FAKULTÄT
FÜR !NFORMATIK

Faculty of Informatics

Predicting Movie Success

with Machine Learning and Visual Analytics

BACHELORARBEIT

zur Erlangung des akademischen Grades

Bachelor of Science

im Rahmen des Studiums

Medieninformatik und Visual Computing

eingereicht von

Philipp Omenitsch
Matrikelnummer 1025659

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Mag. Dr. Silvia Miksch
Mitwirkung: Bilal Alsallakh, M.Sc

Wien, 26.02.2014

(Unterschrift Verfasser)

(Unterschrift Betreuung)

FAKULTÄT
FÜR !NFORMATIK

Faculty of Informatics

Predicting Movie Success

with Machine Learning and Visual Analytics

BACHELOR'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Bachelor of Science

in

Media Informatics and Visual Computing

by

Philipp Omenitsch
Registration Number 1025659

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Ao.Univ.Prof. Mag. Dr. Silvia Miksch
Assistance: Bilal Alsallakh, M.Sc

Vienna, 26.02.2014 ________________________ ________________________
(Signature of Author) (Signature of Advisor)

Acknowledgements

I want to thank my advisor Bilal Alsallakh for your endless patience and commitment. Also, thank you Markus Bögl for your help with statistics problems.

Abstract

Predicting a movie's opening success is a difficult problem, since it does not always depend on its quality only. External factors such as competing movies, time of the year and even weather influence the success as these factors impact the BoxOffice sales for the moving opening. Nevertheless, predicting a movie's opening success in terms of BoxOffice ticket sales is essential for a movie studio, in order to plan its cost and make the work profitable. I introduce a simple solution for predicting movie success in terms of financial success and viewer recipience. As a result, this approach achieved decent estimations, allowing theatre planning to a certain extent, even for small studios.

Contents

CHAPTER 1

Introduction

Predicting the outcome of events and the success of products is a fundamental problem in data mining and predictive analytics. A variety of techniques have been proposed to address real-world prediction problems arising in different domains. In this work, I address the problem of predicting movie success based on two indicators:

- Boxoffice income: the gross revenue of the movie combined for all theatres showing the movie on the opening weekend
- Average Rating: the rating users provided on the Internet Movie database after the opening weekend

Many models have been proposed in order to predict viewer ratings and box office incomes, for example with the help of social media [Oghina et al., 2012] or news analysis [Zhang and Skiena, 2009]. The success of a movie can be measured by many different aspects. The main criteria tough are quality, how the audience liked the movie and box office, as in economical success.

Movie success prediction has a lot of use for companies to plan their resources. For example, a Hollywood studio, that expects its newest movie to be highly successful will rent more theatre rooms in advance, increasing revenue if the prediction turns out to be true. If it rent less theatre rooms, not all viewers might have been able to watch the movie in its opening weekend.

Before the advent of the internet, critics published their opinions on the quality of a movie, so a movie would be broadly rated on the opinions of a small elite educated audience or polling the audience after the movie. The internet not only changed the way we consume movies but also how movie quality is determined. Thanks to the IMDb (Internet Movie Database) [IMDb, 2014], one of the first online movie databases and the biggest today, it is possible for everyone to rate movies, the positive effect being it is more or less anonymous and statistically more significant, because of the bigger sample size (currently 45 mio registered users).

The VAST (Visual Analytics Science and Technology) Challenge 2013 Mini-Challenge 1 [Cook et al., 2013] was motivation to tackle the problem of movie success prediction with the help of a different approach, Visual Analytics, in order to get more insight into complex data structures. Visual Analytics is an emerging field of computer science, intertwining the areas of visualization and and analytical methods (like, statistics, machine learning), in order to create deeper insights into datasets. Often standard machine learning models cannot represent complex data well enough to make good predictions, also it can be a very difficult task to identify important features. Visual Analytics can help by first visualizing the data to examine trends and determine good features for the later prediction step. Therefore, it is necessary to have a human in the loop, who obtains information about connections of various aspects of the data by first visualizing and then analysing them.

Within the challenge, the success of a movie is measured by two indicators, box office income and viewer rating on IMDb. The challenge participants were allowed to use data from the IMDb API [IMDb, 2014] as well as the data from relevant tweets via Twitter and click-through counts of the popular URL shortener service bitly. Submission for predictions was always due one day prior to the starting weekend every weekend for the time window between 11th of January 2013 and 26th of July 2013. During the challenge teams did receive continuous feedback and recognitions from the challenge committee. At the end of the challenge best teams receive awards based on their submitted predictions. The award ceremony was held during the IEEE VIS conference in Atlanta, Georgia USA which took place in October 2013 [IEEE VIS, 2013]. In order to give the reader a better understanding of the weekly submitted reports, I provide a sample submission on the next page.

White House Down

Philipp Omenitsch, Bilal Alsallakh, Marcus Bögl TU Wien, Austria

Abstract— A short description of our movie prediction for “White House Down”

Results

Predicted Rating: 6.3
Predicted Box office: 21.000.000$

Data

White House Down (2013)
Release-Date: 2013-06-28
Genres: Action Drama Thriller
Director: Roland Emmerich (average rating 6,48)

Actors	Channing Tatum	Jamie Foxx	Maggie Gyllenhaal	Jason Clarke	Richard Jenkins
Ratings	6,26	6,59	6,66	6,26	6,28
Billing position	1	2	3	4	5

Expected calculated Ratings

We computed a weighted sum of the avg. ratings for each actor and director where newer movies have a higher weight . These ratings and the log of the estimated number of ratings are taken as input for a simple 1-layer neural network which calculates a rating for the movie to predict.

Result: 6.68

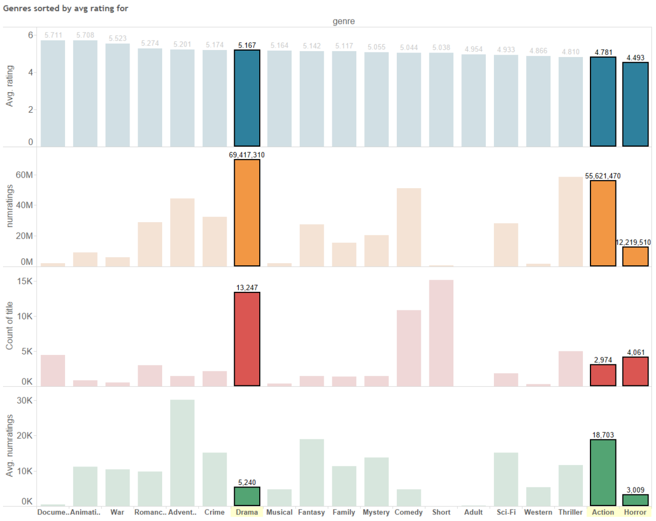


Fig. 1. Genres for White House Down highlighted

Conclusion for Rating

The genres speak for a less than average rating, including 2 of the worst ranked genres for ratings. That is why we lower the predicted rating from our neural network by .4

Final Rating: 6.28

2 Box office

2.1 Twitter

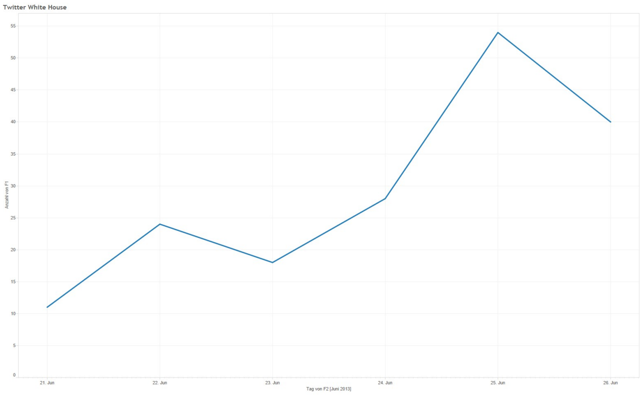


Fig. 2. twitter trend 4900 tweets (data incomplete)

