Volume 10, Issue 2, February – 2025 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14987459

Predicting Film Box Office Performance Using Wikipedia Edit Data

Niraj Patel[1]

[1]St. Clair College

Publication Date: 2025/03/11

Abstract: This study explores the potential of Wikipedia edit data as a predictor of opening box office revenues for films released in the US. After analyzing films from 2007 to 2011, we developed a predictive model based on Wikipedia article edits using gradient boosting trees as the primary algorithm. Our model incorporates features such as the frequency of Wikipedia edits, the size and content of article revisions, and the revenues of similar films. The results demonstrate that Wikipedia activity can serve as a rough indicator of film popularity, though the model's predictive accuracy is limited. We find that Wikipedia-based features, particularly edit runs and content changes, significantly contribute to the model's performance, achieving an R^2 of 0.54 for films released in 2012. This suggests that while Wikipedia data offers valuable insights into social interest, it is best used in conjunction with other predictors for more reliable revenue estimates.

How to Cite: Niraj Patel (2025). Predicting Film Box Office Performance Using Wikipedia Edit Data. *International Journal of Innovative Science and Research Technology*, 10(2), 1951-1956. https://doi.org/10.5281/zenodo.14987459

I. INTRODUCTION

➢ *Wikipedia as a Gauge of Social Interest*

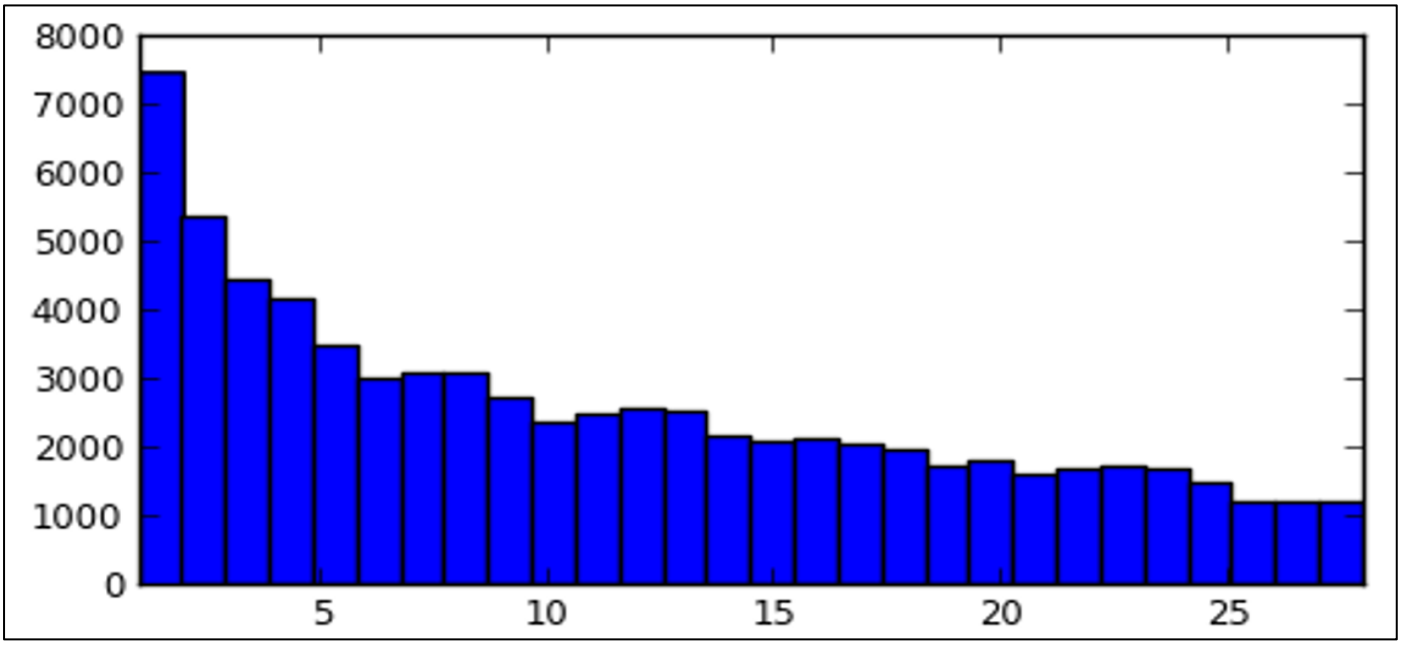

Fig 1 Count of Wikipedia article edits for the films used in this paper's training dataset over the 4 weeks prior to each film's respective release date, bucketed by days before the release date that the edits occurred. This graph shows the uptick in editing activity that typically accompanies a film's release.

