SENG 474 Data Mining

Project Report

Haoming Zhang
Fei Xiang
David Goede
Sean Hoessmann

April 8, 2020

Introduction:

The production of a movie is an incredibly costly endeavor. There is a massive commitment of time, capital, and personnel with no guarantees of a worthwhile return. And yet, despite this risk every year hundreds of movies are still produced. Predictive modeling plays a key role in this aspect of the movie industry. Knowing which movies are bound to fail and which will succeed before the movies have been produced is incredibly valuable information for studios to have in deciding which projects to go through with. With this studios can allocate their resources accordingly such as determining which movies will receive a sequel based on the first movies profits and historical data of the success of sequels of similar movies. Other applications include determining the budget of higher risk lesser known movies such that they maintain an acceptable profit margin as well as determining marketing and advertising budgets.

In this project we aim to replicate this process and predict the success or failure based on several characteristics. In order for this model to have any practical usefulness we will limit the features to those that could be known before the movie is officially released to theaters. Some of the main criteria we will focus on will be Rating, Number of votes, Metascore(imdb), Metascore(rotten tomato), YouTube Trailer views, YouTube like-dislike Ratio and so on. With this information we will implement various methods described in detail below to predict the expected ratings and revenue of the given film. We had originally intended to include other attributes such as producers, writers, and studio backing but have found that many datasets do not include this data.

The inclusion of ratings and metascore may seem strange for the prediction of a movie's revenue and success, how can these be known before the movie is released? While the actual public opinion on a movie can not be known before release, production companies employ many methods to gauge public reaction during production well before the release date. Through the use of focus groups, test screenings and sneak previews studios get their first indication of public reception to the film and can roll back production as needed based on this. This step plays a critical role in the decision making behind a movie and as such we felt it important to include in our models. While the actual data collected by the production companies in their focus group testing is not publicly available the ratings post release are. As the early release feedback collected from these focus group tests is meant to sample the post release feedback we believed it to be similar enough to use in our analysis.

For datasets we have utilised the collection of 1000 movies released between 2006 and 2016 available from kaggle. After cleaning the dataset and removing incomplete data we have 838 movies. We have elected to use movies of a similar timeframe as there can be major differences that render the data useless for future predictions if they were released very long time ago. For example 1960's Spartacus was the highest grossing movie of 1960 and was Universal Studios biggest moneymaker for a decade and has an all time worldwide box office of 60 million dollars. Comparatively the highest grossing movie of 2010 was Toy Story 3 and had a box office of over 1 billion dollars. While both movies were box office successes at the time of their release the massive discrepancy in what constitutes a successful revenue makes older movies less useful in predicting future movies success. As such our current dataset is skewed towards more recently released movies.

The Problem:

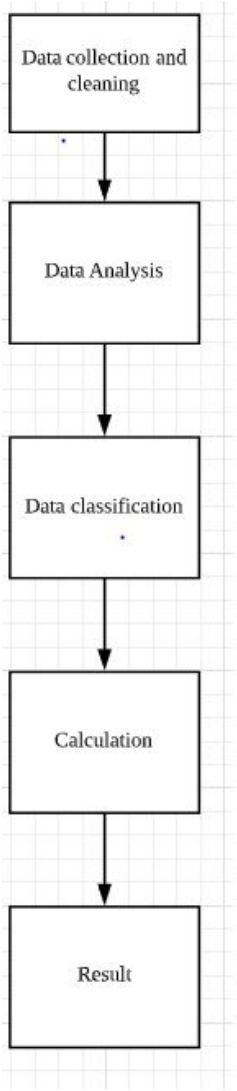
Different people can have varying opinions on the same movie, as such it can be hard to quantify a successful movie on something as subjective as opinion. Therefore, we consider the revenue generated when defining a success. For example, if a particular movie's revenue is higher than 90% of other movies in the data set, it is safe to say the movie was a success. We collected our data from multiple sources mainly imdb and rotten tomato. Furthermore, we only consider the actual "data", so we remove attributes such as the descriptions and movie runtimes.

We apply the linear regression method and calculate the OLS regression table to find the correlation between revenue (success of a movie) and other attributes. Then we made a model to predict the success of movies.