2.2 Conclusion

The current calender week suggest a strong weekend, because schools are slowly closing for the summer, increasing the available audience greatly averaging 15M for an opening weekend in the last 13 years. The movie combines 3 genres with the absolute most number of ratings for single genres, showing the huge market potential the movie has. Twitter shows less than average anticipated engagement, being a bad short term indicator

Predicted Box office: 21.000.000$

CHAPTER 2

Related Work

A large interest in the analysis and prediction of movie success started when NetFlix announced the NetFlix Prize [Netflix, 2005] in 2006, which offered 1.000.000 $ to the best team for the imporvement of their movie rating algorithm Cinematch. NetFlix is a big online streaming platform, where users pay a monthly fee to watch movies and series on demand. They are interested to keep the user as long as possible watching movies, therefore it is crucial to provide each user with good recommendations on what to watch next after he finished watching a movie, based on the user's taste and of course on the quality of the movies. The challenge lasted until 2011 and was very successful. At first, around 2005, a lot of groups tried to infer movie success from static attributes of the movies, such as actors, directors, genre or plot. Later on with the advent of social media, more and more analysis of the connections between social media and movie success was performed [Oghina et al., 2012]. One can use social media trends but also other sources, like Google search trends, to predict movie success to a certain extent.

The features used and proposed by other groups [Oghina et al., 2012, El Assady et al., 2013, Jäger et al., 2013, Yafeng Lu and Maciejewski, 2013] can roughly be separated into two groups - movie intrinsic features and all external features. To the first group belong features, such as the actors, directors, genre of the movie and so on, which determine the movie itself and can be brought into correlation with it's quality.

The second group consists of external features and real world conditions, which do not influence the movie itself, but can massively do so for its reception and popularity, hence the revenue. These features include the release date, time of the year and other movie releases at the same time, the weather conditions, marketing campaign, audience movie anticipation and many more.

A lot of related work has been done by the other participants of the 2013 VAST Mini-Challenge 1. The goal of many approaches is, to combine automated prediction tools, i.e. machine learning, with domain specific knowledge of the analyst. 'This is a smart way to improve accuracy, because machine learning models tend to be too general while analysis with the help of visualizations can be highly subjective' [Jäger et al., 2013] p.2 6-8. Thus, combining them tries to use the strengths of the approaches, resulting in better predictions.

2.1 MooVis

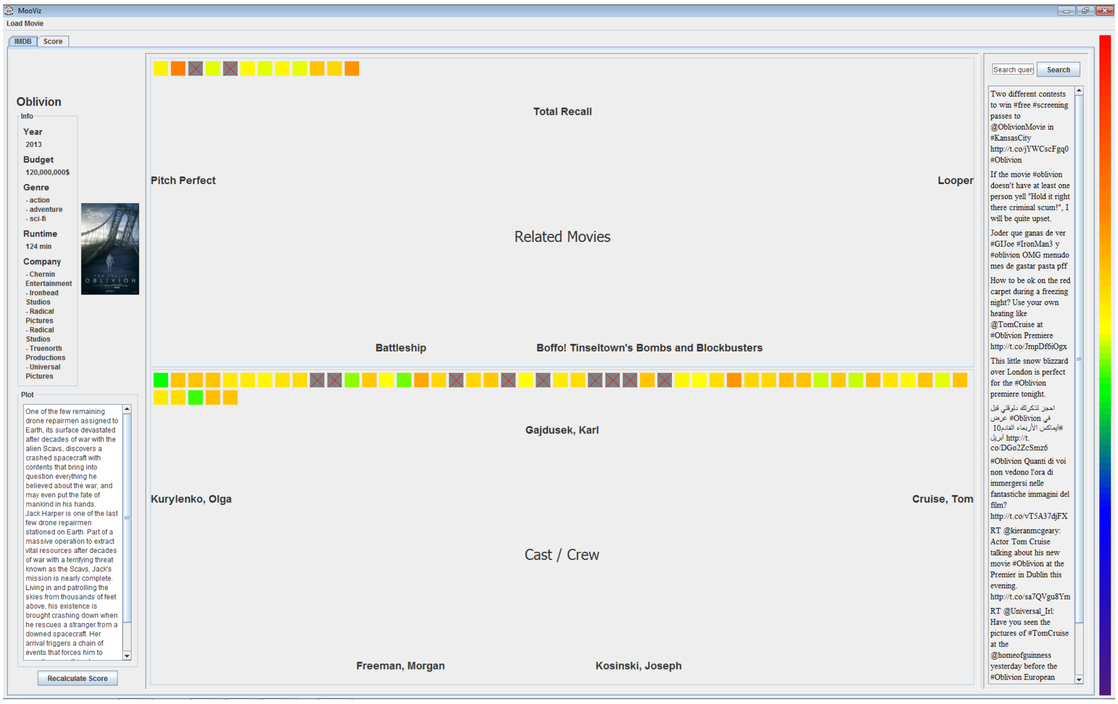
MooVis [Jäger et al., 2013] makes their predictions by letting the analyst choose related movies, supported by historical IMDb data and twitter sentiment analysis. The rating can then be predicted with the help of a neural network which only takes into account ratings of crew members and actors of relevant movies. They use two tools which were developed in the previous works:

(1) VISONE (Visual social networks) is a tool to analyse and visualize graph structures of social networks.

(2) KNIME (Konstanz Information Miner) which contains different data mining techniques and the ability to export trained models in the PMML (Predictive Model Markup Language) format.

When starting the program the user is presented with a general overview about the movie as can be seen in 2.1

Figure 2.1: Main screen of MooVis analysis tool



Oblivion

Info

Year
2013

Budget
120,000,000$

Genre
- action
- adventure
- sci-fi

Runtime
124 min

Company
- Chernin Entertainment
- Ironhead Studios
- Radical Pictures
- Radical Studios
- Truenorth Productions
- Universal Pictures

Plot

One of the few remaining drone repairmen assigned to Earth, its surface devastated after decades of war with the alien Scavs, discovers a crashed spacecraft with contents that bring into question everything he believed about the war, and may even put the fate of mankind in his hands.
Jack Harper is one of the last few drone repairmen stationed on Earth. Part of a massive operation to extract vital resources after decades of war with a terrifying threat known as the Scavs, Jack's mission is nearly complete. Living in and patrolling the skies from thousands of feet above, his existence is brought crashing down when he rescues a stranger from a downed spacecraft. Her arrival triggers a chain of events that forces him to

Two different contests to win #free #screening passes to @OblivionMovie in #KansasCity http://t.co/jYWCscFgq0 #Oblivion

If the movie #oblivion doesn't have at least one person yell "Hold it right there criminal scum!", I will be quite upset.

Joder que ganas de ver #GIJoe #IronMan3 y #oblivion OMG menudo mes de gastar pasta pff

How to be ok on the red carpet during a freezing night? Use your own heating like @TomCruise at #Oblivion Premiere http://t.co/JmpDf6iOgx

This little snow blizzard over London is perfect for the #Oblivion premiere tonight.