- According to its article about itself (as of this writing), Wikipedia is "a collaboratively edited, multilingual, free Internet encyclopedia" launched in January 2011. [6] Its articles can be edited by anyone, either anonymously (though the editor's IP address is logged) or with a registered user account. The edit history of each article is saved with a timestamp. Interested users can view any past version of an article, and an article's edit history exhibits an evolving record of Wikipedia's "knowledge"1 of its subject.

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14987459

- As such, Wikipedia's edit history can be viewed as a barometer of social interest. For example, when a person is in the news, editing activity in his or her article often spikes. In fact, Wikipedia has template warnings indicating when an article is likely to be in flux due to a relevant current event. Edit activity on Wikipedia, in this sense, is akin to mentions on social networks like Facebook or Twitter, although perhaps with a smaller participating audience (although many people read Wikipedia, not nearly so many participate in its creation).
- One area where we can try to gauge the degree to which Wikipedia activity reflects social in- terest is in film box office performance. Films have relatively well-defined release dates prior to which we can measure activity on Wikipedia. They also have well-defined, measurable outcomes - revenues at the ticket booth - that are clearly sensitive to popular interest. Theater owners obviously have a direct fi- nancial interest in knowing how well a film is going to perform. Advertisers and publicists, sellers of tie- in products, and film journalists have a slightly more indirect but still strong interest; they will want to know how they should spend their time and money. Can we use Wikipedia to usefully predict films' opening box office performances?

II. FORMULATION OF PROBLEM AND DATA SOURCES

➢ *The specific question I set out to answer was how accurately, using Wikipedia's help, can we predict the domestic per-theater box office gross of a film released widely in the US over the first three days2 of its release.*

- Of course, Wikipedia's highly open policy means that it contains a stunning breadth of information from contributors with wide- ranging expertise and that said information is sometimes unreliable. For an example that was in the news not long before this paper was written, see [5], or for Wikipedia's own list of Wikipedia hoaxes, see [4].

- Films traditionally open on Friday, and their "opening" often refers to their gross over the first Friday, Saturday, and Sunday that they are playing. However, there are plenty of non-Friday openings. Consequently, I've stated the problem in terms of the first three days' worth of grosses.

➢ *The Data Sources I used to Answer this Question were:*

- Box Office Mojo (http://www.boxofficemojo. com/) - contains detailed box office data. I used it to select the universe of films to analyze and as my source for theatrical release dates, number of opening theaters, and revenues. There is no API - I scraped the data with the Python package Beautiful Soup.

- Rotten Tomatoes (http://www.rottentomatoes. com/) - a popular movie review aggregator. I used it to obtain descriptive information about films: genres, runtime, MPAA rating, cast and directors, and so on. It offers an API if you register for a key (which is free as of the present writing).

- Wikipedia (English-language) (http://en.wikipedia.org/)- MediaWiki, the name of the web application upon which Wikipedia is based, offers an API, no registration or key necessary.

- Much of the work involved in data retrieval and formatting was to ensure that data retrieved from these three sources corresponded to the same film; data from Rotten Tomatoes and Wikipedia was obtained by using their APIs' search functionalities, which can lead to incorrect hits if you are not careful. For example, we want to make sure that Rotten Tomatoes data for the 2012 film "The Lucky One" is not mapped to the 2008 film "The Lucky Ones," or that for the 2010 film "Salt" we do not examine the Wikipedia article for salt, the mineral.[3]

- The universe of films that I considered were those listed on Box Office Mojo as having opened in at least 1000 theaters. I manually excluded a handful of films that were re-released or had limited engagement special features. I trained my algorithms on films released between 2007 and 2011, inclusive. In total, 689 films were in the training dataset. Data from films as far back as 2002 were used for some of the feature calculations; see the next section for more details. I tested my algorithm on films released in 2012, of which there were 124.

