Final Project
Computational Data Science
Professor Nitin Upadhyay

Prepared by
Maharshi Vyas 201401414
Keval Shah 201401418

Is movie a Box Office Success? Analyzing Regional Search Queries to Predict Commercial Success

Table of Contents

ABSTRACT

Movie production is a risky and time consuming task. significant resources are invested in creating, editing, distributing and showcasing a movie across the world. In this industry, the most common and high-priority question to be answered is "*whether a particular movie with the specified cast in the selected genre will be a box office success or not?".* This will allow the stakeholders associated with the movie to explore the opportunity space in promoting, branding and repositioning the strategies to make gain commercial success while maintaining the overall gist of the movie plot and other core features. In recent days, viewers are extensively using search engines to know about the features of a movie such as genre, cast, studio etc. These platforms in turn provide more reliable insights on the generated search engine data that can be traced to societal trends and behaviour. Therefore the massive data generated from the search engines has widened the perspective of the market research and analysis. In this paper, authors provide movie analytical framework to predict its success based on the the available massive search query data.

1. INTRODUCTION

There has been a great amount of work for analysis of search engine queries which can provide extensive social applications.

In this paper we are trying to predict how much a particular movie will gross in a selected market a country. We have selected different countries as described in section.3.4 and predicted how much money a movie will pull from the country. In the entertainment industry, it is important to pool in factors such as cast, director and genre popularity. A viewer will decide to watch the movie if the combination of these three fits into his personal taste. These factors may have different importance in different countries so we are analysing country wise, which can be used for market research for a movie stakeholder to decide how to distribute a movie to maximize profits. So our market prediction strategy can be used directly by distributors.

2. RELATED WORK

Data scientists have used search engine data to to create models that predict the various features and that may be used as a judging criteria for movies. Depending on the user requirement, one can make predictions based on factors such as movie critic ratings, user ratings, box office collection and cast popularity.Nithin[3] has predicted movie success based on IMDB data. Deniz Demir[1] has predicted IMDB movie rating using the Google trends which provides access to google search engine data. Demir[1] has implemented this taking the movie viewers as primary users. Lee[2] entirely focusses on movie performance in markets based on metrics such as box office collection, number of theaters and ratings. Their primary source for data collection is the google trends query database. Both the authors have used linear regression for modelling. This paper will extend Lee[2] and Demir[1] work to include country-wise analysis along with improvisation in some methods and addition of other practical factors that define a movie popularity and success

One figure to show the sources of data, characteristics

One figure to show the analytic process -- can be a flow/system/method diagram

4. DISCUSSION AND FINDINGS

5. RECOMMENDATION TO MOVIE STAKEHOLDERS

3.FEATURE SELECTION AND DATA CHARACTERISTICS

3.1 Google Trends

provides a search popularity index(SPI) of a Google query in a specified time period and geographical region. This index is based on the query share, i.e. the total number of searches for a particular query divided by the total search volume seen in the specified geographical region and within the given time period. Afterwards,it performs another round of normalization and maps the highest and lowest values with 0 to 100 integers. Thus Google shows the normalized relative search statistics,and the normalization constant varies from query to query. More in choi[4].
So SPI = (Relative Frequency of a query)/(Span of the relative frequencies)
Note that this fact does not allow us to compare the popularity of two different queries using the trends data only. Two movies may have SPI 100 for same geolocation but may have huge difference in volume of overall queries.
Deniz Demir[1] fixes this problem using AdWords data which shows the exact amount of frequencies for a query. But there is no surety if google uses the same data for Google AdWords and google trends. So they also have large false predictions because of that. Lee[2] does not provide any solution to this. It may lead to incorrect predictions if lee[2] has worked with the SPI provided by the google trends. The weights governed by Lee[2] also supports this fact.

We have come up with a solution to this problem. (To the best of our knowledge, not used before. Write this after some research.)

