an actor is how much profit an actor has made on average for every movie he or she has acted in before the movie about to be released. Each actor's Average ROI (Return on Investment) will be calculated by taking the sum of the profits for all the movies he or she has acted in and dividing that by the number of movies he or she has acted in. This factor was weighted in as 65% of the value of a star power due to the high correlation between this factor and the output variable, which is the return on investment of a movie. It has proven to be a good feature to incorporate for numerous iterations of the predictive analysis.

2. **Popularity** of an actor reflects how much an actor is well known or is a fan favorite in the entertainment world and social media. This determines the star power, as we would be able to gauge how the audience will perceive the performance of the upcoming movie. We obtained the popularity of an actor by ascertaining the average monthly searches for their names on google. The average monthly searches for each actor is collected beginning from September 2013 up until September 2017. This factor was weighted in as 20% of the value of a star power. Because of the dynamic change in trends in the entertainment world, we have to keep up with which

actors are gaining traction with the audience. For that reason, we will keep this as a significant factor to the star power.

3. **Number of Movies acted in** represents how much experience an actor has in the film business. This is calculated by the total number of movies the actor has acted in over the course of his or tenure, within the scope of the data set. This factor was weighted in as 15% of the value of a star power. We believe that that some form of experience will help in the performance of the actor in upcoming movies.

II. **Director Star Power:**

A director's potential to produce the highest profits, just like that of an actor's, and it is also weighted the same as way as the actor star power. It is based on three different factors:

1. **Average Return on Investment** of a director denotes his or her previous success. This is how much profit a director has made on average for every movie he or she has directed in before the upcoming movie. Calculate by taking the sum of the profits for all the movies he or she has directed in and dividing that by the number of movies he or she has directed in. This was weighted as 65% of the director power.

2. **Popularity** of a director, similar to an actor's, tells us how much he or she is well-known or a favorite in the entertainment world and social media. We obtained the popularity of a director from getting data on the average monthly searches for their names on google. The average

monthly searches for each director is collected from the beginning of September 2013 up until September 2017. This was weighted in as 25% of the director power.

3. **Number of Movies directed in** tells us roughly, how much experience an actor has in the film business. This is simply the total number of movies the actor has acted in over the course of his or tenure, within the scope of the data set. This was weighted in as 10% of the director power.

4.3.2 Integrated Features

The integrated features are derived from our star features. Star power features that we have included in our model above represents the success factor of the actors and directors. Aside from having actors and directors with a high star power, we should consider having the right mix of actors and directors. A truly great product may come out of this consideration, and that is synergy. In order to determine if such synergy exists, we would have to look at past collaborations between an actor and director. This integrated feature would indicate if the first leading actor has had a collaboration with a director and will be represented as binary. In our model, the synergy factor would look at two different categories: first leading actor's name and director's name. The variable would check whether the two have worked in the past. If they have, the variable would output as "1", else as "0".

4.3.3 Descriptive Features

Another aspect to consider when determining a movie's success is the general description of the movie. The audience normally knows this information prior to a movie's

release coming from its advertising efforts such as movie previews. These descriptive features will include sub-features like:

1. **Genre** of a movie, these are often classes or categories such as action, horror, and science fiction.
2. **Rating** of a movie is going to be whether it is all ages friendly like a G rating, requires adult supervision, PG-13, or maybe a more restricted one like R.
3. **Creative Type** is going to be more specific than the genre of a movie, some examples are, superhero, fantasy, kids' fiction, factual etc.
4. **Production Method** would be attributes such as animation or live action.
5. **Sequel**: If a movie has decided to continue the earlier storyline.

4.3.4 Time-Based Features

The movie industry has to account for seasonality as the market always spikes and slumps in different periods. This will play a definitive role to the success of its release. Some time-based features that we have included in our model are as follows:

1. **Release Dates** has information about the time of release of a movie, this is not limited to just the date but this piece of data will tell us the season of the year and if the release date is in line with a holiday or not. There are tradeoffs to summer or holiday releases despite a higher likelihood of getting more sales. Releases during these dates will demand a higher production

budget to cover for higher advertising expenses due to high competition. The release date is not going to be precise, but it does give an estimate for the date ranges to release it in.

2. **Opening Weekend Theatres** for a movie is decided based on the agreement between the movie producers and the theatre company releasing their films. Most of the time the theatre companies determine the amount, as they may want to be cautious when showing a movie that may be unheard of before. However, in certain cases where it is a joint agreement between the two parties the amount of opening weekend theatres may affect the potential sales of a movie if it has reached a higher level of audience than expected.

3. **Maximum theatres** is simply the maximum amount of theatres the movie is released in at a single point in time. This may be adjusted based on if the movie has still received more marginal benefit than marginal cost during the course of its release. It is as follows:

$$Maximum\ Theatres = Max\ (number\ of\ theatres\ on\ opening\ week, \dots, number\ of\ theatres\ on\ last\ wee$$

4. **Theatrical Engagements** is the running sum of the number of theatres that that film screened per week. To illustrate, let us say the movie Spiderman is shown in 200 theatres this week then 150 the next week, the theatrical engagements would be 350. It can be equated by:

$$Theatrical\ Engagements = \sum (number\ of\ theatres\ on\ opening\ weekend,\ number\ of\ theatres\ on\ second\ week \dots,\ number\ of\ theatres\ on\ last\ week)$$

Chapter 5

Methodology

5.1 Early Models

The raw data had 10,161 entries as mentioned in chapter 3, after narrowing down the years involved we arrived at 5,311 entries. We then cleaned the data to incorporate entries that have completed values in every feature except the actors and directors, coming down to 1,759 entries. Using this new dataset, we created a correlation matrix for our preliminary analysis. These results turned out to be poor, as a majority of the values in the director column was missing.

Table 4. Correlation Matrix for Preliminary Analysis

Input variables \ Output variable	**Return on investment (ROI)**
Production Year	0.063
Sequel	0.007
Opening Weekend Theatres	0.050
Maximum Theatres	0.065
Theatrical Engagements	0.058
Creative Type	-0.091
Source	0.013
Production Method	-0.009
Genre	-0.0071
Production Budget	-0.056
Box office revenue	0.047
Opening weekend theatres	0.045

A correlation matrix, shown in table 4, tells us the correlation of the attributes relative to our return on investment. The absolute values are all below 0.1, telling us that these

attributes do not have much effect on the outcome of ROI based on this dataset. Choosing not to ignore the missing values in the dataset entries, we had to further filter the dataset and made a list of fully complete entries. Table 3.3 has shown statistics of the final dataset and thus the final number of entries, which is at 536.

5.2 Measure of Success

To determine the success of a movie, we decided to look at its profitability. Since we have multiple revenue streams and production budget data for every movie in the dataset, we will be able to compute the amount of profits (i.e., total revenue stream – production budget) for every movie. The monetary value of profits may seem to be a reasonable indicator of the success of a movie, however, when looking at individual profits we would still have to consider the amount invested in the movie. A profit of $50,000 may look good on the balance sheet. However, given a production budget of $2 Million, that profit does not look that much lucrative anymore. Therefore, we decided to proceed with the return on investment (ROI) as our success metric or output. ROI is given as:

$$ROI = \frac{Total\ Revenue - Production\ Budget}{Production\ Budget}$$

ROI in Figure 5, shows a right skewed distribution. 34% of movies are close to the breakeven point (profits are zero), but there is a difference of 54% of movies from the sample that are profitable compared to those that make a loss. Of the list of movies, 24% of them have returned a loss. This is still a considerable number when stakeholders are trying to enter the movie industry, adding more reasons for the need for predictive models for movie success.