احجز تذكرتك دلوقتي قبل عرض #Oblivion في #أيماكس الأربعاء القادم10 أبريل http://t.co/DGo2ZcSmz6

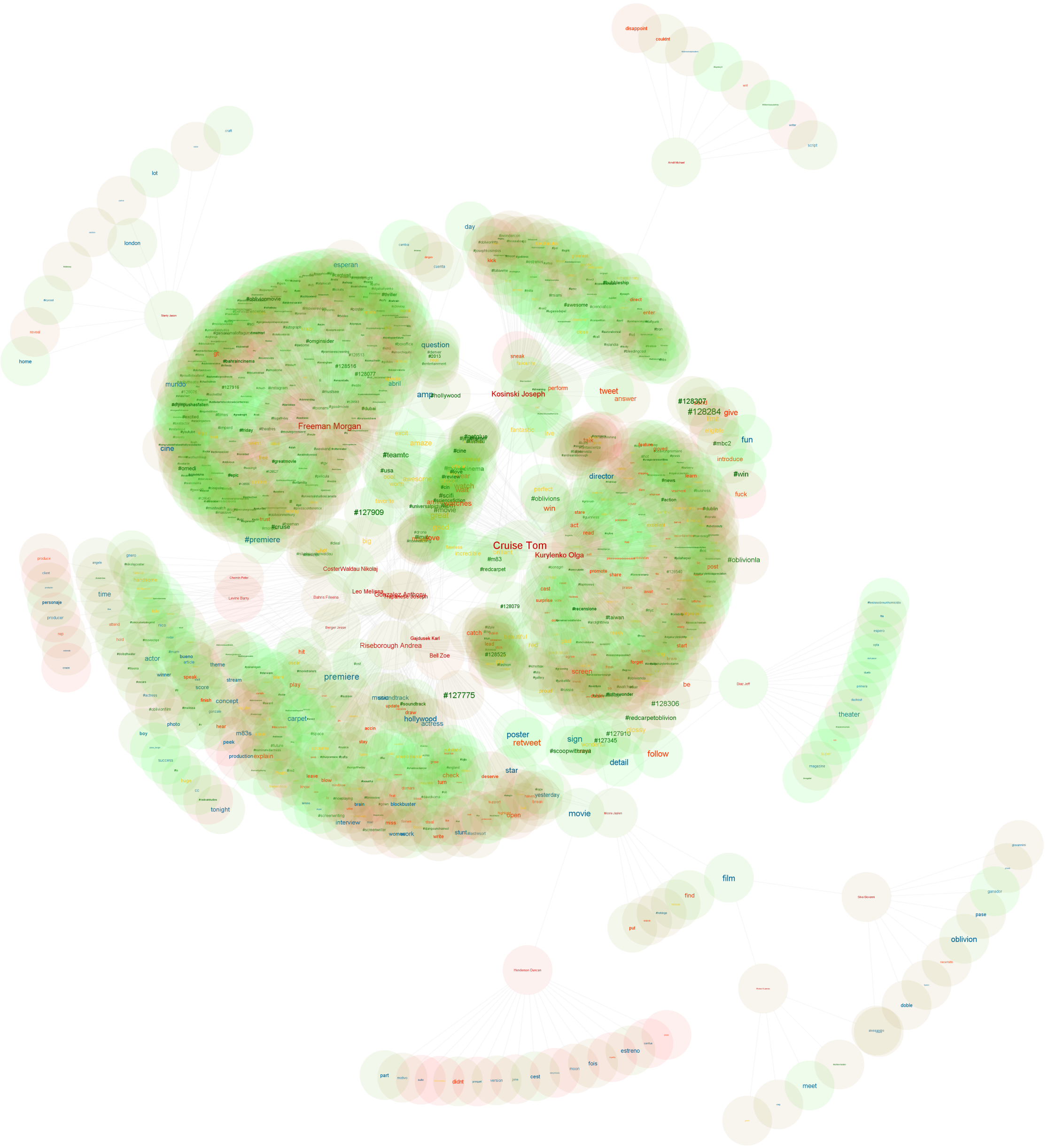
#Oblivion Quanti di voi non vedono l'ora di immergersi nelle fantastiche immagini del film? http://t.co/vT5A37djFX

RT @kieranmcgeary: Actor Tom Cruise talking about his new movie #Oblivion at the Premier in Dublin this evening. http://t.co/sa7QVgu8Ym

RT @Universal_Irl: Have you seen the pictures of #TomCruise at the @homeofguinness yesterday before the #Oblivion European

In the left panel general movie intrinsic features are displayed for the analyst. In the middle related movies are displayed, the coloured tiles represent the average rating of related movies (top) as well as the average rating for crew members in which they stared (mid). Later on the analyst can reorder, add or remove related movies and weight them for future prediction with neural networks. Another very important part consists of the sentiment analysis of the tweet data. Tweets and their hashtags are shown regarding cast members and the sentiments they are shown with, this can be seen in figure 2.2. The green areas show positive, red negative sentiments. More important casting crew members can also be found in the graph representation

Figure 2.2: Twitter sentiment analysis with MooVis for movie 'Oblivion'



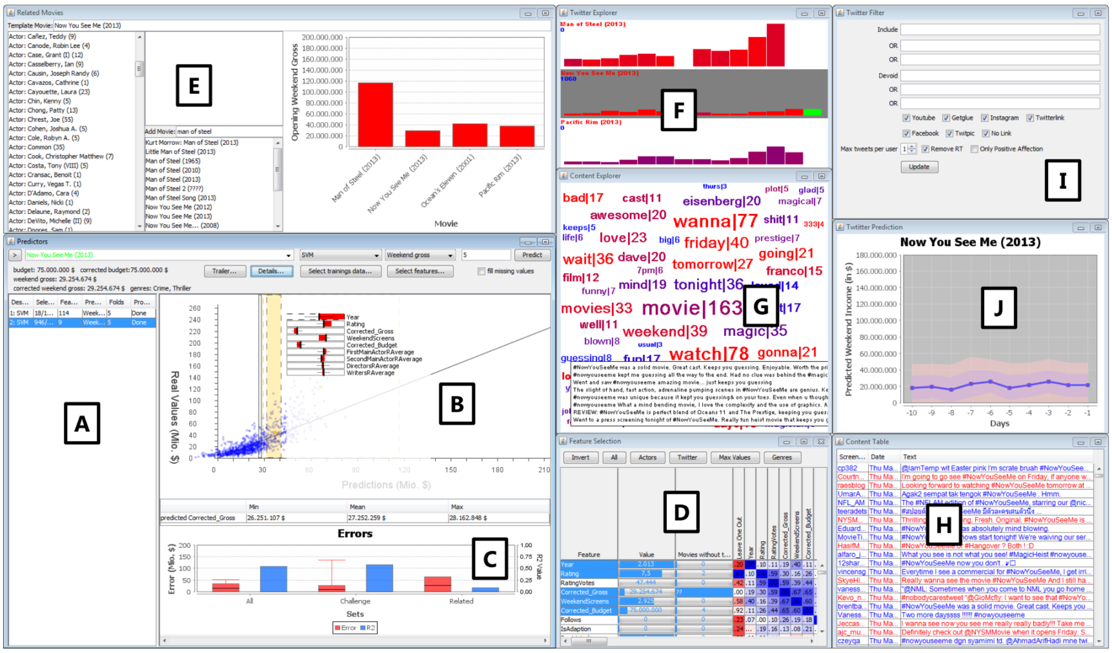
and used for sorting crew members, also for later machine learning algorithms, similar as for related movies.

Finally, 7 different predictions are calculated for the movie rating as well as the box office. The analyst then has to weight and combine them to reach the final prediction values.

2.2 Team Prolix

Team Prolix [El Assady et al., 2013] of the University of Stuttgart uses an iterative approach so analysts could combine features from structured and unstructured data. The analyst can select features, which s/he thinks will be most important for the movie to predict and also a prediction model (Support Vector Machine, Multi Layer Perceptron or a logistic regression model). Furthermore, s/he gets feedback via error measures by predicting historical data with the currently selected model. Once again Twitter data is used to explore features via sentiment analysis.

Figure 2.3: System Overview for Team Prolix: The views A to J provide complementary information about movie prediction such as details about the automated predictor (A to C), Feature Selection (D), related movies (E), and information about related Twitter messages (F to J).