3.2 Relative Popularity Index and Benchmarking

Comparison between two queries done in google trends shown in fig[1]c also have different normalization constants for different time intervals and geolocations. But RPI defined here is gives proper insight to the volume of queries.
RPIx means RPI of x.

RPIx = (SPI of x when x compared to y)/ (SPI of y when x is compared to y)

From formulae SPI, it results to:
(Relative frequencies of x)/ (Relative frequencies of y)

So two different movies when compared to a common movie, may have different SPIs, but the one with more frequencies will have more RPI . And also RPIx/RPIz = RPIx when compared to z.
So all the hits of a movie are compared with a common benchmark movie, now calculated RPI is the proper index for comparing movie queries.
The movie taken for benchmarking should have very less variance across the time we have considered for all queries or the variance should be such that its average value when compared a query nullify its variance because again there is no absolute concept for variance also. A variance can be much high when compared to rare queries and can be negligible when compared to popular queries.
We have taken benchmark movie as The Shawshank Redemption and benchmark personality as Charlie Chaplin for RPIs of actor and director queries.

Note that a benchmark should not have equal frequencies for all the countries we are taking. We are training a different model for a different country, so the trained weights will adjust themselves as the frequencies and RPI vary. It will vary equally for everything for that country. Though, we the frequencies of the benchmarks are nearly equal among all countries taken here.

3.3 Feature Selection:

We have selected following features to predict earnings of movie among different countries. The selection is supported by results, how important each feature is. The features which does not have any trustworthy data source providing separate data for different countries are not taken into consideration,for e.g. facebook page likes. Rather than generating inconsistencies due to use of common features, we have used features which are separate for each country.
1. Movie Query RPI(M): This is relative popularity index as defined above of a movie query compared with the shawshank redemption. Mostly at the time of release, a movie gets its peak value,it is followed by decay which is 3 to 4 times longer than its rise as shown in figure 1a. There are some cases when there is no proper decay, these are the movies which last long in the theatres e.g. *The Frozen* as shown in figure 1b.In these cases, we have taken the weeks when movie is still in theatres, and neglect further queries.