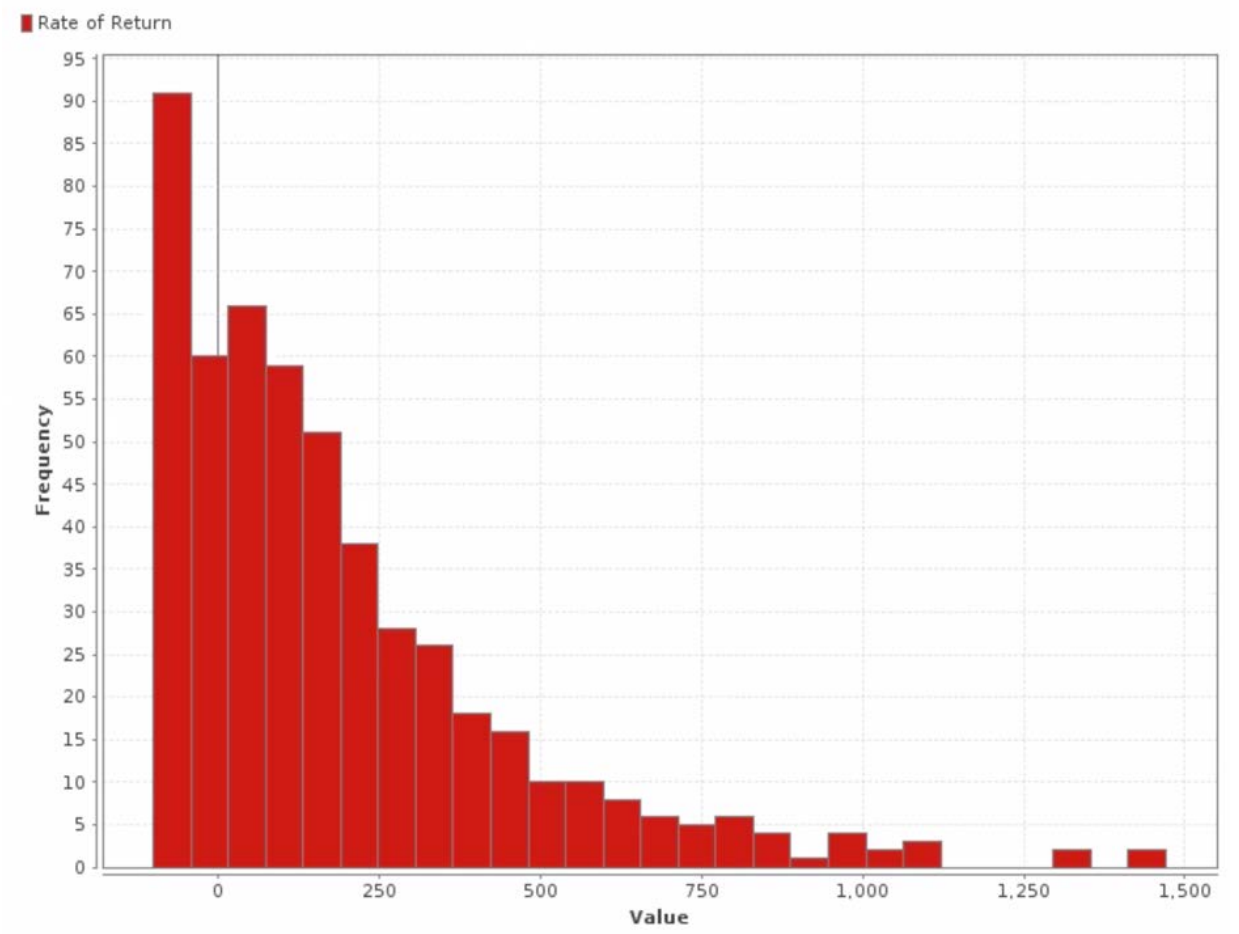


Figure 5. Distribution of ROI (in %) for films in our dataset

Our data shows that the ROI is not affected by box office revenue. Returning a meager correlation coefficient of 0.046 between the ROI and the box office revenue. By this fact, we can develop a hypothesis that a high box office revenue may not always bear a high ROI.

5.3 Return on Investment Classification

Our prediction model for movie success will be studied as a problem of classification. These data points (i.e., the movies) will be discretized, where they are classified

into multiple categories or in statistical terms, bins. The movie industry does not have a predetermined ideal ROI, but of course any investor would want to reap massive dividends from a blockbuster movie, especially if millions of dollars' have been devoted to the movie along with uncertainty. Based on these observations, we adopted both binary and multi-class predictions to interpret the decision threshold or ranges between movies that are profitable enough or not.

An array of algorithms was tested for both the classification methodologies; some of the algorithms we got involved with were Deep Learning, Decision Tree, Random Forrest, Naïve Bayes, and Support Vector Machines. A 10-fold cross-validation was used to get the experimental results here, 90% would represent the training data and 10% would be the testing data. Using 10 folds for cross-validation will set aside 10 subsamples, and from those 1 subsample will be kept as the testing data for the model, the remaining 9 will then be used as training data. The cross-validation process will be iterated 10 times (i.e. the folds), and every 10 subsamples will be used exactly once as the testing data. Cross-validation allows all the observations will undergo training and testing, and each observation would be used for testing once. 10 folds with a 90%-10% split for training and testing will assure that the variance of the results will be reduced as we are averaging over 10 different subsamples, by doing so they performance measure is less reliant to the splitting of the data. Given that we have a relatively low sample size this number of folds would fit suitably. We measured their performances and then selected the best one based on four metrics:

I. **Classification Accuracy**: It is the percentage of accurately predicted instances out of the total predicted instances. This metric is a the most commonly used one.

$$Accuracy = \frac{Correctly\ Predicted\ Instances}{Total\ Predicted\ Instances}$$

II. **The Area under the Receiver Operator Characteristic curve (AUC)**: A Receiver Operator Characteristic curve lays out an XY graph, where the true positive rate being the Y-axis and the false positive rate the X-axis. True positive rate is the ratio of true positives to all positives. False positive rate is the ratio of false positives to all negatives.

$$True\ Positive\ Rate\ (TPR)\ = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$False\ Positive\ Rate\ (FPR) = \frac{False\ Positive}{Flase\ Positive + True\ Negative}$$

The closer the AUC is to 1, the closer it is to reaching a perfect classification. On the other hand, if it is closer to 0.5, then it is leaning towards a random guess. A weighted average of the AUCs will be taken for multi-class classifications. Weighted average of the AUC calculates each class' AUC, and then weighting each of them depending on the number of occurrences that fall under each of the classes proportionate to the overall number of occurrences. This performance metric measures how the TPR and FPR trade off. An AUC will typically be able to handle cases where we have a very skewed sample distribution. A sample of an AUC curve is shown in Figure 6.