Goal and architecture of Project:

The goal of the project is to predict the performance of movies. In this project, we determine a movie's success by its revenue. In this work, data mining techniques will be used to extract patterns which can be useful in anticipating a film's success. This approach is significant since these data mining techniques can recognize the connections among different factors.

The project's architecture:



Data collection and cleaning: collecting data and remove invalid or incomplete data

Data analysis: apply method of linear regression and find correlation between revenue and other attributes

Data classification: Add weight to each valid attribute

Calculation: calculate the mean of weight and make a prediction

Result: make predictions on different movies

Data collection and cleaning:

We are using the movie dataset consisting of data from imdb, Rotten Tomato and YouTube. There are 1000 movies initially and the attributes that we collected include the following:

. . . . . .
-- Movie name
-- Release year
-- Number of votes
-- Actor credits (imdb)
-- Directors Credit (imdb)
-- YouTube like-dislike Ratio
-- Metascore (Rotten Tomato)
-- Revenue
. . . . . .

After we have the datasets, we do the data cleaning by deleting lines with missing/invalid values and replacing empty cells with a specific value

Data analysis:

We are using movie data from movies released between the years 2006 and 2016 with the majority of the movies being released in 2015 to 2016 as shown in figure 1 below. Each movie has data for each of the attributes defined above.

We want to know how each attribute connects to the success of movies (revenue). Correlation of these attributes and success of movies (revenue) can be found by linear regression analysis between them.

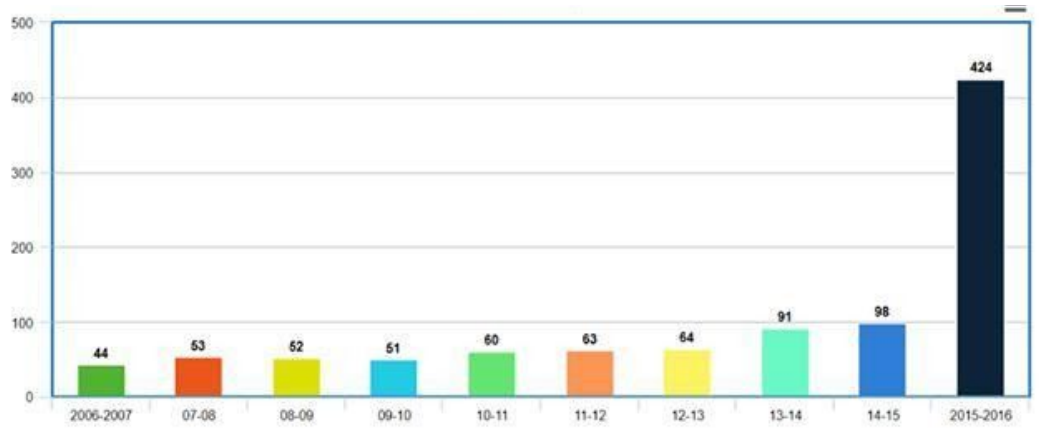


Figure 1: *Number of movies released each year in our dataset*

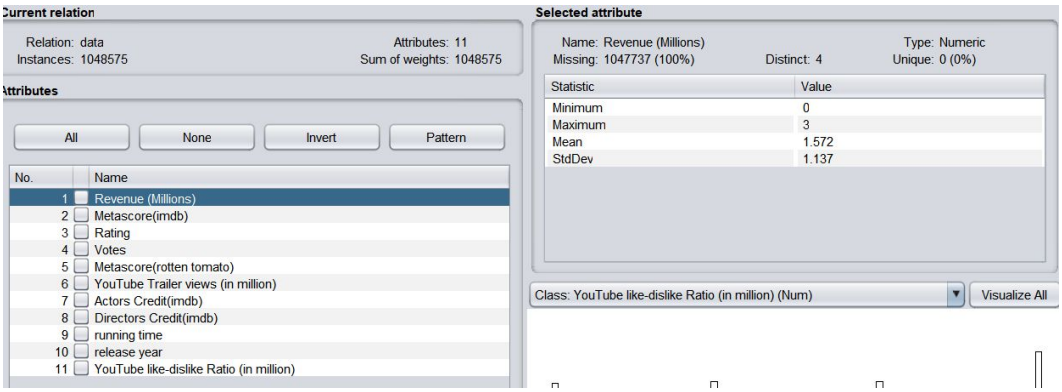
The linear regression line can be presented as: Y = a + bX where X is the explanatory variable and Y is the dependent Variable. Furthermore, the model for multiple linear regression can be presented as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

For i = n observations, y_i = dependent variable, x_i = explanatory variables, β0 = y-intercept (constant term), p = slope coefficients for each explanatory variable and **ε** =the model's error term (also known as the residuals).