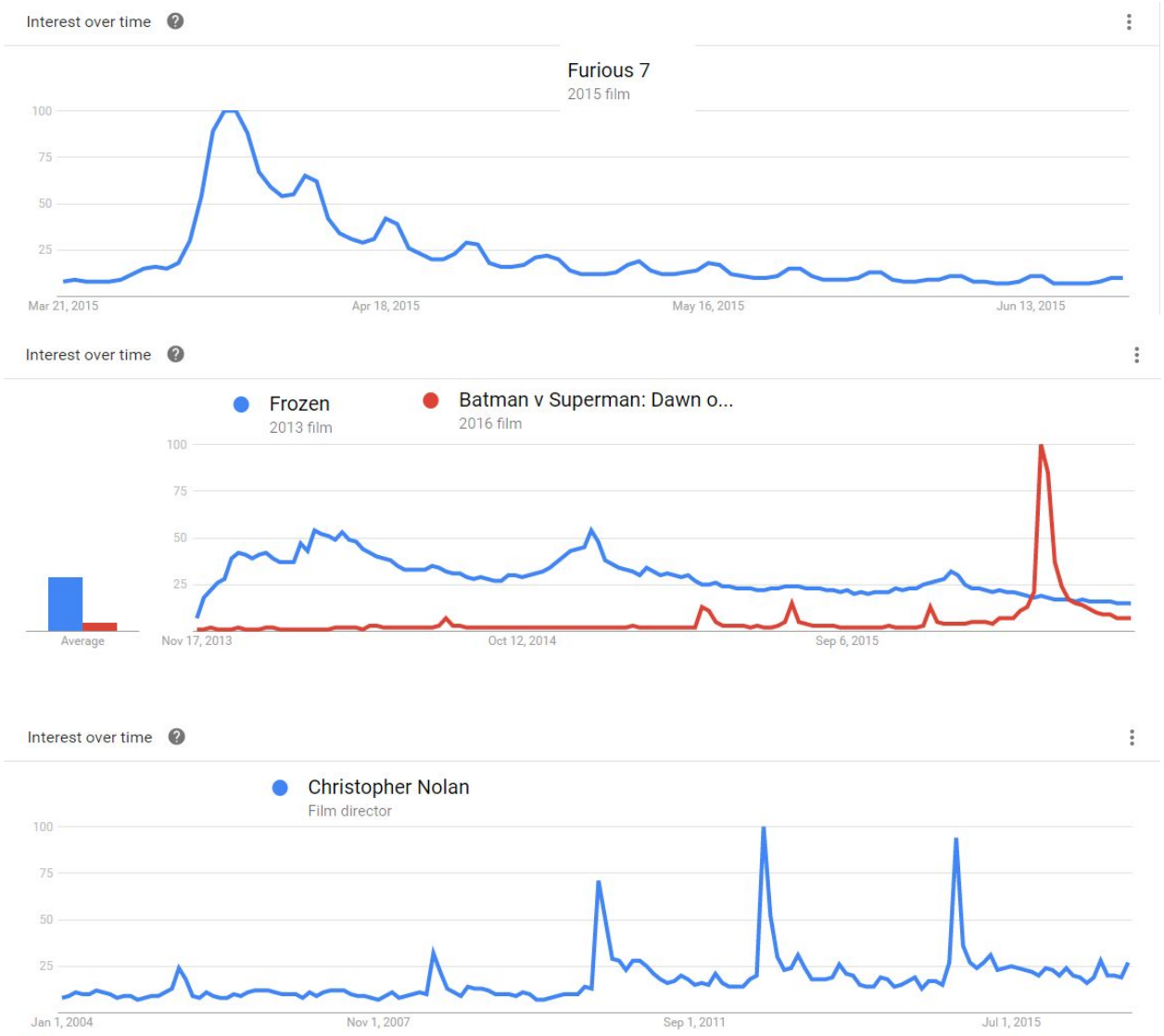


Figure 1:(a) Frequencies of a movie query across a time interval of 4 months,20 days before the movie release and rest afterwards. (b) Shows a comparison between 2 search queries mapped from 0 to 100 with maximum and minimum frequencies of either query being compared. (c)Shows query frequencies for a well known film director. It has a spike whenever his movie is released. There is an increase because of its popularity and increase in number of google users also.

Now each time we are looking at the average values of RPIs for the target country for the decided time interval.
As you can see in figure 1b, popularity of the movie *Batman vs Superman* was nearly double than *The Frozen* but still it has earned 872 millions compared to 1.276 billion of *The Frozen*. The popularity in few months before the movie release affects only the

starting weeks of a movie. The total gross depends more on how long the movie stays in cinemas. *The Frozen* has overcome *Batman Vs Superman* in terms of Total Gross and average.
This is opposite to method used by Lee[2].So a movie query RPI = Average of SPI over time t,where t = min(t where SPIt <=(1/10) Peak SPI Value in that region, t for which movie lasted in theatres).

2. Director Name Query RPI(D): Director name query RPI which is average of only few weeks before the movie release when its RPI starts to rise.
3. Actor #1 and Actor #2 Name Query RPI(A1,A2): Average of RPI of name queries of Actor#1 and Actor#2 as described in the IMDB which is average of only few weeks before the movie release when its RPI starts to rise. There is a rise in the search queries for directors and actors also when their movie gets released which will help in more accurate prediction. As you can see in figure 1c , there is an increase whenever Christopher Nolan releases a movie.Noteworthy point is this curve has peak value at 2012(*The Dark Knight Release*) , before it there wasn't a huge use of google, so it had a peak, but afterwards in 2014(*Interstellar Release*), he wasn't as popular as in 2012, and *interstellar had a lower box office compared to The Dark Knight Rises.* So these attributes contribute in the predictions.
4. Genre#1, Genre#2 and Genre#3 Query RPI(G1,G2,G3): Popularity of genre of a movie in a country contributes in overall success of movie in that country. Average of RPI of queries of Genre#1,#2 ,#3 for 1 year because release of a movie doesn't affect a popularity Genres.
6. Number of theatres released and no. of weeks spent in cinemas(T,W).
Here, note that we have taken each feature and collected data for different countries.