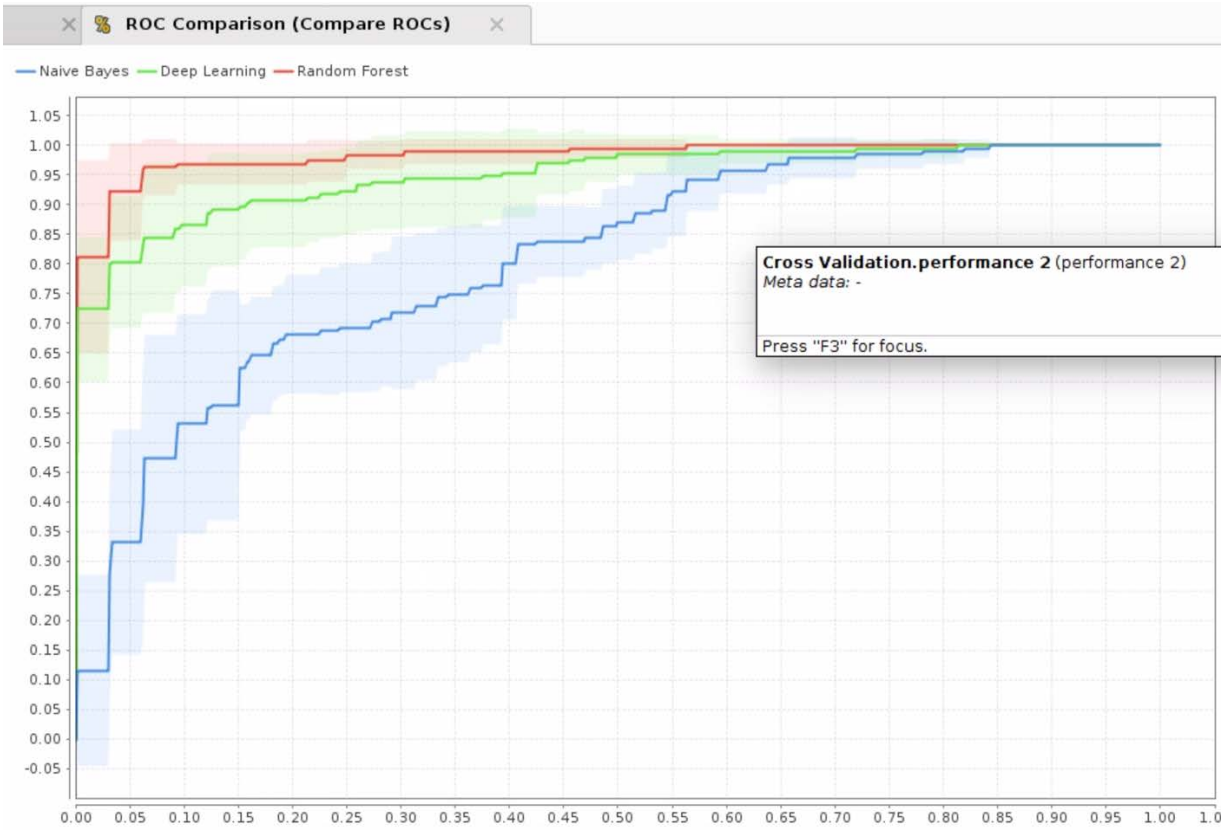


Figure 6. AUC Example

III. **Precision (for the positive class)**: It is the sum of occurrences classified to be positive, and are indeed successful, based on our decision boundaries, (i.e. true positive) over the sum of occurrences classified as being successful (i.e. true positive and false positive).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

IV. **Recall (for the positive class)**: It is the sum of occurrences classified to be positive and are indeed successful (i.e. true positive), over the sum of occurrences that are actually successful (i.e. true positive and false negative).

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

5.3.1 Comparisons

Together with analyzing the algorithm with the best performance, we analyzed if features we proposed in this thesis contributed to better prediction. Features we proposed were previously mentioned such as, “Star”, “Integrated”, “Descriptive” and “Time-Based”.

5.3.2 Binary Classification

This form of classification will take a movie, classify it into one of two classes: “profitable enough” or “not profitable enough” movie. Following that, we will determine after an evaluation of two decision ranges, then calculates these ranges. The two ranges will guarantee if a movie is considered profitable enough based on the threshold value. A total amount of two decision boundaries is analyzed here:

1. **Decision range number 1** observes that a film is thought to be profitable enough if its ROI value is ≥ 50%. This threshold was decided as we found that this value would be a safe option for most investors to be content with. Table 5 shows the performance of the best three algorithms, with the Random Forest Classifier generating the highest accuracy, AUC, and Precision, and the Decision Tree classifier is the highest in Recall as well as accuracy. Generally, the best performing algorithms would hit an accuracy and AUC of 0.900 or above. This decision range has shown to produce good performances.

Table 5. Best 3 prediction results of the binary classification model (Range 1)

Classifier	Random Forest	Decision Tree	Deep Learning
Accuracy	0.946	0.946	0.886
AUC	0.979	0.949	0.942
Precision	0.943	0.931	0.860
Recall	0.906	0.916	0.833

2. **Decision range number 2** observes that a film is profitable enough when the ROI value is ≥ 150%. This threshold ensures that 46% of the movies were deemed profitable enough, this improvement increases the standard for a profitable enough movie. Table 6 shows the performance of the best three algorithms, with the Random Forest still generating the highest accuracy, and also precision, but the deep learning model has prevailed when producing the highest AUC, and recall. Compared to the previous decision range, the second decision range has returned higher results for the deep learning model overall. While the Random Forest and the Decision Tree returned an accuracy, AUC, and precision of over 0.9, these algorithms dropped slightly in performance.

Table 6. Best 3 prediction results of the binary classification model (Range 2)

Classifier	Random Forest	Decision Tree	Deep Learning
Accuracy	0.921	0.909	0.905
AUC	0.956	0.915	0.965
Precision	0.919	0.914	0.862
Recall	0.894	0.895	0.928

5.3.3 Multiclass Classification

This classification now creates 4 categories or classes, which are “low ROI”, “good enough ROI”, “good ROI” and “very good ROI”. These give investors a more thorough

understanding of where their movies will lie depending on ROI. Our learning algorithms or classifiers does not incorporate any exact true positive values or false positive values. For example, when our model predicts a “good ROI” movie as a “good enough ROI” as opposed to predicting it to be a “low ROI” we consider the severity of the two occasions to be different. Logically they are both wrong predictions, but there is still a level of correctness. Thus we have used a rationalizing technique to overcome this. In a multiclass classification, we have certain expected predictions for a movie to be placed in the correct class. The movies being placed in a specific class have certain levels of confidence associated with them. To illustrate, a movie is expected to hit the “very good ROI” class (the expected accuracy), what can be observed however is that a movie may be placed in any of the other 3 groups. We would expect that a movie will not be placed more than 2 classes below it, and therefore we have a specified a level of confidence that it will not be in the wrong group for each of the remaining groups. Previously, we have used the AUC, precision, and recall for the performance metric of a binary classification. In the case of a multiclass classification, we have substituted those 3 metrics with a metric called the Kappa Coefficient. The Kappa coefficient will take into account the observed accuracy with the expected accuracy. Here, we have four categories of success, and how we defined it:

1. Films were separated into four classes, our “low ROI” has an ROI value ≤ 50%, “good enough ROI” has an ROI value ≤ 100%, “good ROI” has an ROI value ≤ 250%, “very good ROI” has an ROI value > 250%. The decision variables were set based on typical expectations of investors’ ROI on a movie and the distribution of ROI in our dataset.

Table 7. Best 3 prediction results of the multiclass classification model

Classifier	**Random Forest**	**Decision Tree**	**Deep Learning**
Accuracy	0.864	0.845	0.754
Kappa	0.816	0.650	0.778

Examining our multiclass classification model's performance on three separate learning algorithms, Random Forest is the best classifier for this set of decision ranges. It produced the highest accuracy value and Kappa of 0.864, and 0.816 respectively. The table 7, displays those performance metrics. While the accuracy of the Random Forest Classifier may not have reached 0.9, it is important to note that a Kappa value of 0.81 or above is usually associated with a near perfect agreement on the accuracy of the model.

Chapter 6

Discussions

6.1 Key Features for Movie Success

In the previous chapter, we compared the performances of the different classifiers. By doing so, we ascertained that the random forest classifier outshined the other classifiers when we applied our current model. Nevertheless, this is not the only variable that can alter the scores of the predictive model. One main driver that can affect the score of the model is the features of movies in the dataset. To observe how much a feature impacts the output variables, and intuitively if they are representative of movie profitability, we would be using a weighting technique called the Gini index. In addition, we also ran a linear regression model that will predict continuous values for the movie ROI if a classification method to split movies into 2 or more discrete categories is not favorable to make investment decisions.

6.1.1 Regression Analysis

Four different algorithms, Deep Learning, Linear Regression, Lasso and Ridge Regression were tested for our regression analysis.

Table 8. Results for ROI prediction with different algorithms

Metric \ Algorithm	Lasso	Linear Regression	Deep Learning	Ridge Regression
RSME	185.336	185.616	187.661	258.118

Table 8 examines in contrast the root mean squared errors (RSME) of the 4 algorithms, with Lasso having the lowest RSME, meaning that it is the best model for the prediction of continuous values of a movie ROI.

6.1.2 Features and their Weights

To evaluate the performance or impacts of our set of features, we will use a statistical measure called the Gini index. The Gini index ranges from a value of 0 to 1. A weight value closer to 1 would mean that the feature has made a more significant impact. We listed the top 8 features with the highest Gini Indices in table 9. Most of the features from the "star features" group make the majority of the list. The star feature is mainly associated with one consolidated feature, the "star power", but here it was broken down into its smaller components. For example, the Total ROI for Director is a component of the combined star power, but the star power itself for a director has two other components (mentioned in chapter 4). Furthermore, the increase in film's ROI has been supported by features such as the theatrical engagements and sequel.