- Box Office Mojo data had to be scraped from HTML, but the HTML was regular and consistent. Rotten Tomatoes has a nice JSON-based API for data retrieval, but its ranking of returns is quirky, sometimes retrieving obscure films or films with similar names (example: Oliver Stone's 2008 biopic "W." was unfindable through search query, even though the website's front end; I had to go to Stone's Rotten Tomatoes page just to find the relevant web page). Wikipedia has a nice API and solid/consistent lookup, which is all the more impressive given that it contains articles on anything, not just films.

III. FEATURES

A. *Descriptive Features*

- The descriptive features considered were the year of release, runtime, MPAA rating, whether the film was released on a Friday, and membership in gen- res as defined by Rotten Tomatoes. Rotten Tomatoes has 18 genre labels, listed below. A film can belong to any number of these genres.

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14987459

- Action & Adventure
- Animation
- Art House & International
- Classics
- Comedy
- Cult Movies
- Documentary
- Drama
- Horror
- Kids & Family
- Musical & Performing Arts
- Mystery & Suspense
- Romance
- Science Fiction & Fantasy
- Special Interest
- Sports & Fitness
- Television
- Western

B. Wikipedia-Based Features

- For each Wikipedia article, I measured the number of edits runs that occurred during the period 0 to 7 days prior to midnight on the day of the film's release, as well as during the period 7 to 28 days prior. I defined an edit run as a sequence of consecutive edits from the same author (identified by IP address if anonymous). Sometimes, on Wikipedia, the same author commits several edits in a row, presumably as part of a single effort to edit the page, which I wanted to correspondingly treat as a single edit. I generally found this to be a slight improvement over raw edit count in terms of predictive power.

- I also extracted a few features from the content of the article revisions themselves. One feature I used was the average size, in bytes, of revisions in the 28-day window. Other features were obtained by scanning the text of the revisions for certain textual patterns. One was a count of the number of article section headings, another was a count of the number of external file references (typically an image or sound file inserted into the article), and the last was a case- insensitive search for the word "IMAX".

C. Revenues of Similar Films

- A natural approach to predicting the box office performance of a film is to look at comparable films; in particular, the natural benchmark for a sequel is its predecessor. To this end, I created a feature consisting of revenues of "similar" films released in the five years preceding each film's release (hence, data as far back as 2002 was involved, even though the training dataset extended only as far back as 2007). The five-year window was arbitrary, but I think it forms a reasonable bond when comparing expected box office performance.

Iron Man 2	0.4472
Fantastic Four: Rise of the Silver Surfer	0.2462
Iron Man	0.2462
Thor	0.2462
Captain America: The First Avenger	0.1741
Push	0.1741
Sherlock Holmes: A Game of Shadows	0.1741
The Losers	0.1508
Scott Pilgrim vs. the World	0.1508
Sherlock Holmes	0.1508
Shutter Island	0.1508

Fig 2 Example similarity scores: for "The Avengers" (2012).

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14987459

- Similarity between two films was defined as the geometric mean of the Jaccard[4] similarity measures of the films' 1) Rotten Tomatoes genre information and 2) Rotten Tomatoes cast/director information. The Rotten Tomatoes API only returns the first few starring members of each film's cast, so the metric is not distorted by differing cast sizes. Directors were always treated as single people, even if there were co- directors, so for our purposes, the Coen brothers, for example, count as a single person.

- The feature incorporated into the algorithms was, for each film, the opening revenue of all other films in our universe released up to five prior to that film, weighted by similarity. See Figure 2 for an example of similarity scores for one of the films in the test dataset.

IV. ANALYSIS AND PREDICTION

- I tried a few different prediction algorithms; the one that proved the most effective on the test set, as measured by R^2, was gradient boosting trees.[5] Gradi- ent boosting is a general predictive technique pioneered by Jerome Friedman of Stanford in which a predictive formula is generated by summing so-called "weak predictors" that are sequentially fit to the gradient of a specified loss function (for example, squared error). The overall model may be accurate and robust even if each individual weak predictor is very simplistic. Gradient boosting trees refer to gradient boosting with decision trees as our weak predictors. For details, see Friedman's article [2], and also Wikipedia's own page on gradient boosting [3].

- I used the Python statistical package scikit- learn's implementation of gradient boosting trees, using the default learning rate and least squares as my loss function.There are a few other model parameters.

[4]The Jaccard similarity of two sets A and B is defined as.

$$|A \cap B| / |A \cup B|$$

[5]Random forests and ordinary linear regression performed worse, but not by much. Despite the clearly non-normal distribution of the revenue per theater (it has a positive skew), I did not have better success with a generalized linear regression than with ordinary linear regression.