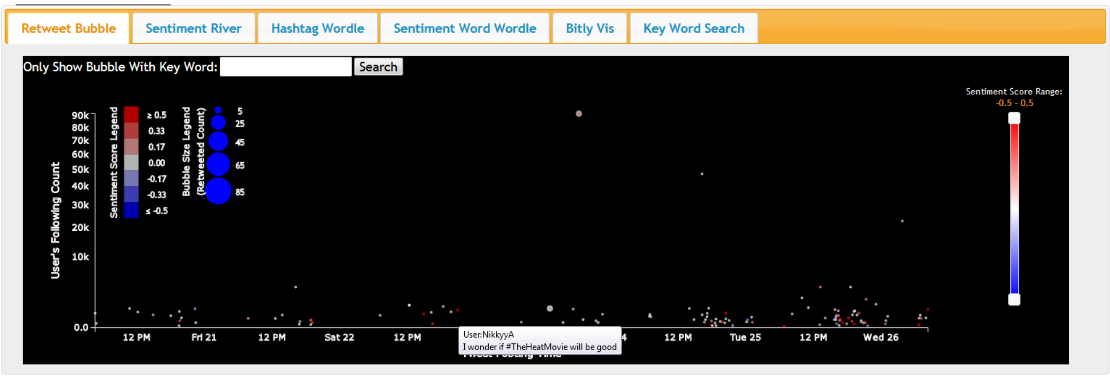


Their visualization tool as seen in the figure 2.3 consists of an UI that is loaded with useful information, everything can be seen at one glance. On the left-hand side static movie related data can be found as well as the selected features and information for similar movies. On the right hand side Twitter analysis can be done. The iterative approach they chose, similar to the one I am proposing, makes it necessary to have all data at one sight so one can see how adjusting one variable affects the others.

2.3 Team VADER

For box office revenue prediction Team VADER [Yafeng Lu and Maciejewski, 2013] uses a linear regression model with 2 variables, budget and number of daily tweets averaged over a two week period prior to the movies release. For training they use a very small training set, movies from January 2013 leading up to the release week of a movie. This has a great advantage, that these movies also already include the exact box office numbers from this year and thus are very representative. Team VADER's approach for predicting the review score of a movie mainly relied on twitter data. The team uses SentiWordNet 3.0 [Baccianella et al., 2010] to calculate positive and negative sentiment scores for tweets. The representation of them can be seen in figure 2.4. Furthermore, their approach relies heavily on similarity visualization. Therefore, they have 9 features which are shown for similar movies. This helps the analyst estimate the quality of their prediction to a big extent.

Figure 2.4: Team VADER Tweet Sentiment Correction



User:NikkyyA
I wonder if #TheHeatMovie will be good

Generally one can say there are three main types of approaches predicting movies with Visual Analytics:

(1) Similarity-driven prediction, this means trying to match the movie to predict with other movies and extract information about the rating and the box office from these historic data [Al-Masoudi et al., 2013]. These features are then used to train a machine learning algorithm and predict movies with help of the algorithm.

(2) Visual exploration of the data to gain complex insights in the data structure. These insights are used to select features and again a machine learning algorithm is applied. Furthermore predictions will be modified most of the time based on the opinion of the analyst [Mat Kelly, 2013] [Fazzion et al., 2013].

(3) Pure Visual Analysis by applying domain specific knowledge to select the most important features for a movie. This approach can be problematic, because it is very subjective and relies on a great experience of the analyst, as well as on the analyst being well informed. On the other hand, it gives a lot of flexibility and a better chance to predict outliers. [Perin, 2013]

2.4 Summary

The advantages of Visual Analytics are obvious, the specifics of each movie can hardly be covered by a simple machine learning algorithm, therefore Visual Analytics greatly improves the probability to detect outliers. This comes at the price of more work for the analyst and him being experienced. My approach can be seen as a combination of similarity driven analysis, visual exploration and Visual Analytics. It strongly depends on the neural network to build a foundation for the further analysis, based on similar actor and director ratings from other movies. After the first estimate is obtained, the analyst refines the estimation with help of Visual Analytics of social media and historical movie data. How this is done will be discussed in more detail in the next section3.

CHAPTER 3

Approach

Analysing and predicting movie success is a task that can be very complex, because of the many external influences on a movie. Thus it is important to start with a simple prediction model. Visual Analytics can help a lot with this, because visualizing data is a good way to understand the most important relationships within them. My approach includes basic movie prediction methods, looking at historic movie data, i.e. actors and directors, as well as social media analysis. First, I will explain how different data were processed, then I will give an overview over the workflow of the interactive iterative visual analysis itself.

3.1 Data Acquisition

I used all three data sources that were allowed by the challenge, namely the IMDb, related tweets, and Bitly link statistics.

- The IMDb has the most complete historic movie data, that is freely available [IMDb, 2014]. It includes over 2,5 million titles and 5 million people and their roles in movies, as well as release dates and ratings. In order to import data from the plain text IMDb files, I extended software written by Kosara et al. for the InfoVis 2007 contest [Kosara, 2007]. In order to use the software for my purposes, I had to fix some bugs in the software first. I filtered the data and only included movies in the data set that were released after the year 2000 and from the US, trying to improve homogeneity. I also implemented functionality to export the data as comma separated values and imported them into a MySQL database for further refining. I wrote some scripts for the database to further enhance quality of the data and construct a weighted rating average for actors and directors.

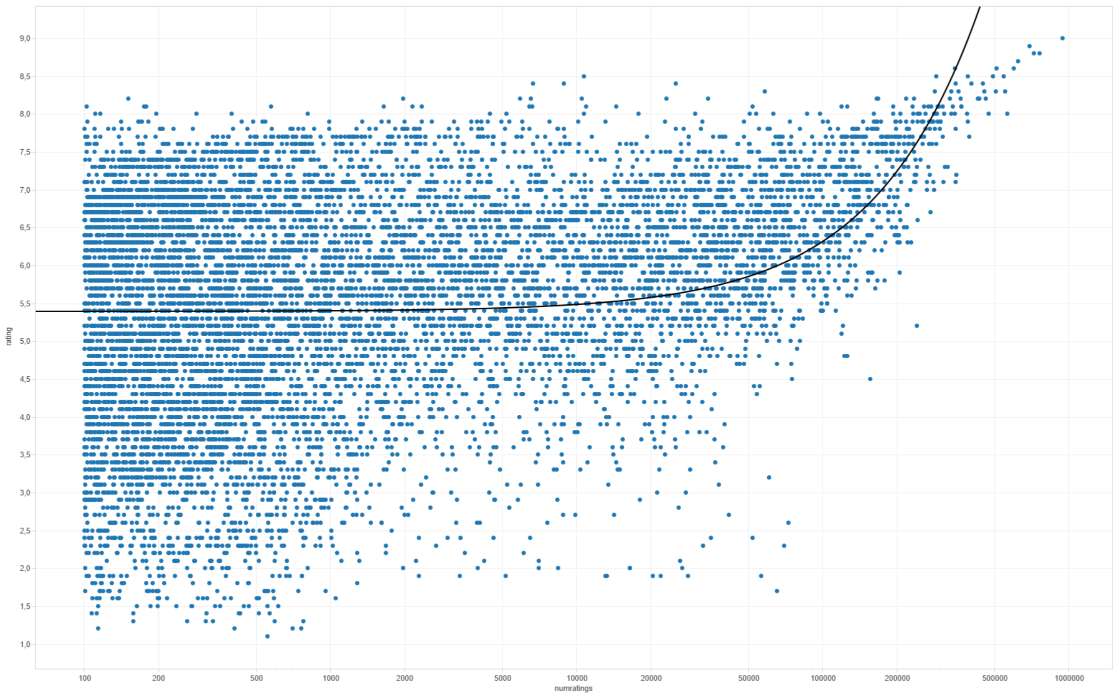
- Due to restricted resources and the Twitter API terms, I could only analyse a small sample of the provided twitter data. The challenge provided a set of tweet ids, with which one can download the tweets via twitter API, unfortunately this is rate limited to 180 tweets per hour.