Table 9. Top 8 features with highest Gini Index from the classification models

Attribute Group	Attribute	Index
Star Feature	Total ROI for supporting actor	0.346
Star Feature	Total ROI for director	0.326
Star Feature	Total ROI for Actor	0.226
Time-Based Feature	Theatrical Engagements	0.138
Star Feature	Director Star Power	0.068
Descriptive Feature	Sequel	0.042
Star Feature	Director Google Search	0.034
Star Feature	Actor Star Power	0.032

6.1.3 Star Powers and Movie ROI

The preceding paragraph tells us that the Star Feature has a strong relevance on a film's success. A research done by Vany & Walls (1999) has verified that a higher Star Feature or Star Power would typically lead to a greater film success, concluding its success from a film's box office revenue with an actor's star power being his or her individual total gross.

We took a look at our how our own star features fared with the ROI of a movie. Our star features, which was highly derived by the historical individual ROI of both actors (leading and supporting) and directors, showed unique discoveries on the film's ROI. Figures 7, 8 and 9 shows that there is a general sense that a rise in the star power of a leading actor, supporting actor, and director will also bring the movie ROI up from a negative value to a positive value. However, the star powers do not have much of an impact on our movie ROI beyond a certain point. This occurrence may be because we have used a small dataset which does not properly distribute the movies and their ROIs. Other reasons may be certain movies like "Avatar" that have made humungous profits without having to have any known actors. Our correlation for the leading actor, supporting actor, and director are 0.042, 0.041, and -0.0009 respectively, based on the graphs alone.