3.4 What to Predict

The prediction from the generated data should show maximum possible correlation to the data and should depend upon the data. Deniz Demir[1] has predicted IMDB movie rating with Google trends. Now, IMDB rating can vary vastly for equally popular movies according to its individual characteristics. So there can be huge variations with different factors like social media and reviews analysis.This data is not proper for that analysis. Nitin[3] has predicted movie success from IMDB data,again that nowhere describes popularity and buzz for a movie.Lee[2] has predicted overall gross of movie with just movie search query of previous months(SPI), rather it can work best for opening weekends only as described in section 3.3.1.

We have collected various data for a movie for countries which are major market for hollywood movies and constantly create more profits. These are US, UK, Australia , France,Germany and Mexico. The data gives us a relative idea of popularity of a movie for a certain time and higher popularity leads to higher average attendance at a cinema.Multiplying average attendance with average price ticket for that country, and multiplying the result with the total number of theatres where movie is released will give us the total amount the movie will earn from the given country. Again,we are training different models for different countries and these multiplications are handled by each model independently. So no need to specify these average ticket prices for each country. Proper thing to predict is the buzz for the movie in people of a country,which as described directly leads to total gross in that country. (So our prediction for each country will be more accurate than overall prediction of Lee[2] which is supported by results.) So a country wise prediction is more robust,reliable and useful for a distributor also. A distributor will put number of theatres in the country where movie will be released, we will predict the nett amount it will earn in the country. So distributors can manage where increasing the number of screenings will make them more money.
Response Variable Movie Income(I): Total gross of movie for each decided country. We have taken natural log(I) to convert the movie income into a scale easy to predict upon.

3.5 Data Characteristics

As described in section 2.3, movie name and related actor, directors queries increase during the release of a movie, but genre popularity doesn't get affected by it. So genre popularity may be very much comparable for a number of movies as there are 3 features for genres. In these cases, features M,T and W will contribute more towards the predictions. Cases when movie names are similar to some regular terms *e.g. The lovers , Home, Interview* can lead to wrong predictions. This will not happen in our case as we are not using a movie name query, we are using a movie as an entity described on google. We look for only those queries where a movie is referenced and shown on the page of google, opposite to the work done by Lee[2].We have done the same for directors, actors and genres. We are taking number of theatres(T) for each country, which contributes a lot in the prediction for those cases when a movie is popular,but has reached to limited screens, which results in lower earnings and happens frequently for lower budget movies.(Try to show it with plot). Number of theatres(T) is a key feature for a movie success, but individual movie characteristics matter a lot when it comes to the total gross. Therefore it can help predict opening weekend gross correctly but to predict total gross, we are taking few weeks after the movie release into consideration as described in section 3.1.1.