- As for Bitly data only the number of clicks per link were available in form of a daily updated CSV (Comma Separated Value) file

3.2 Data Analysis

For visualization I used the software Tableau [Tableau Software, 2014] also via MySQL connection. Visualizing the data in various ways helped me identify important features and correlations, how exactly I will show later. Tableau is a very powerful tool, because it can handle huge data sets (the set I worked with included approximately 35000 movies) and display them in many different ways.

Figure 3.1: The relation between the number of ratings and the actual viewer ratings for a movie.



Neural networks

Neural networks are one class of machine learning algorithms. They can model complex nonlinear behaviour at the cost of many parameters (weights). This is the reason, why one has to use them as black boxes and most of the time cannot fully understand their world model. One more good thing about neural networks is that they cannot only predict discrete class labels but continuous ones as well.

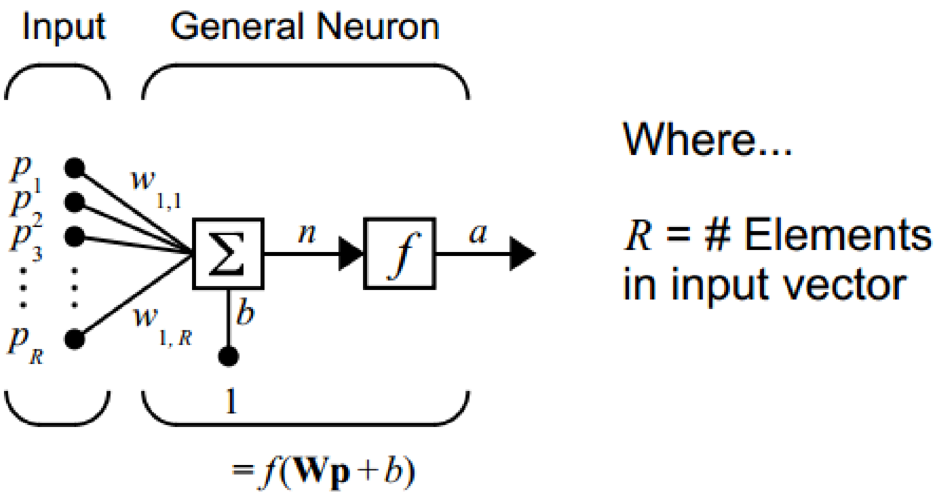
’Neural networks are composed of simple elements operating in parallel. These elements are inspired by biological nervous systems. As in nature, the network function is determined largely by the connections between elements. We can train a neural network to perform a particular function by adjusting the values of the connections (weights) between elements.' [Demuth and Beale, 1993, ch. 1, p. 2]

Usually neural networks are composed of multiple neurons. One neuron has multiple inputs which represent the features and one output representing the target value (e.g. class / predicted value). The task of the neuron now is, to find a weight *w* for each input *p*, so the output is produced. More precisely this is done by multiplying all weights *w*

$$bias + \sum_{0<i<n} w(i) * p(i) \tag{3.1}$$

with all inputs *p* and summing up these and the bias, a static term (often 1). The transfer function takes this sum and produces the output. The transfer function is used to obtain a normalized output, normally between -1 and 1, although many different transfer functions exist [Demuth and Beale, 1993, ch. 2, p. 4]. In order for the neurons to gain their weights, a neural network needs to be trained, so each neuron can obtain its weights. There are multiple algorithms for training neural networks, the backpropagation algorithm being the most popular. To model an even more complex behaviour, neural networks can also consist of multiple layers, so that the outputs of one layer of neurons are the input for the next, higher level of neurons.

Figure 3.2: One neuron with multiple inputs [Demuth and Beale, 1993, ch. 5, p. 3]



To keep my model simple and understandable, I used only one neuron and backpropagation for training. The model can only represent linear behaviour, thus after training I obtain the weights for my model.

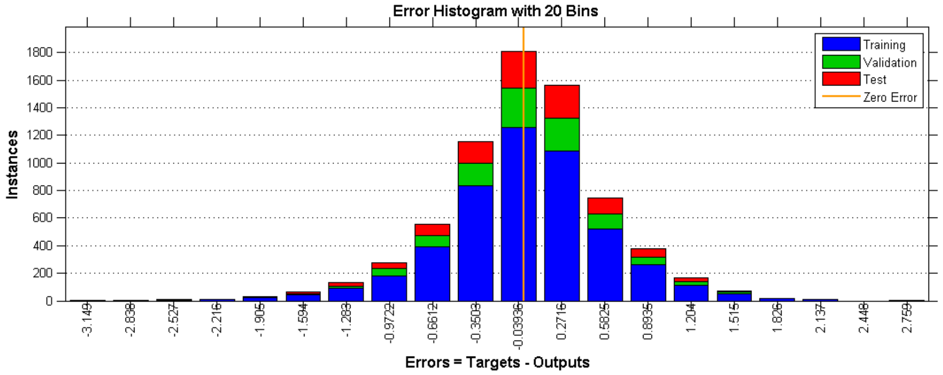
3.3 Viewer Rating

On the IMDb website users can vote for each movie and rate it from 1(very bad) to 10 (very good). They can vote from 1 very bad, to 10 very good. The prediction of the viewer rating depends on multiple features. In the end these feature combination gave the best results.

- Rating(1-10): the rating of a movie determined by IMDb users, this is the target.
- Number of ratings: this shows two things, first, how statistically trustworthy the rating is, second, roughly how many people watched the movie. After visualizing the number of ratings and the ratings see figure 3.1, it became clear, that the number of ratings n correlates with the rating, even more for movies with $(n \geq 50,000)$. This is also the reason why I took the log10 of the number of ratings, because only the order of magnitude is important. I can only speculate why they correlate, but my first guess would be, that film studios tend to know quite well how to satisfy their target customers.
- Average director rating: I calculated the average rating of all the movies, the director directed before. I also did weigh new movies stronger than old ones for this average, this tries to emphasize recent movies, coupled with the hypothesis, that recent movies have bigger impact on the popularity of a director. The weights were tuned iteratively by hand, using the error histogram to minimize error.
- Average actor rating: The weighting works the same way as with directors, but I chose the first 5 actors in the billboard listing of a movie, most of the time these also are the main characters of the movie. The reason for taking the top 5 actors was because this gave us the best results.