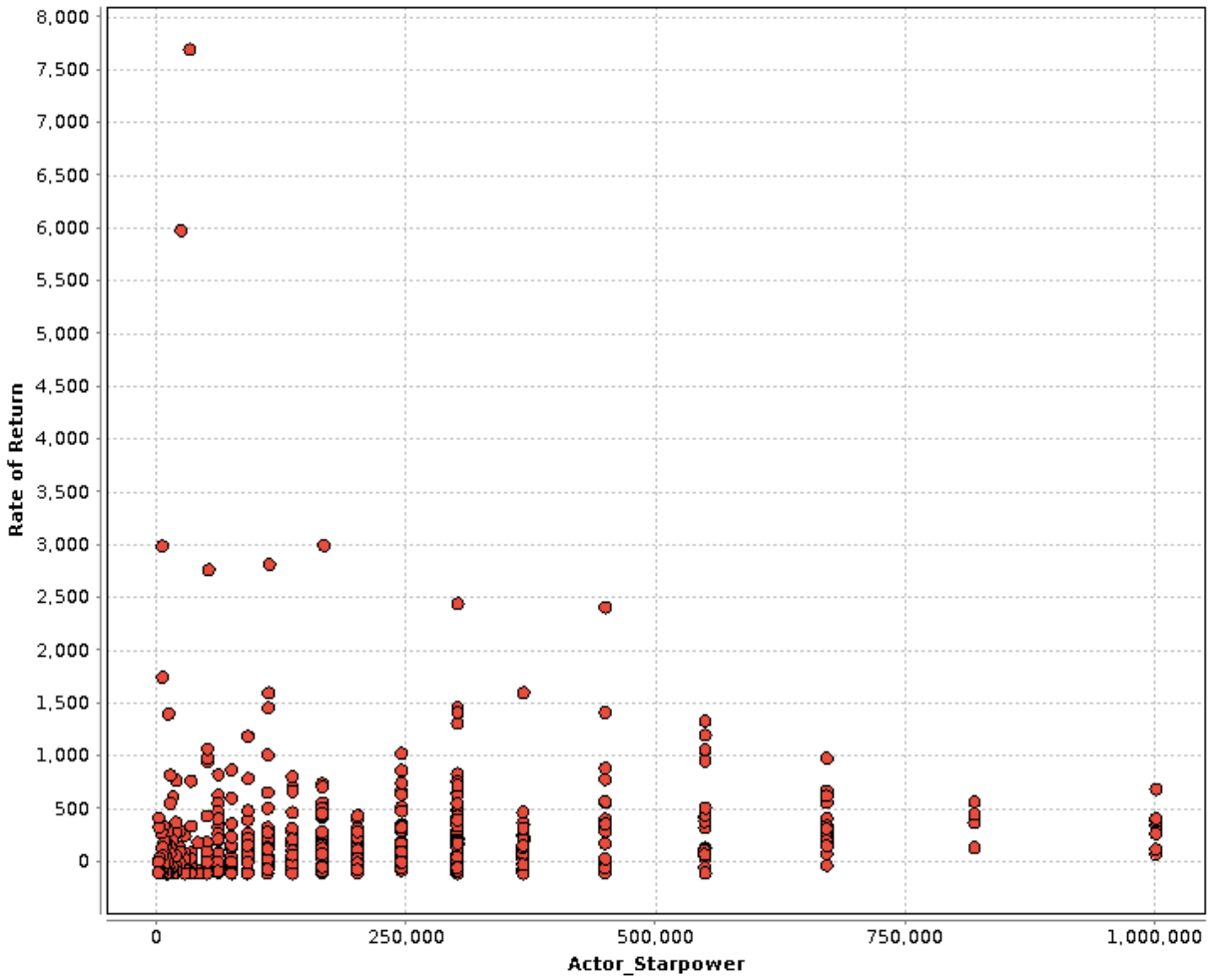


Figure 7. Leading Actor Star Power vs. Movie ROI

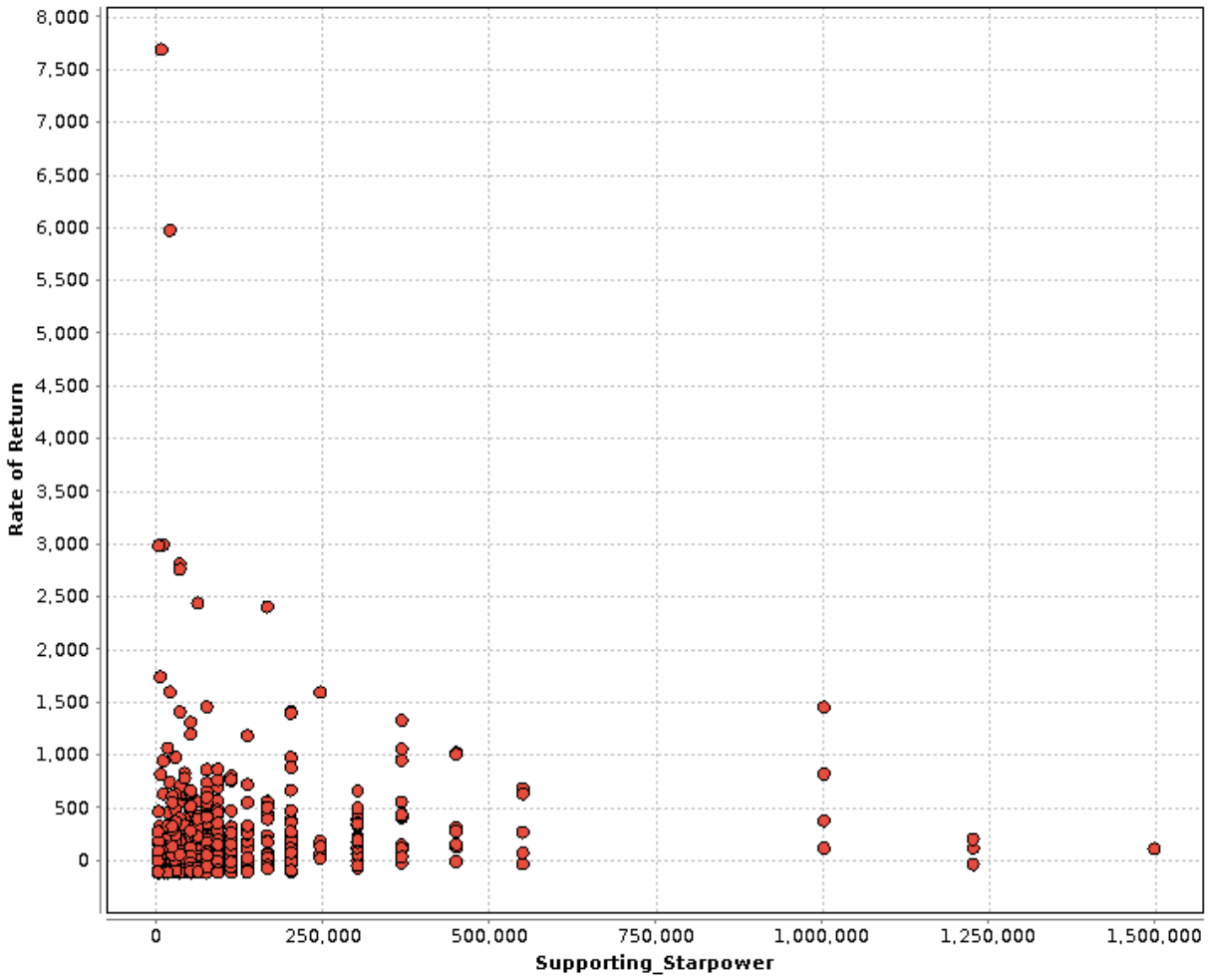


Figure 8. Supporting Actor Star Power vs. Movie ROI

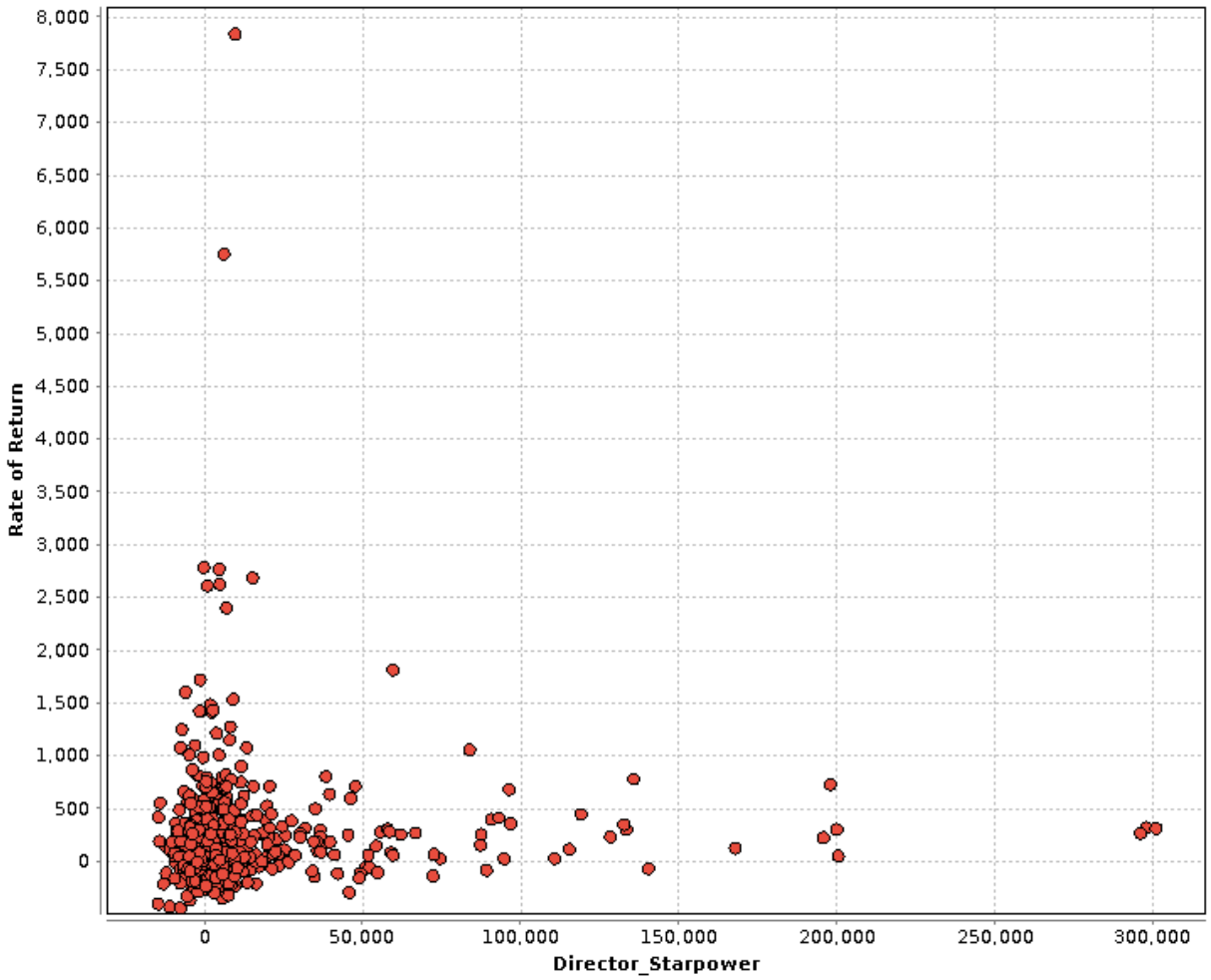


Figure 9. Director Star Power vs. Movie ROI

Since the definition of our star power consists of the individual star's ROI, google search and number of movies involved in, it may have skewed certain data points, making some stars seem that they have a great star power when in fact they haven't produced movies with relatively high ROIs. Table 10 lists the top ten leading actors, supporting actors and directors just based on their individual ROIs. Popular and highly rated stars typically will appeal to the majority of the audience, however, table 10 tells us otherwise when generating profits for a

movie. In conclusion, a star's ability in creating more ROI for a movie should be weighted more than his or her involvement in high generating revenue movies.

Table 10. Top ten stars by individual ROI

Rank	Leading Actors	Supporting Actors	Directors
1	Katie Featherston	Eric Young	Tod Williams
2	Kirk Cameron	Molly Ephraim	Alex Kendrick
3	Lena Headey	Rhys Wakefield	James Demonaco
4	Leila Hatami	Shahab Hosseini	Asghar Farhadi
5	Helena Bonham Carter	Guy Pearce	Josh Boone
6	Dev Patel	Anil Kapoor	Tom Hooper
7	Alex Kendrick	Willem Dafoe	Drake Doremus
8	Mick Jagger	Ben Davies	Ben Palmer
9	Felicity Jones	Mike Vogel	Steven Soderbergh
10	Joe Thomas	Christina Aguilera	Pierre Coffin

6.2 Limitations

Certainly our research comes with barriers that hinder us from building a model at full potential or capabilities. Due to the missing values for both the leading and supporting actors as well as the directors we had to make a decision based on two options. The first was to proceed with a dataset that has a few hundred in sample size although fully complete, shown in Table 3.3, or alternatively to proceed with a dataset that has thousands in sample size but has a great deal of missing values for certain important features. Eventually we decided to proceed with a complete dataset that has only a few hundred entries. This may not have gotten the most exceptional model due to certain biases within our dataset.

The second limitation is that the new actors or directors that may not have existed in our dataset would have their impacts disregarded, although they may actually prove to create successful movies in the future. There are so many great actors in the movie industry that have not been involved in blockbuster movies that the dataset may not have scoped. Investors would like to be able to have a greater selection of actors to choose from beyond the typical top billing list.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

This research proposed the foundation of a movie analytics model to predict the success of a movie in order to help stakeholders make sound investment decisions in the industry. We achieved this by analyzing historical movie data from a relevant movie database. Success of a movie is defined by the return on investment (ROI) rather than box office revenue, which is contrary to what most believe is to be the metric of success. Investment agreements have usually been made during the phase of preproduction. Fixed production budgets and genre would be almost impossible to alter. Hence our model will only use data that is provided at the beginning of production, keeping this in mind when establishing our set of features. Features in this model scopes those that are commonplace for every movie and those that we have innovated. These features comprise of four major categories, “star features” include the list cast members and their individual capabilities or potential, “integrated features” go further and take the synergies from the cast members, “descriptive features” give information on the movie, such as creative type, genre and rating, “time-based features” will be when the movie is going to be released and the amount of theatrical engagements during the period of movie release.

Movies spanning from 2007 to 2016 were used for our experiments, the performance of our model was measured by a classification method. By studying and testing the different classification algorithms, we could verify how well our model is doing in providing an

accurate prediction for a movie's return on investment. The model would accommodate investors by allowing them to modify the value of the parameters in the model to improve the profitability.