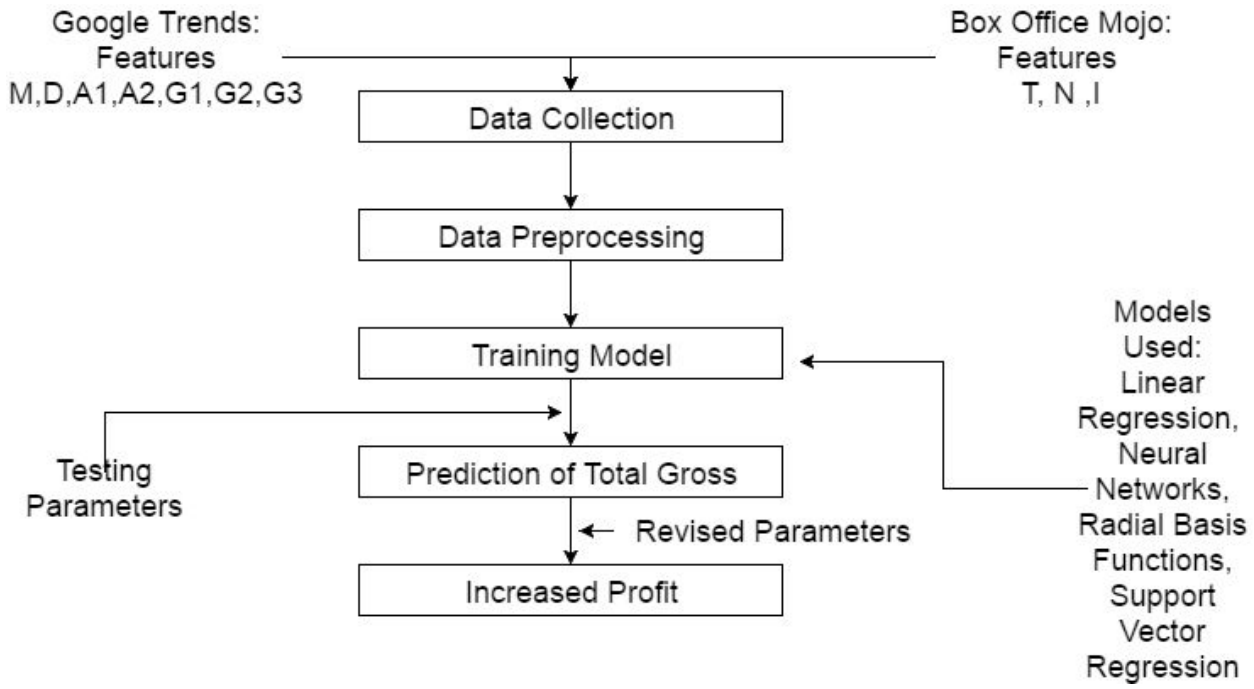
4. OVERALL PROCESS

The overall process consists of:

1.Data Collection: The model uses 8 features among which features M,D,A1,A2,G1,G2,G3 for each movie and for each country are collected from google trends, and then converted RPIs are used in the model. The box office gross of a movie(I) ,number of theatres(T) and number of weeks(W) it lasted,again separately for each country, are taken from box office mojo[8]. All of data is gathered manually so were able to collect data for 112 movies, from 2013 to 2016 so that variations are less.Most of the movies are released after 2014, which decreases anomalies created due to different pricings and all.

2.Data Preprocessing: After collecting raw data, a python script converts into a structured data to feed in the model to be used. If a query doesn't have any entry for our specific country in the csv file from google, then value 1 will be assigned to RPI, which shows equal popularity for a particular query and its benchmark query. The comparison will remove a country from its list only if the country SPI is 0 for both of the queries compared. Entries with SPI value 0 for only a query or its benchmark query are considered to be 0.5 for either case. If a query has less than 0.5 value of SPI, then it is indexed at 0 by google trends. To compare two RPIs these 0 values have to be replaced by a value which indicates relative lower from 0.5 and higher from 0.

3.Training and Prediction: The data after preprocessing is passed through models described in section 3.2 for training. For an unseen movie, features D,A1,A2,G1,G2,G3 can be calculated from google trends. For a movie query, the average before the release and after its rise which is approximately 10-15 days(can vary from movie to movie) is taken as overall average, assuming it will create equal decay after the release.(A lot of movies last longer than expected which can prove our assumption wrong but this is the average case.)



5. PREDICTION METHODS

To predict the function fitting the Income (I) with given features, we have used different models. We have used Linear Regression, Neural Networks (Multilayer Perceptron) and Radial Basis Function Neural Networks to predict the function. As we will describe in conclusions, our data is not linear, so our fitting function is not a linear function.

