Entry Name: "USTUTT-Thom-MC1"
VAST 2013 Challenge
Mini-Challenge 1: Box Office VAST

Team Members:
Dennis Thom, University of Stuttgart, dennis.thom@vis.uni-stuttgart.de PRIMARY
Edwin Püttmann, University of Stuttgart, edwin.puettmann@vis.uni-stuttgart.de
Florian Heimerl, University of Stuttgart, florian.heimerl@vis.uni-stuttgart.de
Harald Bosch, University of Stuttgart, harald.bosch@vis.uni-stuttgart.de
Qi Han, University of Stuttgart, qi.han@vis.uni-stuttgart.de
Robert Krüger, University of Stuttgart, robert.krueger@vis.uni-stuttgart.de
Steffen Koch, University of Stuttgart, steffen.koch@vis.uni-stuttgart.de

Student Team: NO

Analytic Tools Used:
Twitter4J (twitter4j.org) and
Java Movie Database (jmdb.de) for obtaining Twitter and IMDB data.
Knime (knime.org) and
Tableau (tableausoftware.com) for initial tests and data set exploration.
Weka library (www.cs.waikato.ac.nz/ml/weka) and
LibSVM library (www.csie.ntu.edu.tw/~cjlin/libsvm/) for model training and evaluation.
Prolix System, developed by University of Stuttgart during the course of the challenge

May we post your submission in the Visual Analytics Benchmark Repository after VAST Challenge 2013 is complete? YES

Video:
http://youtu.be/6AY4RVAGc8c
Description:

**Approach**

Prediction of box office results has a long history in information mining and machine learning research - traditionally based on movie meta-data, but recently also based on web and social media data. Although there are many successful ideas, the overall performance of pure data modeling approaches seems to have reached an upper limit. Specifically outliers are hard to predict using only structured data and problems of prediction algorithms, which are most often black boxes, are difficult to identify and correct. With the subsequently described Prolix system, we aim at complementing fully automatic methods with interactive visualization to let analysts consider additional information, e.g. from social media data sources, and integrate their expert knowledge to build, refine, and apply models for improved revenue and rating predictions. For this purpose, Prolix facilitates interactive analysis loops for gaining insights directly from textual social media content, exploring and building contexts of movies related to the prediction candidate, and for comparing and selecting features the prediction is based on.

**Data Preprocessing and Feature Generation**

The Internet Movie Database (IMDB) contains comprehensive metadata for each movie. The available text-based IMDB data files were converted to an SQL-database, which allowed further preprocessing in different ways: First, the multi-valued fields were normalized and semi-structured information, like cast order, was parsed. Additionally, the data was cleaned of entries that were duplicate, incorrect, or irrelevant to the MC1 task such as video games/short movies. The huge number of nominal values such as actor names leads to a much higher number of values than training instances, which hampers machine learning. Thus, we calculated numeric values for actors, producers, directors, etc. on a per movie basis by aggregating the maximum, total, and average weekend gross and ratings of previous movies they were involved in. Other attributes with limited value ranges like genre and MPAA rating were represented as binary vectors.

As second data source we used the Twitter messages with the provided Ids. To make the Twitter data usable within our prediction models we applied NLP methods, such as tokenizing, part of speech (POS) tagging, and negation recognition. We then extracted unigrams, bigrams, POS unigrams, POS bigrams, as well as embedded links as features and trained a support vector machine (SVM) to classify tweets according to their sentiment as positive, neutral or negative. Additionally, messages signaling positive interest to watch a certain movie were identified.

**System Overview**

The Prediction View (Figure 1.A) is the core component of Prolix used to define the prediction candidate and select a predictor from the available regression models, i.e. SVM, multi-layer perceptron (MLP), or linear regression (LS). After starting the training and subsequent prediction, a list entry is generated containing information about training set size and number of features used. By selecting a given prediction, the prediction range is shown in a table as well as a grey highlight in the scatter plot (1.B) illustrating how the model performs for each movie in the training set in terms of predicted vs. actual value. Hovering over a movie glyph gives more detailed information, such as absolute prediction error. The movies can also be colored according to a specific feature, revealing dependencies in the model. Additionally, a feature lens can be moved over the scatter plot to see averaged attribute values in distinct areas of the plot (Figure 3). Lastly an error diagram (1.C), based on an n-fold-cross-validation shows mean absolute errors and $R^2$ scores.

The Feature Selection View (1.D) can be used to adapt the feature set for improving the prediction. A color coded matrix illustrates correlation between attributes of movies in the training set (blue indicates high correlation over white (low) to red (negative correlation)), how many movies actually have a specific feature, and the prediction candidate’s attribute values.