Features, such as star power (calculated using our variables), that we recommended applying and eventually formed in the model have proved to made an impact to the overall movie ROI. This study leveraged different techniques of machine learning in order to get our predicted ROI. Machine learning techniques allow the model to train itself based on historical data, and then test model to see if it has done a good job in making accurate predictions. The model sets aside a large portion of the dataset, containing the input variables as well as output variables, in the training phase, and then use a smaller and different portion of the dataset to measure the model's performance. We gathered our input variables or the features and the output variables or the return on investment of movies, and ran these through our model. In the training phase we used different learning algorithms on the training dataset and saw which ones worked best based on the performance data of the model, which we can obtain from the testing phase. The results of the testing phase tell us that the "Random Forest" algorithm has the highest overall accuracy with a value of 92.1%. This methodology allowed us to dynamically alter parameters in order to get more accurate prediction results. The correlation between a feature and the dependent variable was measured using the Gini Index. Our analysis found that the "Total ROI for a Supporting Actor" was the feature that made the most difference in predicting results with an index value of 0.346.

7.2 Future Work

There is still so much potential to raise the quality of the predictive models for further work or research. Beginning with our popularity factor within the star feature, our popularity factor takes into account the amount of google searches for the actors or directors in real time. Another aspect of popularity could be the number of followers each of these stars have on social media accounts such as Twitter, Facebook, Instagram and Snapchat. However, we would also have to weigh in this factor with respect to the tenure of the actor in the movie industry and not just simply take their popularity on social media. One example would be Justin Bieber, who has probably a hundred million followers but he may not be a great movie star just based on this factor alone. A next step would be using a technique to normalize all these considerations. A better google search would incorporate the actor or directors with keywords related to past movies they have done.

Collaboration amongst actors and directors was something we looked at, additionally we could further explore the concept of actor-actor or actor-director collaborations based on what the audience would like to see. This data could be obtained from official movie forums, relevant blogs, focus group discussions and surveys; however, it would be difficult to capture this data due to its random nature and developing standardized surveys are usually time-consuming and expensive.

Consumers change their taste haphazardly, and it is hard to predict what the market wants to see in a movie. Perhaps a storyline or topic that has never been presented before should be considered and let them be relative to what people in the world have been discussing about lately. A paradigm of this idea would be the 2017 movie “Get Out”, the movie addresses

the issue of racism in the modern world. The plot was relative to the issues that people were facing in the world.

Finally, we would like to introduce additional features such as competition, where we would look at movies that are released in the same time period. We must figure out if the other movies involved actors or directors with a high star power. This may affect the viewing base for our movie. Time-based features could be further explored by incorporating movies and their most optimal release dates (i.e. a horror movie to be released before Halloween).

Appendix A

Machine Learning in RapidMiner

RapidMiner is a data science software platform mostly used by data scientists, statisticians and even business analysts. It started as an open-source software where there are sources codes open for redistribution and modification for anyone. Just like any other open-source software, RapidMiner has open interfaces and can be extended using R and Python scripts.

It was first developed by Ralf Klinkenberg, Simon Fischer, and Ingo Mierswa in 2001, and now has millions of downloads and has about 250,000 customers that download the paid edition including Intel and Samsung. RapidMiner provides data mining and machine learning processes that has been verified by the employees of RapidMiner, who have been backed by multiple venture capitalists and received $16 million in funding. In this thesis, RapidMiner will aid the processing of all the data captured. Here are some of the practical functions available in RapidMiner:

- Read and write datasets to and from an Excel (.csv) file
- Data cleaning and consolidation
- Machine Learning techniques for training and testing
- Visualizes performance vectors for predictive models

RapidMiner allows a GUI to craft and run analytical workflows or flowcharts. This feature would allow easy access for those users who aren't as proficient in writing scripts in other programming languages. RapidMiner is used in areas such as education, research, business and commercial companies. The flowcharts were validated and verified by RapidMiner and can was stored.

Appendix B

Algorithm flowcharts

This appendix lists all the flowcharts and functions for the data preparation, learning algorithm performance vector and the predictive modelling.

Data Preparation

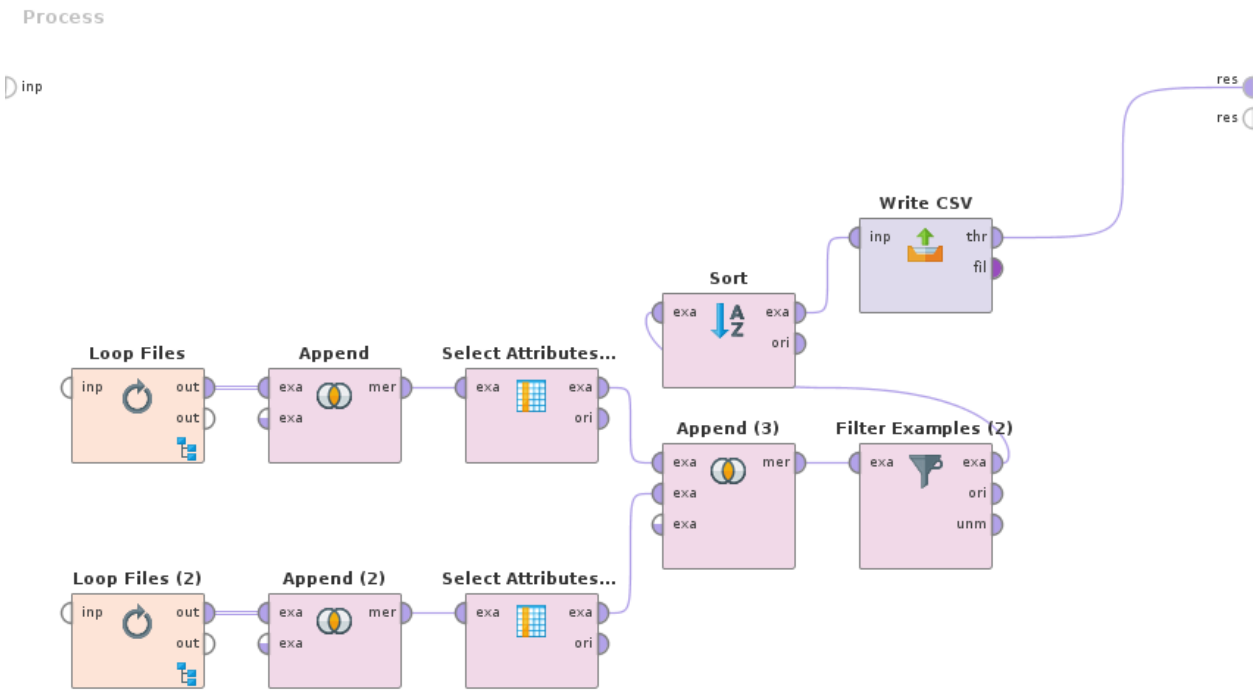


Figure 10. Annex Data Preparation in RapidMiner

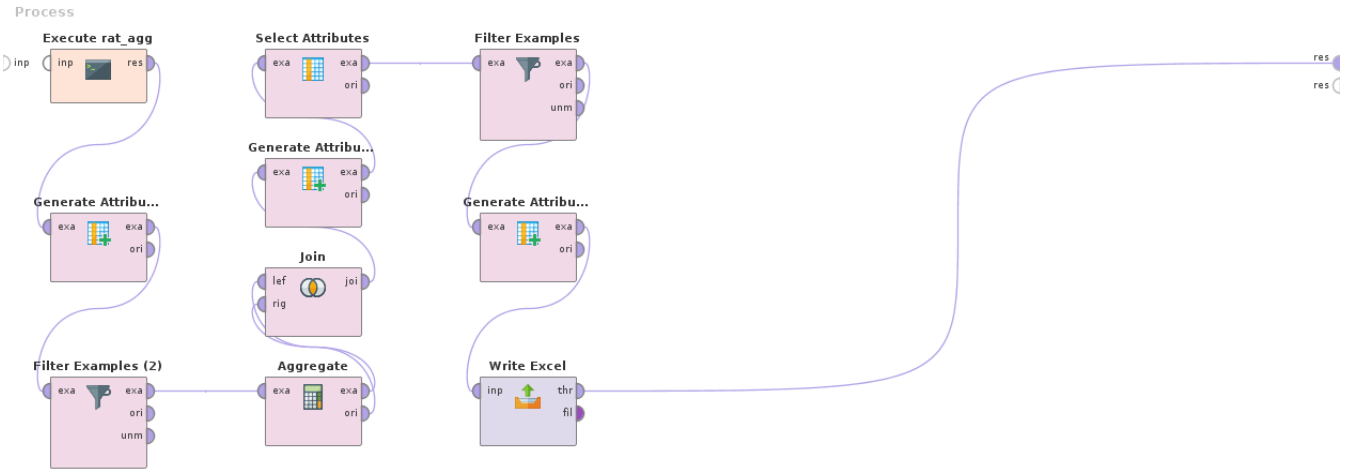


Figure 11. Star Power Consolidation in RapidMiner

Figure 12. Final Dataset Consolidation in RapidMiner (part i)