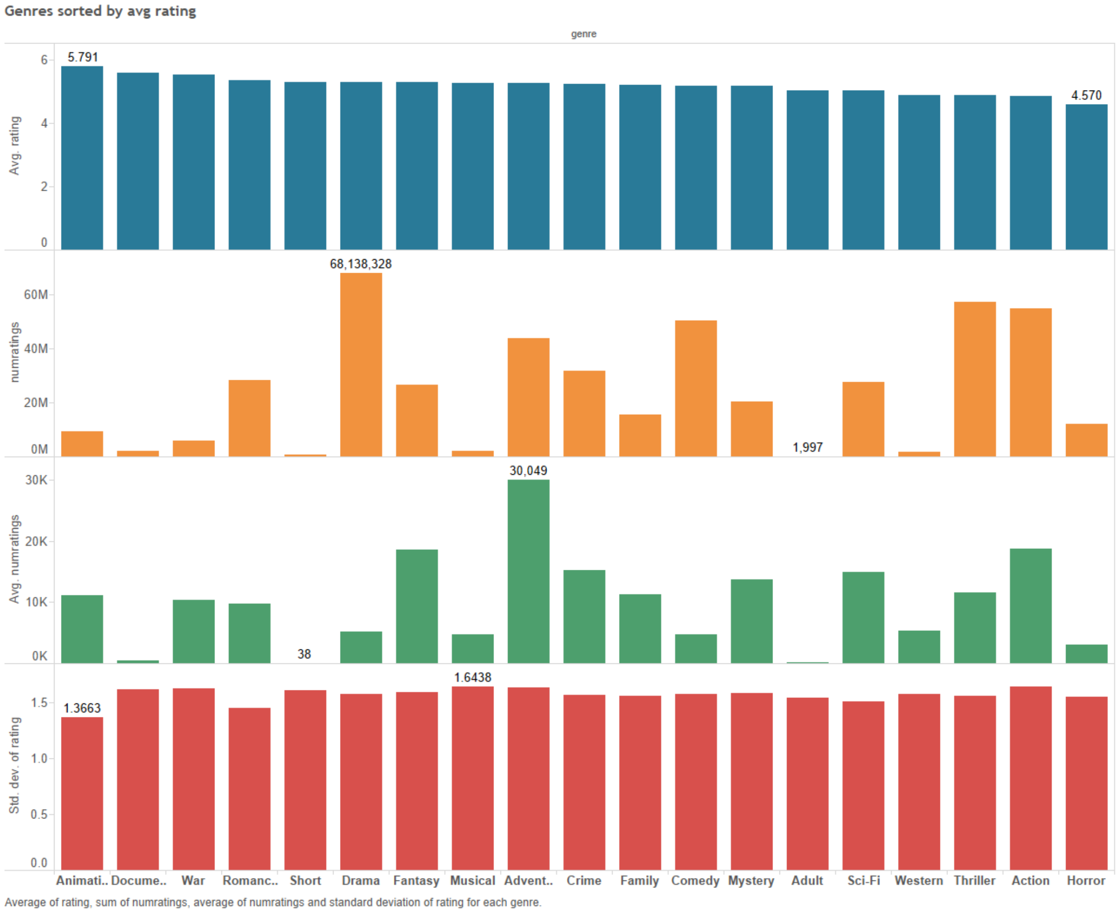
For this to work, I first had to estimate the number of ratings, a rough estimate was enough, because only the order of magnitude is needed. This approach worked quite well, as can be seen in the error histogram in figure 3.3. This simple machine learning approach gave a really good first estimate for the viewer rating to work with.

Figure 3.3: The error histogram for the trained neural network to predict viewer rating



Although this model gives quite low error rates, there is still room for improvement. So I decided to manually adjust the rating, with the help of visualizations. Especially, analysing the genres of a movie can greatly improve prediction accuracy, see figure. The reason why this cannot be easily incorporated into the neural network is that a movie can combine multiple genres, making it hard to train the neural network properly. A small amount of movies might belong to one combination, while another combination includes a lot more movies.

Figure 3.4: Avg. rating, number of ratings, average number of ratings and std. dev. of rating for each genre


Genres sorted by avg rating

Average of rating, sum of numratings, average of numratings and standard deviation of rating for each genre.

The information and order of genres will also change significantly when selecting only movies with the number of ratings being over a certain threshold, one can filter out a lot of noise. There are a lot of rated low budget movies, but since the predictions almost always included big Hollywood blockbusters, low budget movies can easily be ignored. I would therefore set the threshold to the lower boundary of number of ratings I thought the movie would achieve. After all this, the predicted rating was corrected a little bit, normally not more than +/- 0.5 points, based on how high the genres of the currently predicted movie were ranked (see figure 3.4). One has to bear in mind, that this method is highly subjective and thus very dependent on the analyst.

The base of the prediction always relies on hard facts and machine learning, while I manually tried to correct errors induced by outliers.

3.4 Box office

Almost always the turnover resulting from ticket sales at the theatres is referred to as box office. It is widely known that 25 percent of the movies turnover is generated on its opening weekend [Simonoff and Sparrow, 2000] p.1, which is therefore a crucial indicator of the movie success. Big movie studios can spend up to 300 mio dollars in the making and marketing of a movie [Numbers, 2013]. Having an box office estimate is therefore essential to the survival of a studio. The box office has a lot of external dependencies and especially on the opening weekend tends to be more important than the quality of the movie itself. Popular directors and actors tend to draw people to go to the movies they participate in, but there is no easy way to see popularity of i.e. an actor from the provided IMDb data. Therefore, I mainly focused on the external dependencies, them being the release date, budget, number of theatres the movie is shown on the opening weekend, social media trends and even the weather. The number of theatres is important due to the fact, that it presents an upper boundary for the possible box office earnings. Due to the restrictions of the challenge, participants were only allowed to use IMDb data, unfortunately they do not provide budget information and the number of theatres a movie is shown in. This shifted the focus of box office prediction to the analysis of social media data and the release date, also taking the genres into consideration.

Features

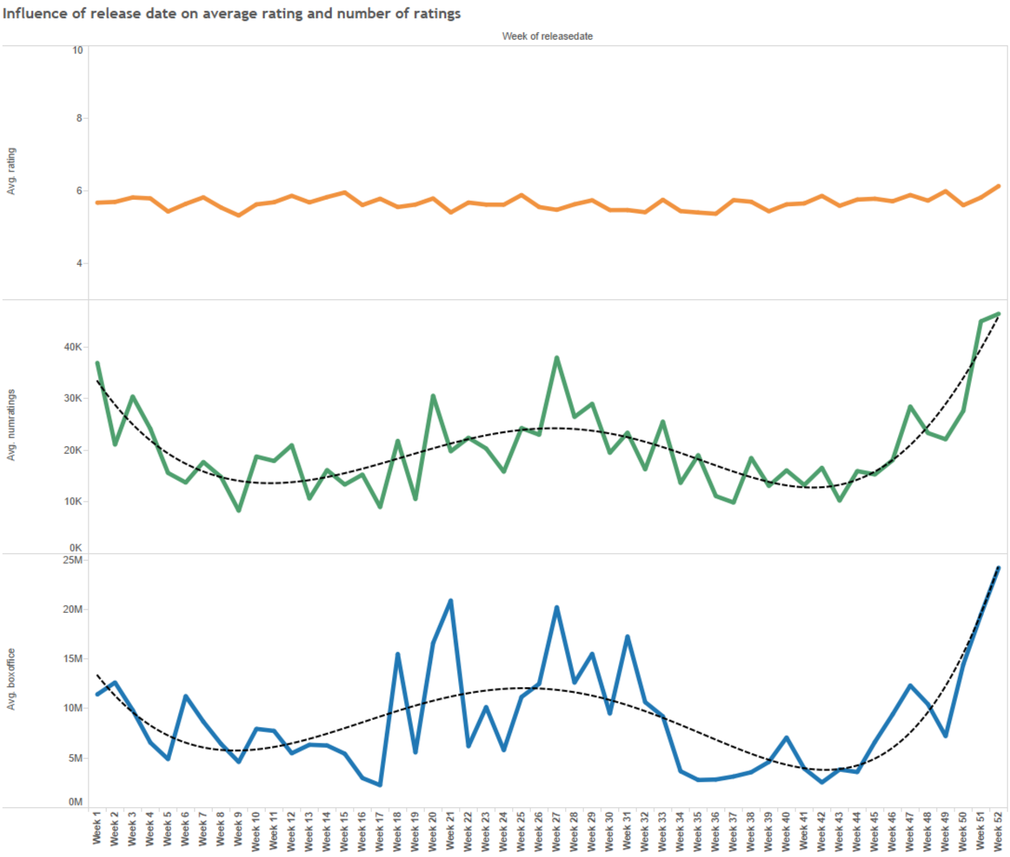
Twitter

The days leading up to the release of a movie are very important, because people often go to the movies do so spontaneously and based on recommendations of friends or fitting marketing. Since people nowadays interact with their friends via social media quite a lot, predictions about interest in the movie can be quite accurate. I used Twitter4J to download tweets and imported them via CSV into Tableau. Due to limited resources and time a sentiment analysis of the tweets was not possible. So I used the number of tweets filtered for some keywords by the challenge committee as well as the trend, that is the number of tweets for each day, from two weeks before, leading up to the release of the movie. Due to rate-limiting from Twitter, I could only download 180 tweets. For this reason 180 of all the provided ID's were chosen randomly and the tweets for them were downloaded. Nevertheless, the sample size is big enough to extrapolate and make trend predictions.