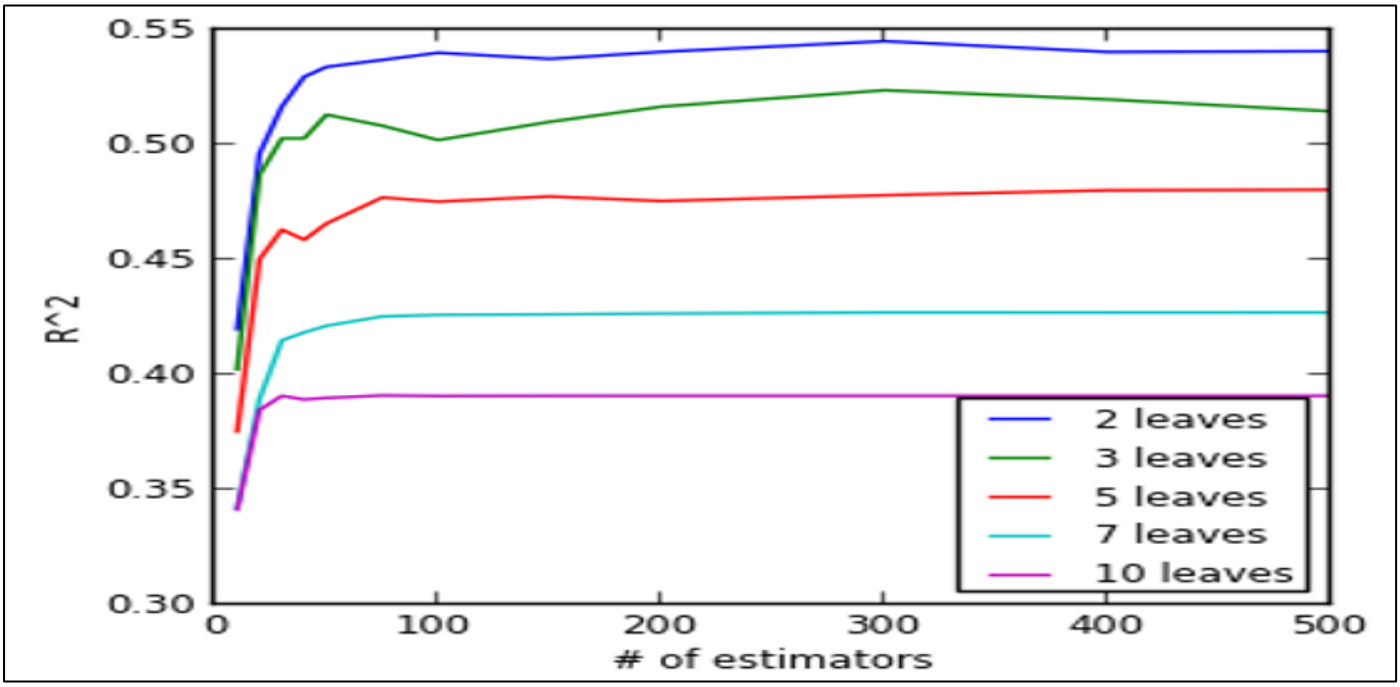

Fig 3 Estimators

Fig. 3: R^2 of gradient boosting tree models on the test dataset as a function of the number of estimator iterations. The different curves represent different numbers of leaves in the weak learner decision trees. The simplest weak learner, a 2-leaf tree, performs the best. Using stochastic gradient boosting trees, in which a subsample of the features is used to fit the decision trees, improved the high-leaf models to some degree. This suggests that the inferior performance of the higher-leaf models may be due to overfitting. that can be controlled by the user; the most important ones are the number of estimators (the number of weak predictors to fit) and the depth of the trees (how many leaves are in each decision tree - this parameterizes the complexity of each individual weak predictor).

- Adapting the example in scikit-learn's docu- mentation [1], I calculated the R^2 of gradient boosting trees at different iterations and tree depths. I fit the model using different parameterizations to the test data. Figure 3 illustrates the results and shows that this model fits the test data best with about 100 iterations (this is, in fact,, scikit-learn's default value) and a very simple 2-leaf functional

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14987459

form for its weak predictors.