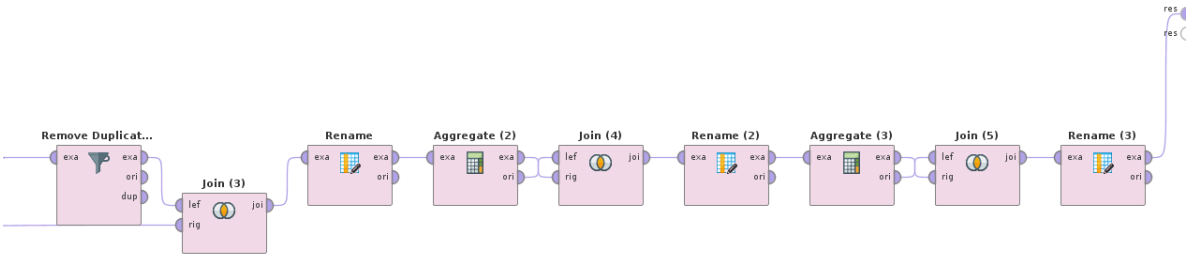


Figure 13. Final Dataset Consolidation in RapidMiner (part ii)

Predictive Modelling

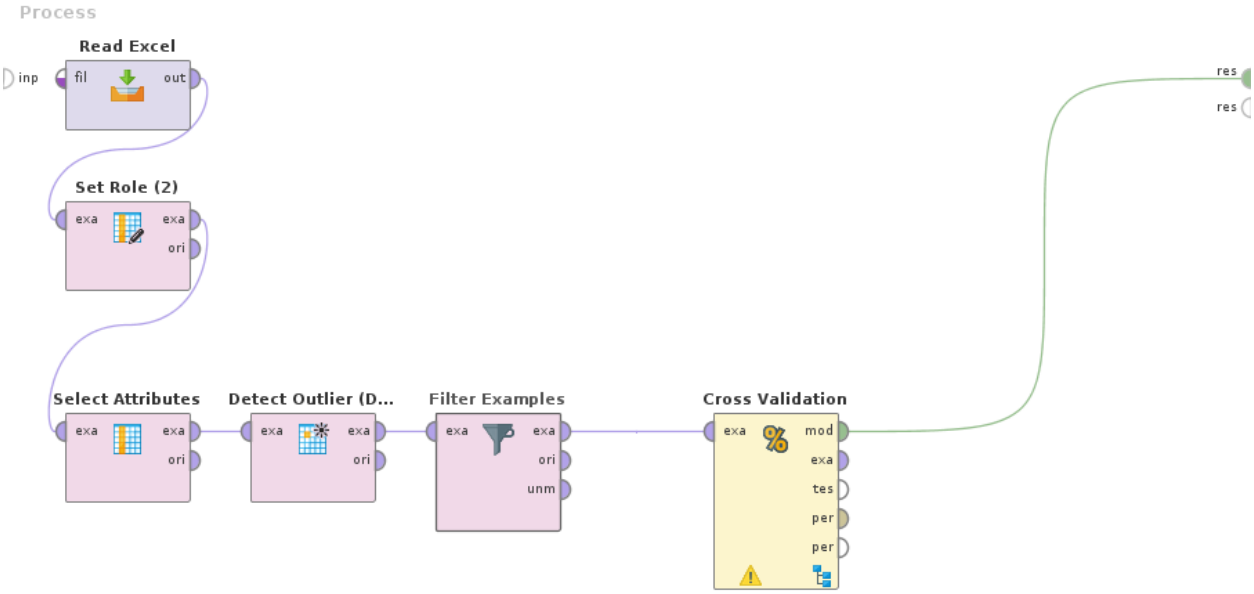


Figure 14. Example of Cross Validation in RapidMiner

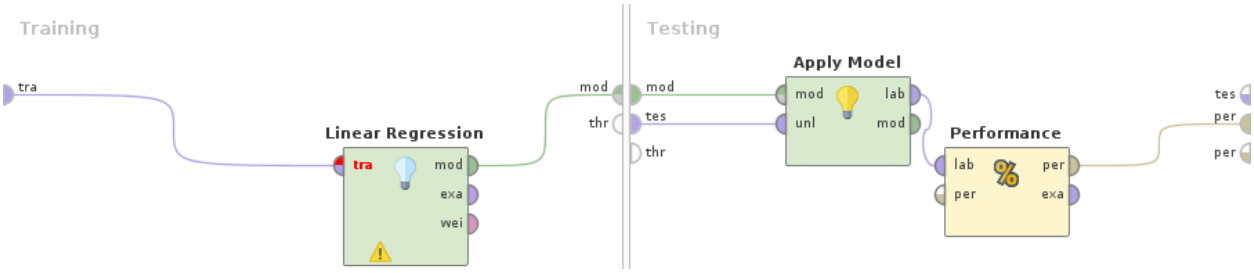


Figure 15. Example of Learning Algorithm and Modelling in RapidMiner

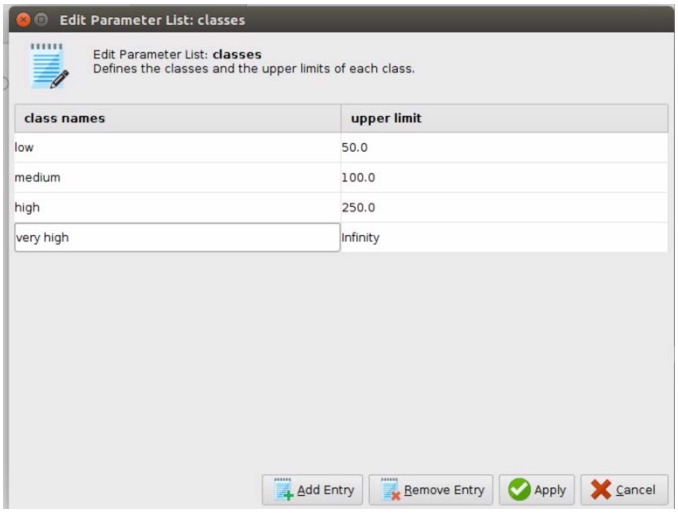


Figure 16. Multiclass discretization example in RapidMiner

Performance Vector

Table 11. Additional prediction results of the binary classification model (Range 1)

Classifier	Naïve Bayes	SVM
Accuracy	0.657	0.822
AUC	0.873	0.945
Precision	0.937	0.787
Recall	0.505	0.997

Table 12. Additional prediction results of the binary classification model (Range 2)

Classifier	Naïve Bayes	SVM
Accuracy	0.797	0.859
AUC	0.931	0.932
Precision	0.914	0.816
Recall	0.603	0.933