We will use linear regression for showing the correlation between movie success (revenue) and attributes such as imdb rating, the number of votes (imdb) and imdb metascore. Also, the result could be checked by observing the OLS regression table.

First, we input the dataset to weka and apply linear regression (notice we split the revenue into 4 levels which makes our experiment easier):




```
=== Summary ===

Correlation coefficient                  0.8998
Mean absolute error                      6.5647
Root mean squared error                  8.2409
Relative absolute error                 41.1952 %
Root relative squared error             43.5375 %
Total Number of Instances              838
```

The relative absolute error is 41.1952%
The root relative squared error is 43.5375%

Attributes that are unrelated to the output variable can also negatively impact performance. To improve performance, we could add or remove attributes of datasets. For example, "release year" does not strongly correlate with the revenue (success of movie). We could remove the attribute "release year" to get a better performance

```
=== Summary ===

Correlation coefficient                  0.948
Mean absolute error                      0.2853
Root mean squared error                  0.3616
Relative absolute error                 27.9913 %
Root relative squared error             31.8378 %
Total Number of Instances              838
```

Now, we reduced the relative absolute error to 27.9913% and the root relative squared error is 31.8378% by adding/removing attributes.

For now, we will focus on specific attributes and analyze the correlation between revenue and them.

We first see the effect of Votes on revenue below in figure 2.

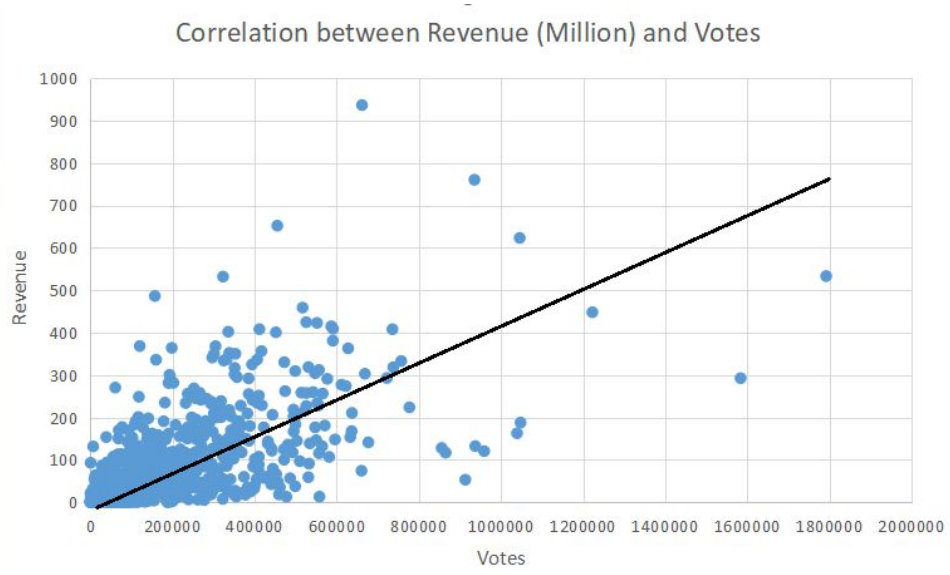


Figure 2: *Revenue (in millions) and number of votes*

To investigate whether the number of votes has a strong influence on the movie's revenue. We use the OLS regression table below (figure 3) to check.

	Coefficients	*tandard err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	17.95736	3.941718	4.55572	6E-06	10.22054	25.69419	10.22054	25.69419
Votes	0.000345	1.44E-05	23.88216	1.64E-96	0.000316	0.000373	0.000316	0.000373

Figure 3: *OLS Regression Table, # of votes and revenue*

As we can see, the p-value is very small. Therefore, we know Revenue and the number of Votes are independent and the number of votes has a strong influence on the movie's revenue. Secondly, the effect of Metascore on Revenue.