- Using a gradient boosting tree model with 100 estimators and two leaves in each weak learner and training on films from 2007 to 2011, as mentioned previously, I was able to achieve an R^2 of 0.5400 on the 2012 dataset. The predictions and results are listed in an appendix at the end of this paper. Figure 4 shows a scatter of predictions and actual values.
- The frequency with which a feature is used in the model's decision trees is representative of its importance in generating predictions; highly relevant features will be frequently involved in trees, and ir- relevant features will be involved rarely or not at all. Figure 5 shows the top 10 features. Several features had frequencies of 0, in particular the boolean variables for several of the genre categories, indicating that they could have been completely omitted without impacting the outcomes of this model.

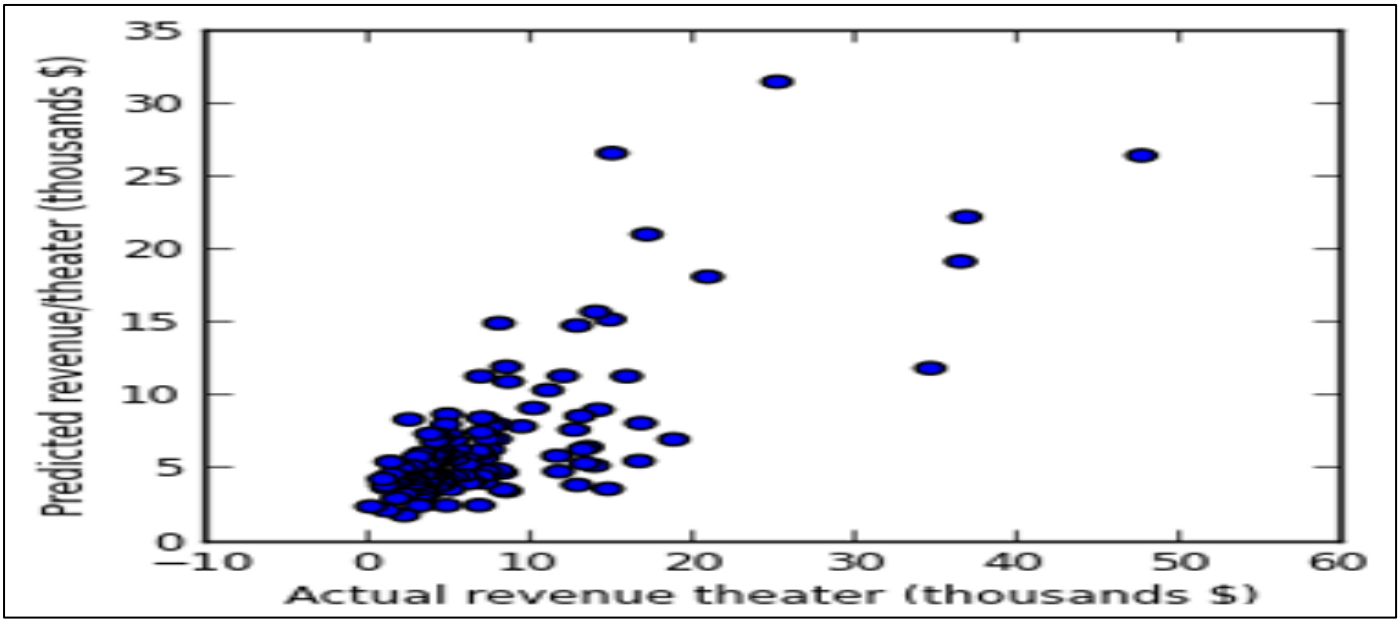

Fig 4 Predicted values vs. actual values.

Table 1 Top 10 Features in the Gradient Boosting Tree Model.

Feature	Frequency (%)
Wikipedia edit runs 7-28 days prior	18.31
Film runtime	14.60
Opening per-theater revenue of similar films	13.30
Wikipedia frequency of headers/subheaders	12.07
Wikipedia edit runs 0-7 days prior	10.97
Wikipedia average size of revisions	9.73
Wikipedia frequency of word "IMAX"	5.07
Wikipedia frequency of external files	4.62
Is comedy	3.74
MPAA rating is PG-13	3.17

- The importance of the Wikipedia data in this model can also be seen by removing the Wikipedia features and rerunning the model, which produces a considerably lower R^2 of 0.3434.