Table 13. Additional prediction results of the multiclass classification model

Classifier	Naïve Bayes	SVM
Accuracy	0.643	0.669
Kappa	0.490	0.513

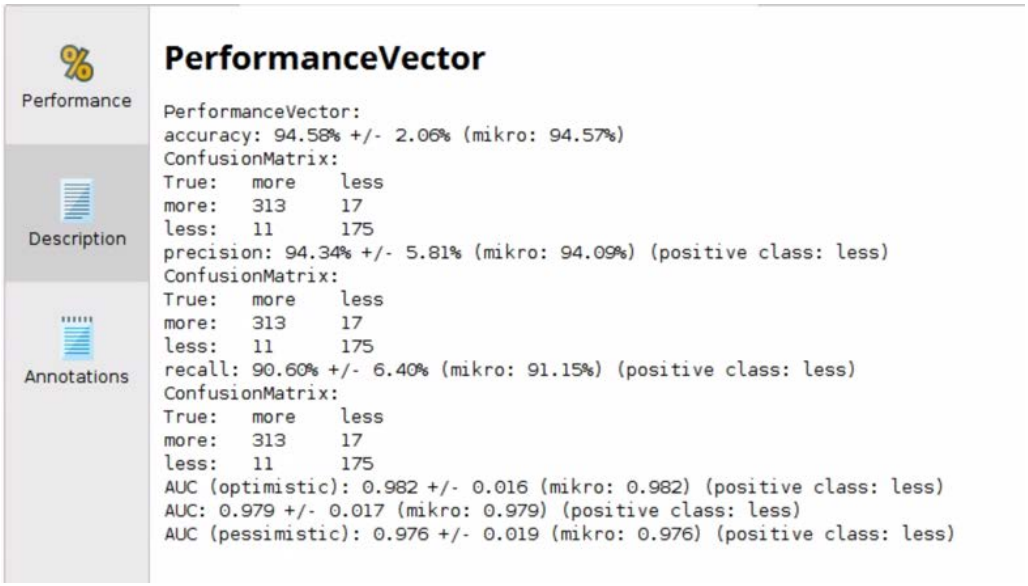


Figure 17. Display of performance vector in RapidMiner

BIBLIOGRAPHY

Lash, M. T., & Zhao, K. (2016). Early Predictions of Movie Success: The Who, What, and When of Profitability. Journal of Management Information Systems, 33(3), 874-903.

Global box office revenue from 2016 to 2020 in billion U.S. dollars. (2017). Global box office revenue 2016 | Statistic. Retrieved from https://www.statista.com/statistics/259987/global-box-office-revenue/

Leading box office markets worldwide in 2016, by revenue in billion U.S. dollars (2017). Leading box office markets worldwide by revenue 2016 | Statistic. Retrieved from https://www.statista.com/statistics/243180/leading-box-office-markets-workdwide-by-revenue/

Klinkenberg, R., & Hofmann, M. (2013). RapidMiner: Data Mining use Cases and Business Analytics Applications. CRC Press.

Litman, B. (1983). Predicting Success of Theatrical Movies: An Empirical Study. Journal of Popular Culture, 16(4), 159-175.

Litman, B.R., & Kohl, L. S. (1989). Predicting financial success of motion pictures: The ‘80s experience. Journal of Media Economics, 2, 35-50.

Terry, N., Butler, M., & De’Armond, D. (2005). The Determinants of Domestic Box Office Performance in the Motion Picture Industry. Southwestern Economic Review, 137-148.

Wallace, W. T., Seigerman, A., & Holbrook, M. B. (1993). The role of actors and actresses in the success of films. Journal of Cultural Economics, 17, 1-27

Nelson, R., & Glotfelty, R. (2012). Movie Starts and Box Office Revenues: An Empirical Analysis. Journal of Cultural Economics, 36(2), 141-166.

Sharda, R., & Delen, D. (2006). Predicting Box-Office Success of Motion Pictures with neural networks. Expert Systems with Applications, 30(2), 243-254.

Ghiassi, M., Lio, D., & Moon, B. (2014). Pre-production Forecasting of Movie Revenues with a dynamic artificial Nerual Network. Expert Systems with Applications, 42(6), 3176-3193.

Wagura, D. (2016). Movie Actor Success Prediction. The Pennsylvania State University Schreyer Honors College, Harold and Inge Marcus Department of Industrial and Manufacturing Engineering.

Vany, A., & Walls, W. (1999). Uncertainty in the Movie Industry: Does Star Power reduce the Terror of the Box Office? Journal of Cultural Economics, 23, 285-318.

ACADEMIC VITA OF PRESTON A. SOEPRANOTO

115A South Butz Street, State College, PA 16801 | Mobile: 973-629-2493 | Email: pas5517@psu.edu

EDUCATION

The Pennsylvania State University | **Schreyer Honors College** Expected Graduation: **Dec 2017**
Bachelor of Science: **Industrial Engineering**

WORK EXPERIENCE

The Boston Consulting Group **Jakarta, Indonesia**
Summer Associate *05/17-08/17*
Deep and Broad Case/Project Experience for a National Oil and Gas Company:

- Formulated a marketing digitalization roadmap and new business initiatives
- Conducted a deep-dive research on key customer segments through a customer ethnography which involved surveys, benchmarking, focus groups, workshops, store visits and dealer & customer interviews that made initial list of concepts
- Generated and prioritized concepts to develop into Minimum Viable Products (MVP), and worked with the client's cross-functional teams to build working prototype
- Composed a list of questions for focus group discussions to get early customer feedback for MVP Piloting
- Created criterions to select a location for the piloting of MVP and was involved in initial kick-starting phase
- Designed a quantitative analysis report, using data analytics, from ethnography data and gathered key insights which led to an increase in the market share and upselling of products for the client's business

Mataharimall.com (#1 E-Commerce in Indonesia), Lippo Group (#1 Retailer in Indonesia) **Jakarta, Indonesia**
Software Engineer *05/15-08/15*

- Developed dynamic pricing platform for consumer electronic segment of Mataharimall.com
- Targeted specific consumer electronics' pricing to create loss leaders
- Simulated dynamic pricing and loss leader creation to determine viability for a loss leader pricing strategy on popular electronics
- Acquired a third party company, Prisync to track and visualize competitor pricing on identified loss leaders
- Presented recommendation for differentiated pricing of iPhone 6s models to raise the sales of more profitable models

RESEARCH EXPERIENCE

Schreyer Honors Thesis on Predictive Modelling for Movie Success **University Park, PA**
Honors Scholar *08/17-present*

- Objective is to identify potentially successful movies before its release, based on factors such as actor, director and budget
- A projected box office is the metric assessed for a successful movie
- Main motivation for this model is to aid movie producers in developing a movie prior to its production
- Methodology formulated will be based upon machine learning techniques
- Intended outcome would be a fully validated, working and accurate model that can perform its functions

Penn State Laboratory for Intelligent Systems and Analytics (LISA) with Dr. Soundar Kumara **University Park, PA**
Honors Research Assistant *03/17-present*

- Led a team in creating a predictive model on a hospital's arrival rate using PivotTable to determine different trends and seasonality

Penn State Information Sciences and Technology Department with Dr. Vasant Honavar **University Park, PA**
Research Assistant *05/16-05/17*

- Modeled macromolecular interaction with naïve Bayes based classifiers built in Python with Scikit-learn
- Created data pipeline to manage, sort and organize protein data sets from multiple sources
- Integrated R based motif finding protein sequence databases

Penn State Industrial Engineering Department with Dr. Guo Dong Pang **University Park, PA**
Research Assistant *01/16-05/16*

- Improved voice recognition (IVR) system for ADT Security customer service call center
- Better matched callers with self-service solutions in the IVR system by applying stochastic modeling techniques
- Decreased average agent queue by estimated 15%

CAMPUS ORGANIZATIONS

Nittany Data Labs; *Head Business Strategist and Software Engineer* *08/15-05/17*

- Developed data visualization for project with Penn State Housing and Food Services
- Taught workshops on machine learning in Python
- Assisted with member development and management

Penn State Institute of Industrial Engineering (IIE); *Committee Head* *08/15-05/16*

- Directed the recruiting program for organization

- Assisted in organizing speakers and events

Engineering Club; *Vice President* *01/15-05/15*

- Taught workshops on AutoCAD and Arduino controlled systems
- Lead fall and spring recruitment

Student Government Association; *Senator* *08/14-05/15*

- Head of Governmental Affairs Committee
- Coordinated career fairs, formal nights and misc. student events

SUMMARY AND SKILLS

***Engineering Skills**: Big data Analytics, Machine Learning, Mathematical Modeling, Stochastic Modeling, Probability and Statistics, Engineering Design*

***Business Skills**: Management Consulting, Completed Wall Street Oasis, Supply Chain Management, Managerial Accounting, Engineering Economics*

***CS Skills**: Data structures and Algorithms. Languages: C++, Java, Python, LaTex, SQL, R, HTML/CSS, JavaScript, Excel Modeling; Software: X-code, AutoCAD, Visual Studio, Weka, MATLAB, Simio, Tableu, SolidWorks, MS Excel, MS PowerPoint*