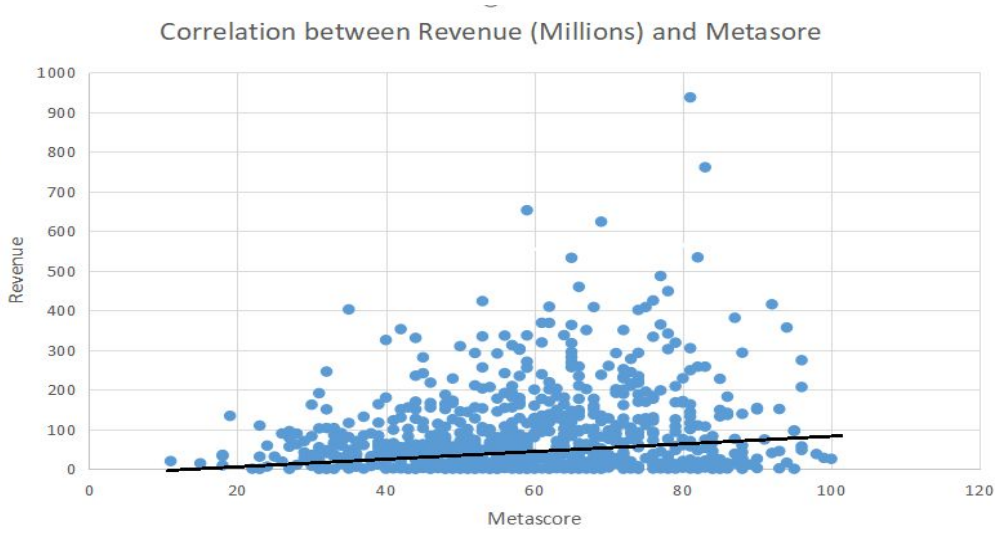


Figure 4: *Revenue (in millions) and metascore*

We then apply OLS regression table (Figure 5) on it.

	Coefficients	*tandard err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	32.2606	13.07285	2.467755	0.013796	6.601131	57.92007	6.601131	57.92007
Metascore	0.877949	0.211066	4.159604	3.52E-05	0.463668	1.292229	0.463668	1.292229

Figure 5: *OLS Regression Table, metascore and revenue*

The p-value is very small and therefore the Revenue and Metascore are strongly correlated.

We apply the same strategy on the correlation between Revenue and Rating (Figure 6) and the OLS regression table (Figure 7).

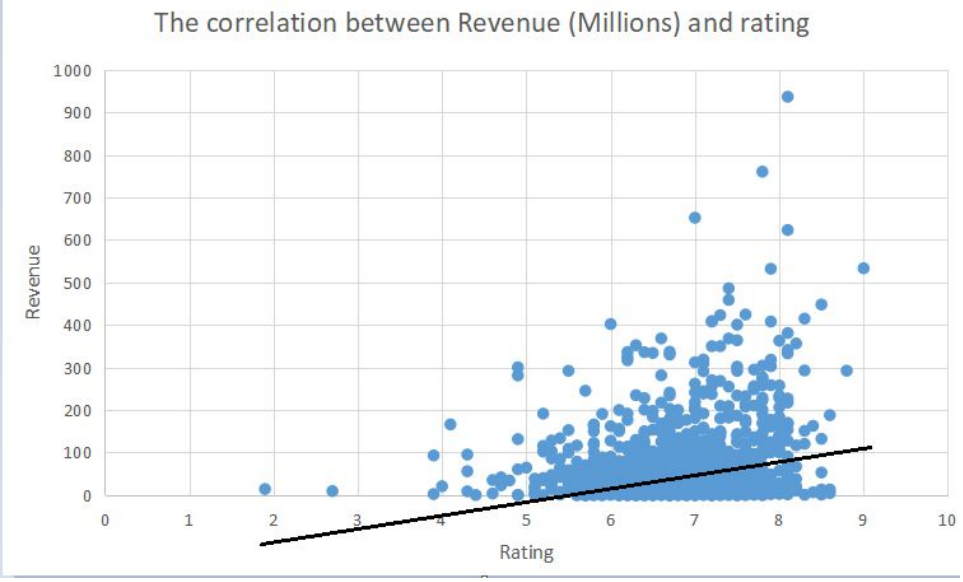


Figure 6: *Revenue (in millions) and rating*

	Coefficients	*tandard err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	17.95736	3.941718	4.55572	6E-06	10.22054	25.69419	10.22054	25.69419
Votes	0.000345	1.44E-05	23.88216	1.64E-96	0.000316	0.000373	0.000316	0.000373

Figure 7: *OLS Regression Table, ratings and revenue*

Rating strongly affects the movie's revenue as the p-value is still very low.