V. CONCLUSION AND AVENUES FOR FURTHER EXPLORATION

- While the results above do show that Wikipedia activity has some ability to predict box office returns, I do not think the model in this paper is precise enough to be used as anything but a very rough forecasting tool. Wikipedia is just one possible source of data for quantifying social interest; social networks such as Twitter or Facebook are another; frequency of appearance in news headlines is another. There are many conceivable metrics to gauge popular interest in seeing a film, and a comprehensive model would include data from many sources.

[6]In fact, I found that the number of opening theaters itself has significant predictive power on per-theater revenue. I omitted it mainly because I wanted to specifically examine Wikipedia's ability to measure social interest.

- In particular, a many-source approach will help overcome the biases that any one source would have. Although Wikipedia is widely known, read, and edited by a wide variety of people, it will still be biased to whatever extent

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14987459

that Wikipedia editors do not reflect the population of people who go to the movies. It is my opinion that the best way to improve this model would be to obtain more measurements of popular interest, particularly data sources whose audiences overlap little with Wikipedia editors - measurements of interest among moviegoing demographics that use the Internet relatively infrequently, for example.

- Nevertheless, the partial success in predicting box office revenues with Wikipedia demonstrates that it is one potential source of data to consider when gauging interest - and not just in films, but anywhere popular interest is a concern. Wikipedia could be used as input for predictions related to interest in news and current events, ticket sales for events other than films, investor sentiments, and many other areas.

Table 2 2012 Predictions and Errors, Sorted by Actual Revenue per theater.

Title	Actual	Predicted	Error (actual - predicted)
Marvel's The Avengers	47698	26452	21247
The Hunger Games	36871	22247	14624
The Dark Knight Rises	36532	19194	17338
The Twilight Saga: Breaking Dawn Part 2 Skyfall	34660 25211	11890 31496	22770 -6285
The Hobbit: An Unexpected Journey Dr. Seuss' The Lorax	20919 18830	18152 7018	2767 11812
The Amazing Spider-Man	17176	21054	-3877
Ted	16800	8127	8673
Think Like a Man	16693	5536	11157

Table 3 2012 Predictions and Errors, Sorted by Actual Revenue per theater (Part 1).

Title	Actual	Predicted	Error
Abraham Lincoln: Vampire Hunter	5247	5668	-421
The Cabin in the Woods	5245	6979	-1734
Sparkle	5189	4511	677
Mirror Mirror	5032	3589	1444
Red Dawn	4916	7430	-2514
The Three Stooges	4892	5981	-1089
Rise of the Guardians	4869	8725	-3856
End of Watch	4818	2503	2315
Cloud Atlas	4787	8046	-3259
Step Up Revolution	4570	4409	162

Table 4 2012 Predictions and Errors, Sorted by actual Revenue per theater (Part 2).

Title	Actual	Predicted	Error
Alex Cross	4489	3955	533
That's My Boy	4440	6258	-1818
Parental Guidance	4392	4140	252
Diary of a Wimpy Kid: Dog Days	4312	5826	-1514
The Dictator	4245	7210	-2965
The Secret World of Arrietty	4235	6930	-2695
The Man with the Iron Fists	4235	6053	-1818
One For the Money	4207	4619	-411
Rock of Ages	4161	7405	-3244
ParaNorman	4108	6899	-2791

REFERENCES

[1]. "Ensemble methods." Retrieved 13 Jan 2012. http://scikit-learn. org/stable/modules/ensemble.html

[2]. Friedman, Jerome H. (19 Apr 2001). "Greedy Function Approx- imation: A Gradient Boosting Machine." Retrieved 10 Jan 2012. http://www-stat.stanford.edu/~jhf/ftp/trebst.pdf

[3]. "Gradient boosting." Retrieved 13 Jan 2012. http://en.wikipedia. org/wiki/Gradient boosting

[4]. "List of hoaxes on Wikipedia." Retrieved 10 Jan 2012. http:// en.wikipedia.org/wiki/Wikipedia:List of hoaxes on Wikipedia

[5]. Pfeiffer, Eric (4 Jan 2013). "War is over: Imaginary 'Bicholm' conflict removed from Wikipedia after five years." Retrieved 10 Jan 2012.

[6]. "Wikipedia." Retrieved 10 Jan 2012. http://en.wikipedia.org/ wiki/Wikipedia