The Related Movies View (1.E) helps to find movies similar to the prediction candidate. The analyst can select several attributes of the candidate, e.g. main actor, director, genre to generate a list of movies that are similar with respect to these attributes. The opening weekend gross of the selected movies are then shown as bars in the chart, highlighted in the scatter plot, and they have their own error bar in the Prediction View. This helps to better align the predictors to movies similar to the prediction candidate instead of creating a suboptimal one-size-fits-all prediction model.
The Twitter Volumes View (1.F) indicates Tweet volumes related to each movie as histograms over the last 14 days prior to release. The histogram is shown for the prediction candidate and movies selected for comparison, if Twitter data is available. By hovering over the bars the respective number of tweets is displayed and the Word Cloud View (1.G) and Table View (1.H) are updated to indicate most prominent terms and tweets of the selected day accordingly. Based on a specifically trained SVM classifier we can separate positive interest tweets (“Wanna watch #ironman”) from negative tweets and buzz. These affections are consistently highlighted (red=positive, blue=negative/buzz) of overall volumes, tags and individual tweets in the three views. To analyze the available data the user can apply keyword and other filters (1.I), e.g. tweets containing URLs or images, in order to show only distinctive tweets in the Table View and aggregate respective volumes and content in the Volume- and Tag Cloud View.

In order to identify unusual developments, the Prediction Evolution View (1.J) applies Gaussian Processes for predictions based on extracted Twitter features, beginning on the 10th day before the release. For each day thereafter we build a Gaussian Process (GP) model based on the available data and show the evolution of the predictions. GPs produce a probability distribution so that the analyst can not only see the most probable value but also uncertainty ranges. Showing the prediction progress through days enable the analyst to find anomalies to further investigate using the Word Cloud View and the Tweet Table.

Interactive Analysis Process (Loops)
Our analysis process, as depicted in Figure 2, consists of two primary loops, the model training loop (orange/black) and the social media analysis loop (purple) enabling analysts to
combine insights from structured and unstructured data sources.

Assume the analyst wants to predict the gross and rating for the movie “Now You See Me” (NYSM). After data preprocessing (Figure 2.A) she would begin the analysis by selecting the movie in the Prediction View and define an initial feature selection and training set (2.B). For example, she could restrict the training set to newer movies that were released after the year 1995 and select features that she considers relevant. At this point she can identify movies in the Related Movies View that resemble a close similarity to the prediction candidate. In the case of NYSM this could be other heist movies like “Ocean’s Eleven” or magician-related ones like “The Illusionist” that will then be highlighted in other views for comparison. In the next step the analyst would select one or more regression models (SVM, MLP, LR) in the Prediction View to perform training on the training set and subsequent prediction for the candidate (2.C). Each completed prediction is then represented in a list (1.A) as a combination of model, training set and selected features. From this list the analyst can select individual predictions and assess their performance on their training set using the actual vs. predicted gross scatter plot and the error diagrams (2.D, 1.B, 1.C). Furthermore, she can explore the results by hovering over individual movies in the scatter plot or use the feature lens to investigate complete ranges of the plot (see Figure 3).

However, for some movies the analyst might lack confidence in the result of a pure meta-data based prediction, e.g. because there are several outliers that were marked as closely related. At this point she can turn to “external sensors” (2.E) to investigate on Tweeters’ opinions about the prediction candidate. The social media views will already have updated to the prediction candidate and the movies that were marked as similar. By browsing over the daily aggregates in the Tweet Volumes View and using the Word Cloud and Table View the analyst can investigate sentiments, opinions, and content. The insights gained from the social media data exploration can then directly be used to change interpretation on the prediction evaluation and influence the refinement in the further process. In addition, the analyst can directly derive formal features from the Twitter data by using a range of textual and meta-data filters. For example, she could select only messages that contain the keywords “going” and “watch” and that were written 5 days prior to release. Or she could only select messages that have positive sentiment and mention a YouTube trailer. The corresponding dynamic features will then be activated in the Feature Selection View and can be used besides the movie meta-data features by the regression models. The Prediction Evolution View further helps the analyst to identify unusual developments and changes, from which features can be derived.

From what the analyst learned from the social media data, the error measures and the exploration of the predictor’s performance she might now want to refine the prediction process iteratively (2.F). This can be done either by changing the training set and feature selection or by adjusting and combining individual regression models to improve the predicted outcomes.

![Figure 3: Feature Lens - mean values of movies for each feature (vertical line), standard deviation (horizontal line) and deviation to the feature of the movie predicted (red bar). It can be seen that movies predicted lower (a) have lower weekend screens, budget and first actor values, while movies predicted higher (b) behave conversely.](image-url)
Questions:

1. **What data factors, alone or in combination, were most useful for predicting possible outcomes?**

   The importance of individual data factors is directly reflected in the Feature Selection View that shows feature correlation in descending order of importance. Most relevant are weekend screens and budget. By training with only those two features we already achieve an $R^2$-accuracy of 55%. They are followed by producer, production company, and director. For the few movies with Twitter data, Twitter features, like number of video and image URLs in tweets of specific days, show high importance. Using the exploration lens and feature coloring in the scatter plot these coherences can be validated. Additionally, the analysts’ ability to detect facts of a movie that are disseminated through social media channels but which are difficult to extract or express as features, can be used to set adequate weights on combined prediction models to guide them into the right direction.