In conclusion, Rating, the number of votes and Metascore play important roles in movie success. The same method is applied for all other attributes.We found the following attributes are strongly correlation with the success(revenue) of a movie:

Rating, Number of votes, Metascore(imdb), Metascore(rotten tomato), YouTube Trailer views, YouTube like-dislike Ratio, Actors Credit(imdb), Directors Credit(imdb).

Data classification:

Each attribute of the datasets which affects a movie's revenue are classified in 4 categories. Based on the performance of each attributes, each attributes is assigned a classification from 0.2 to 1.0 where:

—Poor - 0.2
—Average - 0.5
—Good - 0.8
—Excellent - 1.0

Attributes	Poor	Average	Good	Excellent
Rating	Less than 6.0	6.1-7.0	7.1-8.0	Above 8.1
Number of votes (in million)	Less than 0.1	0.1-0.25	0.25-4	Above 0.4
Metascore(imdb)	Less than 40	40-60	61-80	Above 80
Metascore(rotten tomato)	Less than 40	40-60	61-80	Above 80
YouTube Trailer views (in million)	Less than 20	20-50	51-100	Above 100
Actors Credit(imdb)	Less than 20	21-40	41-60	More than 60
Directors Credit(imdb)	Less than 20	21-40	41-60	More than 60

Calculation

From the classified data, we can predict the success of a movie with the weight assigned to each attribute.