7. RESULTS

For each country, for each model, we calculated mean squared error and absolute mean error. Prediction will be done by taking exponent of the predicted model output.

		Linear Regression		Multilayer Perceptron		Radial Basis Functions	
		MSE	AME	MSE	AME	MSE	AME
US	Training	0.656006755	0.706006755	0.675944392	0.724504018	0.598553145	0.610811212
	Test	0.806763555	0.796743456	0.804633318	0.971971723	0.768878446	0.932416534
UK	Training	0.842786326	1.321880363	0.85863122	1.384144925	0.754562831	1.183944255
	Test	0.986593115	1.337418571	0.84043669	1.036522114	0.796157591	0.931764797
Germany	Training	0.86457465	1.262280643	0.847115928	1.354640248	0.775404355	1.223634987
	Test	0.957412408	1.338742732	0.783195942	0.991949837	0.738259418	0.886560773
Australia	Training	0.694468862	0.977721815	0.697060679	0.996316943	0.615959965	0.888159089
	Test	0.844290305	1.169708347	0.974795286	1.798722497	0.951269969	1.704012737
France	Training	0.694468862	0.977721815	1.047648168	1.852270297	0.901713737	1.466307775
	Test	0.844290305	1.169708347	0.877028686	1.562037831	0.85917394	1.477809933

Table 1: MSE and AME values for all the models of each country.

As we can see from the table above, in most of the cases, radial basis function works best. If data have a larger linear fraction than linear regression will also give better results, as happened in US. Test data is much smaller so have some higher values, taking larger test sets will decrease test errors also.

Following graph is of movie query RPI(M) vs log(Gross)(I).

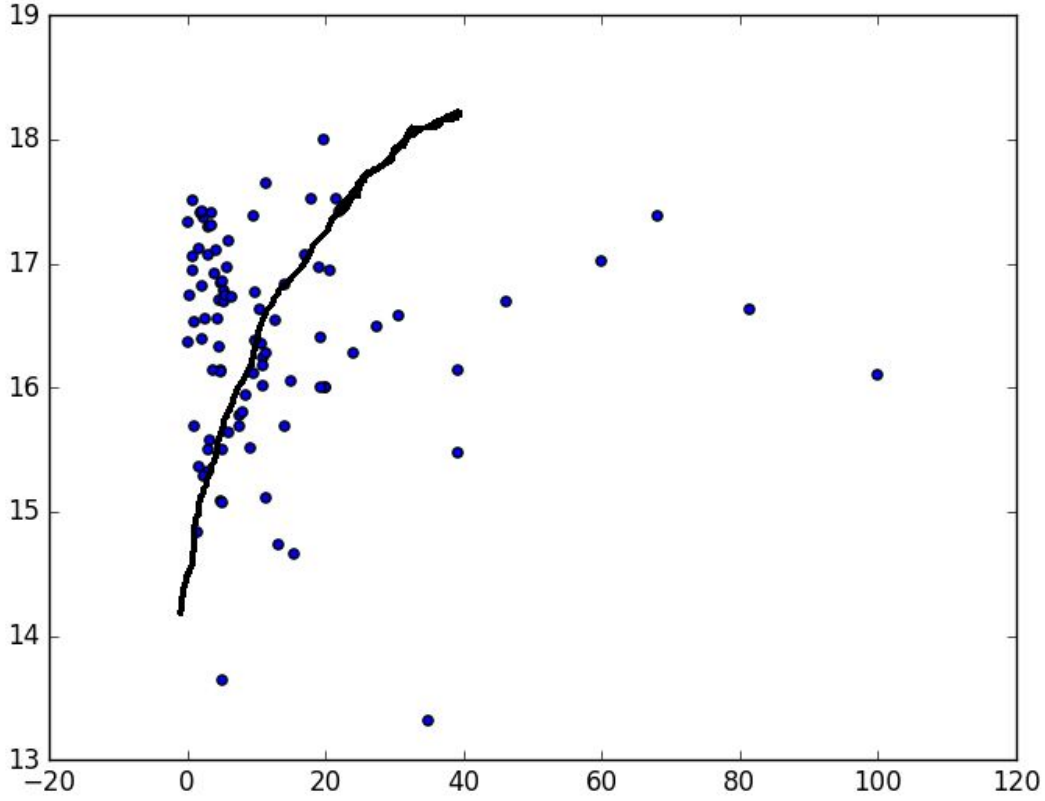
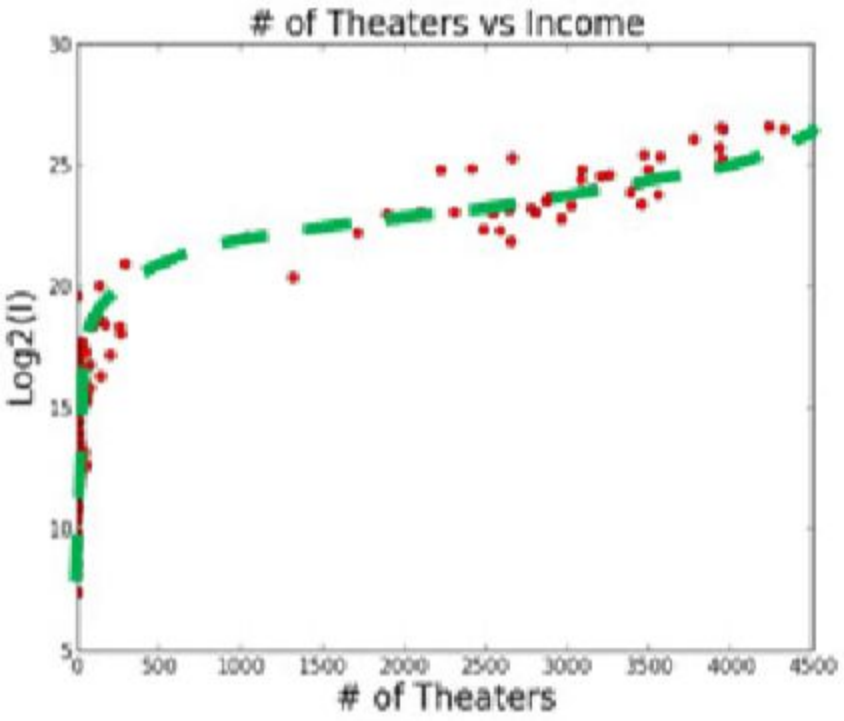


Figure 3a: Movie query RPI(M) plotted against log(Total Gross)(I)

8. CONCLUSIONS

Features used by us for the prediction have different impact over different countries. HIghest weightage value for a movie queries was in Australia whereas highest director weightage value was for USA. So we can comfortably say that people in USA are more involved in who has directed a movie, same with actor #2 also.It also has highest weightage in USA. We analysed that number of theatres affect the total gross directly with minor trend shifts. So overall the features developed by use represents the function very well.

In future, we will try to implement Support Vector Regression and Random Forest Regression also. We will try to include number of weeks in the feature directly and introduce some common factors and see how well they describe the data.



9. REFERENCES

[1] Deniz Demir, Olga Kapralova, Hongze Lai, "Predicting IMDB Movie Ratings Using Google Trends," Dept.Elect.Eng,Stanford Univ., California, December, 2012

[2]PREDICTING MOVIE SUCCESS FROM SEARCH QUERY USING SUPPORT VECTOR REGRESSION METHOD
Chanseung Lee1 and Mina Jung2

[3] Predicting Movie Success Based on IMDB Data Nithin VR, Pranav M, Sarath Babu PB, International Journal of Data Mining Techniques and Applications
ISSN: 2278-2419

[4] Choi, H., Varian, H., (2012). Predicting the present with google trends. Economic Record, 88 (s1), 2–9