Bitly

Bitly data became available only later in the challenge. I only used the total number of links clicked for the links provided. Thus I missed both a proper trainings-set as well as experience, as to which numbers were a good indicator for high box office income. Therefore I mainly used the twitter features, but also found out that one of the twitter data sets was corrupt, for my submission on the 28.06.2013 for the movie The Heat due to the huge discrepancy between twitter and bitly.

Figure 3.5: Avg. rating, avg. box office revenue and average number of ratings shown for each week of the year



The trends of average of rating, average of numratings and average of boxoffice for releasedate Week. The data is filtered on numratings and releasedate Year. The numratings filter ranges from 100 to 944503. The releasedate Year filter keeps 8 members. The view is filtered on releasedate Week, which excludes Null, Week 53 and Week 54.

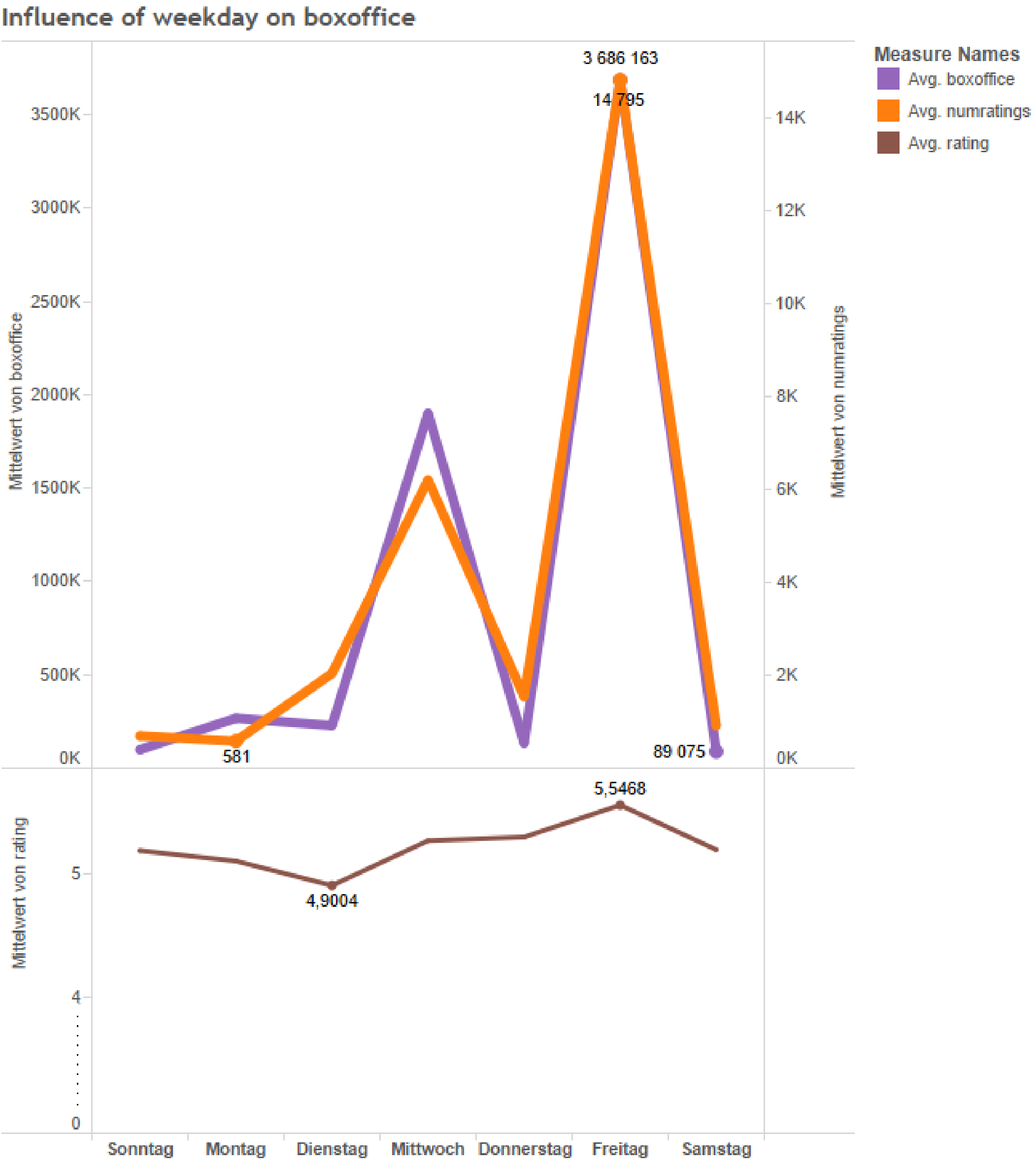
Release date

The release date plays a major role for the box office and the success of a movie. Major studios plan their movies and also the release dates for them years in advance. Differences between release dates can have an impact as big as up to a factor of 3 as the analysis of historical data shows, this can be seen in figure 3.5. This figure is very meaningful for it shows, 3 important things. First, the release date has no influence on the movie's rating. Second, the box office and the number of ratings of a movie highly correlate. And third, a trend for the year, reaching the top box office ticket sales in holiday seasons. In order to compare movies from different years with each other, it makes sense to only look at the week of the year they have been released in and at the weekday. Generally speaking, all big movies premiere on Fridays, because people usually go to the movies on weekends. Of course, days prior to big holidays, i.e. the Independence Day on the 4th of July are also very popular for obvious reasons. Low budget movies tend to be released on Wednesdays in order to distinguish themselves from blockbusters (compare figure 3.6).

Bringing it all together

For the foundation to predict the box office I used the release date, since this indicates what is possible in terms of box office income at a certain time of the year. Analysing the number of tweets and the trend, almost always led to a big correction of the first estimate. In the end I figured out that the box office income can be approximated quite well by simply correcting the prediction based on the release date by the number of tweets multiplied by a factor. I used the forgone submissions to tune this factor. The trend was used to compare movies on the opening weekend with each other. The assumption here was, that only a finite number of users will tweet about a movie, so if one movie has more tweets and a better trend than the other movie, this movie will certainly have a bigger box office on the opening weekend.

Figure 3.6: Avg. rating, avg. box office revenue and average rating shown for each weekday



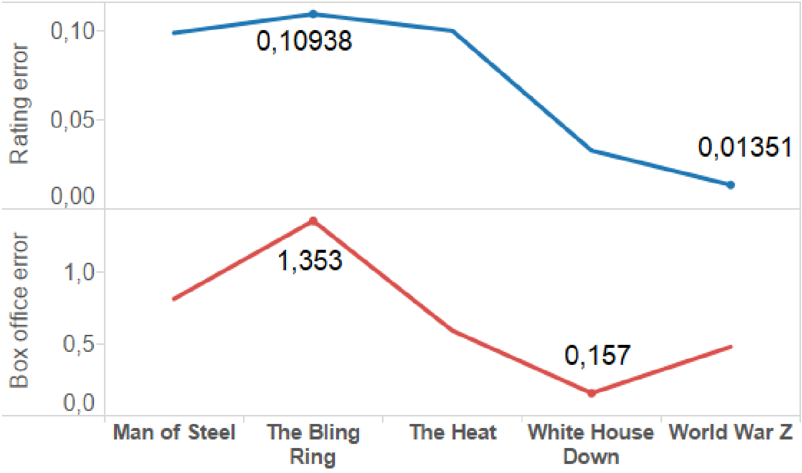
The trends of Avg. boxoffice, Avg. numratings and Avg. rating for releasedate Weekday. Color shows details about Avg. boxoffice, Avg. numratings and Avg. rating. Details are shown for releasedate Weekday. The data is filtered on releasedate Week, which excludes Null, Week 53 and Week 54.

CHAPTER 4

Results

The results of this work can be seen as the discrepancy between the predictions I made and the actual values of the movies' viewer ratings and box office incomes. This table shows the hard numbers for all movies I predicted, I will go into detail on two movies, to show what mistakes I made, that I afterwards avoided.

Figure 4.1: MRAE for the movies predicted



4.1 Viewer Rating

One can see that the error rates are quite low, and improved during the course of the challenge. This can be accredited to the experience effect as this part of the movie success prediction heavily depended on subjective abilities. Especially the twitter data were great help in improving accuracy of predictions, because they provide information about the general interest in the movie, just days before its release.The prediction for 'The Heat' shows very well that corrupt twitter data led to a poor prediction, both for rating as well as box office numbers. In order to really show the reliability of the system, many more movies should be predicted. Even more so as movies can be very hard to predict nowadays, because of many success or failure amplifying effects such as word of mouth or virality of online content [Berger and Milkman, 2012]. Bad aswell as good news can spread within hours.

Table 4.1: prediction, actual outcome and error for ratings of movies I analysed in chronological order

Viewer Rating	Predicted	Actual	Error
The Bling Ring	7.1	6.4	0.109
Man of Steel	7.3	8.1	0.0987
World War Z	7.3	7.4	0.0135
The Heat	6.3	7.0	0.1
White House Down	6.3	6.1	0.0328

4.2 Box Office

Predicting the box office numbers is way harder than predicting the viewer rating of a movie. This has mainly two fundamental reasons. On the one hand, the scale of viewer rating, from 1 to 10 is capped, while there is no upper bound, on the other hand, the box office income tends to vary greatly, depending not only on the success of a movie, but also on factors like weather, season and competing movies on the opening weekend, which is shown in chapter 1. For example 'The Bling Ring' was only shown on the prediction weekend in very few theatres, therefore its box office is an order of magnitude smaller compared with the other movies, since they didn't have the same external preconditions. Generally, it would make sense to include the number of theatres as one feature in the prediction in future works, because with help of these, an upper limit for box office sales can be found. I didn't do it in this work, because the challenge only allowed the use of IMDb data.

Table 4.2: prediction, actual outcome and error for box office of movies I analysed in chronological order

Box Office	Predicted	Actual	Error
The Bling Ring	0.5	212537	1.352
Man of Steel	22	116.6	0.811
World War Z	35	66.4	0.473
The Heat	16	39.1	0.591
White House Down	21	24.9	0.157

The results show the capabilities of the prediction system I used. With low effort and some domain knowledge, good predictions can be accomplished. The error rates suggest a steady improvement in the course of the challenge. This shows that the additional features I implemented during the challenge, paired with the gained experience in predicting movies, substantially improved the prediction ability.

CHAPTER 5

Conclusion

The model I used to predict movie success is quite simple, but still powerful enough to make good predictions. Compared with other proposed methods from the challenge it is by far less sophisticated, though its strengths lie in its simplicity. The challenge was a great opportunity to learn about Visual Analytics and broaden knowledge in many areas. It is always hard to find the right balance between only analysing the surface of a problem and getting stuck in details. The approach to viewer rating prediction is very straightforward and can be easily applied by anyone. The box office prediction requires more experience, because of its dependency on the user to be able to interpret the different visualizations of the data in the right way and infer the right conclusions from it. Visual Analytics is not widespread in the field of movie success prediction, so the challenge was a good way to fuel interest in this research topic. I am very confident Visual Analytics will become more and more integrated in our lives for the great possibilities it enables. The human mind is very good at extracting information from visualizations and thus can analyse data quicker and in a more complex way. Particularly in our times where the amount of data we produce is skyrocketing.

Bibliography

[Al-Masoudi et al., 2013] Al-Masoudi, F., Seebacher, D., and Schreiner, M. (2013). Similarity-driven visual-interactive prediction of movie ratings and box office results. `http://bib.dbvis.de/uploadedFiles/almasoudi.pdf`. [Accessed 26.02.2014].

[Baccianella et al., 2010] Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC 2010)*, pages 2200–2204. European Language Resources Association (ELRA).

[Berger and Milkman, 2012] Berger, J. and Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, 49(2):192–205.

[Cook et al., 2013] Cook, K., Grinstein, G., and Whiting, M. (2013). VAST Mini Challenge 1: Visualize the box office. `http://boxofficevast.org/`. [Accessed 12.12.2013].

[Demuth and Beale, 1993] Demuth, H. and Beale, M. (1993). Neural network toolbox for use with matlab.

[El Assady et al., 2013] El Assady, M., Hafner, D., Hund, M., Jäger, A., Jentner, W., Rohrdantz, C., Fischer, F., Simon, S., Schreck, T., and Keim, D. A. (2013). Visual analytics for the prediction of movie rating and box office performance. In *IEEE Int. Conf. on Visual Analytics Science and Technology (VAST Challenge Paper)*.

[Fazzion et al., 2013] Fazzion, E., Las Casas, P., Gonçalves, G., Melo-Minardi, R., and Meira Jr, W. (2013). Open Weekend and Rating Prediction Based on Visualization Techniques. In *Proc. IEEE Int. Conf. on Visual Analytics Science and Technology (VAST Challenge Paper)*.

[IEEE VIS, 2013] IEEE VIS (2013). IEEE VIS conference. `http://ieeevis.org/`. [Accessed 12.11.2013].

[IMDb, 2014] IMDb (2014). The internet movie database. `http://www.imdb.com/`. [Accessed 26.02.2014].

[Jäger et al., 2013] Jäger, A., Hafner, D., and el Assady, M. (2013). Moovis - a visual analytics tool for the prediction of movie viewer ratings and boxoffice. `http://bib.dbvis.de/uploadedFiles/MooVisSummaryFinal.pdf`. [Accessed 26.02.2014].

[Kosara, 2007] Kosara, R. (2007). InfoVis contest 2007 data. Eagereyes Weblog. http://eagereyes.org/blog/2007/infovis-contest-2007-data [Accessed 26.02.2014].

[Mat Kelly, 2013] Mat Kelly, Michael L. Nelson, M. C. W. (2013). Graph-Based Navigation of a Box Office Prediction System. In *Proc. IEEE Int. Conf. on Visual Analytics Science and Technology (Poster)*.

[Netflix, 2005] Netflix (2005). Netflix challenge. http://www.netflixprize.com//rules. [Accessed 03.11.2013].

[Numbers, 2013] Numbers (2013). The numbers budgets. http://www.the-numbers.com/movies/records/budgets.php. [Accessed 26.02.2014].

[Oghina et al., 2012] Oghina, A., Breuss, M., Tsagkias, E., and de Rijke, M. (2012). Predicting imdb movie ratings using social media. In *ECIR 2012: 34th European Conference on Information Retrieval*, page 503–507, Barcelona, Spain. Springer-Verlag, Springer-Verlag. [Accessed 11.05.2013].

[Perin, 2013] Perin, C. (2013). CinemAviz. In *VAST Challenge 2013*, Atlanta, GA, United States. [Accessed 26.11.2013].

[Simonoff and Sparrow, 2000] Simonoff, J. S. and Sparrow, I. R. (2000). Predicting movie grosses: Winners and losers, blockbusters and sleepers. *CHANCE*, 13(3):15–24. [Accessed 10.05.2013].

[Tableau Software, 2014] Tableau Software (2014). Tableau. http://www.tableausoftware.com/. [Accessed 26.02.2014].

[Yafeng Lu and Maciejewski, 2013] Yafeng Lu, F. W. and Maciejewski, R. (2013). Excellent Visual Analysis of Structured and Unstructured Data - Team VADER. In *Proc. IEEE Int. Conf. on Visual Analytics Science and Technology (VAST Challenge Paper - Award for Effective Analytics)*.

[Zhang and Skiena, 2009] Zhang, W. and Skiena, S. (2009). Improving movie gross prediction through news analysis. In *International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pages 301–304. IEEE.