2. **How did you combine factors from structured data with factors in unstructured data and what was the impact on the results? Did you see correlations? How can a user of your system explore this combination?**

   As indicated in our analysis process, we support two different ways to integrate unstructured and structured data: Dynamic Features or Insight & Interpretation. One can derive new structured features from Twitter by defining filters and using the resulting message counts. These new features can be combined with IMDB data as (1) additional features in a single prediction, (2) in an own prediction that is later combined with other predictions, or (3) as features for predicting the error that will be made using only IMDB features. The first approach usually suffers from the limited amount of training instances that have Twitter data, leading to a problem in learning the impact of the many IMDB features. The second approach allows the training based on an optimal training set for each type of feature and a subsequent linear combination for the movies that have both. By interactively adjusting the combination weights, the analyst can observe the influence on error measures and scatter plots. The third approach performed well for selected movies but was not guaranteed to increase the accuracy. In addition to the dynamic features, the analyst can gain insights from investigating the content of unstructured data. These insights can then be used to rate or refine the final prediction results. There are some interesting correlations between our extracted Twitter features and movie genres: Action, Adventure, and SciFi have a high positive correlation with Twitter features, while Comedy, Crime, and Family have negative correlations.

3. **Do the important factors vary by class, such as movie genre?**

   The genre of a movie has impact on its success (especially weekend gross). Figure 5 shows training results, colored by different genre types. While action movies are located in low to high gross range (A), romances perform in a much lower range (B).

   One can observe that the importance of features (as measured by the Information Gain Ratio between the gross and the specific feature) vary by movie class (C). For example, budget in general is an important feature but using our tool one can observe that this is not the case for horror movies. Writer is a very important feature for SciFi and animation movies, less so for comedy movies. The production company/director is more important for animation movies.
4. Did you use data on previous movies to help analyze/predict outcomes for later movies? If so, how?
Using previous movie data is essential for our approach, since we apply supervised learning methods. We build our models from previous movie data using it as training and test instances in an n-fold (10-fold per default) cross validation. These models are then used to predict a new movie’s weekend gross and rating.

5. For any prediction that you had a significant margin of error (for our challenge, this would be a high mean relative absolute error), explain possible sources of error.

*After Earth* (RAE: 137%, predicted too high), was an early submission, where only some IMDB- and no features from Twitter were used. Our current system predicts this movie in the range of 36-45 M, which is much closer. Investigating the available tweets, the number of messages was pretty low but mostly positive. Reading the media however reveals that there was bad advertisement and rumors about Scientology affiliation, which both was underrepresented in the allowed-to-use data. Searching for “Scientology” (Content Table View) shows only two results: “More and more people saying AE is about Scientology” and “AE is scientology movie, right?” which we may have underrated.

For *The Lone Ranger* (RAE: 86%, predicted too high), our expert opinion based on the promotion-strategy “like Pirates of the Caribbean with Johnny Depp but western” was not successful. Our model takes previous success into account and thus the movie’s features for actors, directors, etc. had quite high values. Mostly this works, but sometimes there are factors not contained in these features. Inspecting the Twitter data reveals that there was only little positive sentiment for this movie. However, due to interruptions during the collection of Twitter data, this source was not reliable.

With *R.I.P.D.* (RAE: 67%, predicted to high) we had similar problems since it comes with prominent cast and high budget. The provided tweets gave no further hints for bad performance, too. Probably, the competition of similar movies was too high and not covered by our features.

6. What data trends if any were you able to identify? How did the identification of trends affect / shape predictions? Did you see instances where early data about a movie was contradicted by later data/factors?

*Ratings*: We observed that ratings are very pronounced right after the release, but then even out over the weeks, often by more than one unit. As our models are built using old movie ratings, it tends to underestimate the ratings.

*Season Trend*: We investigated the trend of weekend gross values over the course of a year and revealed an interesting pattern (Figure 6.A). Thus we extracted a week of year feature which slightly improved prediction results.

*Release years*: As with more training instances the prediction gets better we also want to use movies that are a somewhat older, saying from the 90s. But back these days the gross values were much lower (Figure 6, B). We therefore corrected the gross and budget values using the price increases of the USA entertainment sector, which improved prediction.

![Figure 6: Temporal trends in IMBD data](image)

**Typical Twitter time series:** We observed that Twitter volume time series have typical periodic behavior. It usually peaks on about the day before movie release and the 7 days before release. We calculate the mean Twitter volume for each day from all the movies (Figure 7). In the contrary we observe that the sentiment ratio (num positive/ num negative tweets) of tweets is stable across days. Our predictors learns such trends and makes predictions based on that.

![Figure 7: Trends in Twitter data](image)

**Contradictory Instance:** There are some complicated cases like “The Internship (2013)” which has a relative low sentiment ratio on the beginning but just before the release the ratio jumps up. This may indicate a possible ad campaign.