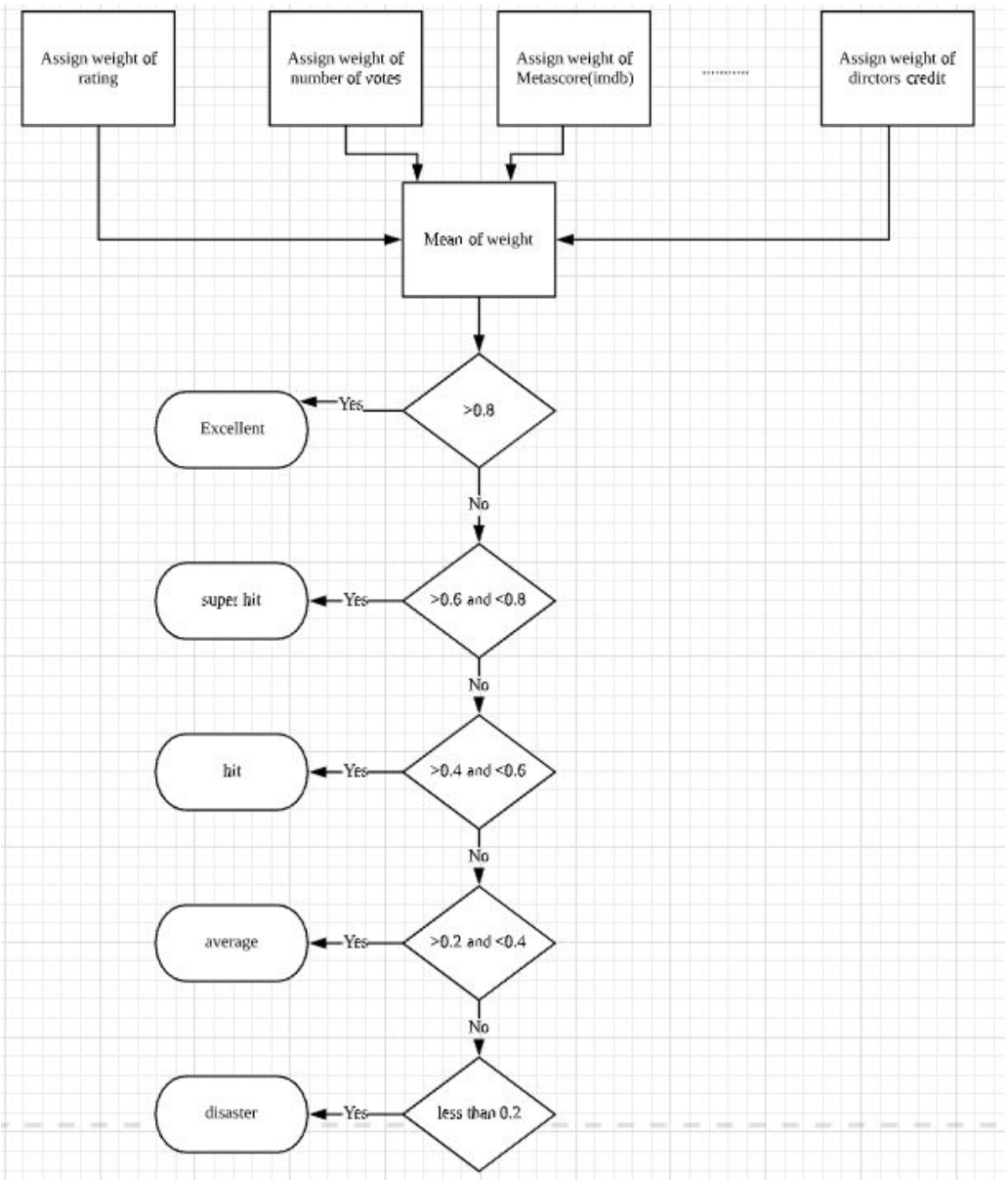
The prediction model is

$$X = \sum_{i=0}^{n} (w_i) / n$$

n = 7 in this case because we have 7 weights of attributes

We then classify the result into five categories:

— Disaster - less than 0.2
— Average - between 0.2 to 0.4
— Hit - between 0.4 to 0.6
— Super hit - between 0.6 to 0.8
— Excellent - above 0.8



Result

We apply above model and do experiment on the movies below

Tittle	Score (mean of weight)	prediction
Little Woods	0.52	hit
Grown Ups 2	0.81	Excellent
Fantastic Four	0.45	Average
Aladin	0.63	Super hit
Vampire Academy	0.32	Average

The results are shown above. The model can successfully predict the success of a movie as the results are close to the real situation

Conclusion/Discussion

In this project, we successfully developed a model to predict the performance of a movie. However, a movie's success does not only depend on the above attributes. In some specific year, the political conditions and economic stability of a country may affect the success(revenue) of a movie. For example, revenue of a given movie in 2008 was generally less than a given movie released in 2007 due to the economic crisis at the time. A similar result will likely be seen for movies released during the current pandemic. In future work, we should consider those factors as well as we could use different data mining techniques to analyze the problem of the project.

Linear regression is a simple method of data mining. We also apply different data mining techniques on the dataset.

We use 80% of the training set and 20% testing set on different approaches.

Logistic regression:

```
=== Summary ===

Correctly Classified Instances         152               90.4762 %
Incorrectly Classified Instances        16                9.5238 %
Kappa statistic                          0.8714
Mean absolute error                      0.0771
Root mean squared error                  0.1926
Relative absolute error                 20.3688 %
Root relative squared error             43.8987 %
Total Number of Instances              168
```

MultilayerPerceptron:

```
=== Summary ===

Correctly Classified Instances         154               91.6667 %
Incorrectly Classified Instances        14                8.3333 %
Kappa statistic                          0.8868
Mean absolute error                      0.0736
Root mean squared error                  0.1825
Relative absolute error                 19.4439 %
Root relative squared error             41.6075 %
Total Number of Instances              168
```

In future work, we will use those two approaches to analyze our problem.

References

Jason Brownlee. Linear regression for machine learning. https://machinelearningmastery.com/linear-regression-for-machine-learning/. Accessed: 2020-03-07.

Thomas W Dinsmore Michele Chambers. Predictive analytics techniques. pages 5–17, 2014.

Deborah J. Rumsey. What a p-value tells you about statistical data. https://www.dummies.com/education/math/statistics/what-a-p-value-tells-you-about-statistical-data/. Accessed: 2020-03-07.

Iawei Han, Jian Pei, and Micheline Kamber. Data MiningConcepts and Techniques, 2012. http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

Linear Regression,http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm