**Predicting Amazon Book Sales Rank using Customer Reviews**

First capstone project report

Yi Zhao

**MAR 20TH 2017**

# Background and Description of Problem

Amazon is a popular ecommerce site that sells a wide range of products. Books are among the many categories. Amazon sells books in hardcover, softcover, and electronic versions, both new and used. Because of the popularity of Amazon sites, there are also a large number of reviews by readers. Some books have even tens of thousands of reviews. In addition to the large number of reviews available for each book, there is also a sales rank number available for each book. This rank is based on the number of items sold in a most recent period of time. The more recent sales there are, the higher the rank is. A rank one would mean the best selling book, and a larger number would mean a lower rank. We hypothesize that information found in readers’ reviews would indicate the rankings of books.

The mythical clients of this project are publishers looking to choose to publish manuscripts that resemble better sales ranking potentials. The goal is to extract features from the reviews of readers to predict the sales ranking of corresponding books.

# Data and description of Features

There were 126 books that represent 5 of the following booklists on Amazon: New York Times best sellers, books to read in a lifetime, biographies & memoirs, children’s books, mysteries & thrillers.

First of all, book lists[[1]](#footnote-1) were scraped to obtain a data frame of meta data that includes the following information of each book in the list: book name, total number of book reviews, book list name, and the URL to the book product page, as well as the Amazon Standard Identification Number (ASIN) that is extracted from this URL. For each book, the product page was scraped to obtain book publisher name, publish date, the average rating from all reviews, the breakup percentage of each review star from 1 to 5 star that is indicative of the distribution of review stars, and the sales ranking in Amazon.

In addition to the product page, all the book review pages were first sorted by dates and then scraped by page. All the book titles were stored in one list, and the review content were stored in a separate list. For books with over 500 pages of reviews, only the first 500 pages were scraped. Less than five books had so many reviews.

The total number of book reviews was formatted by separating comma sign at every thousand digit, so this column was converted first into a string variable, replaced the comma with an empty string, and then converted into an integer.

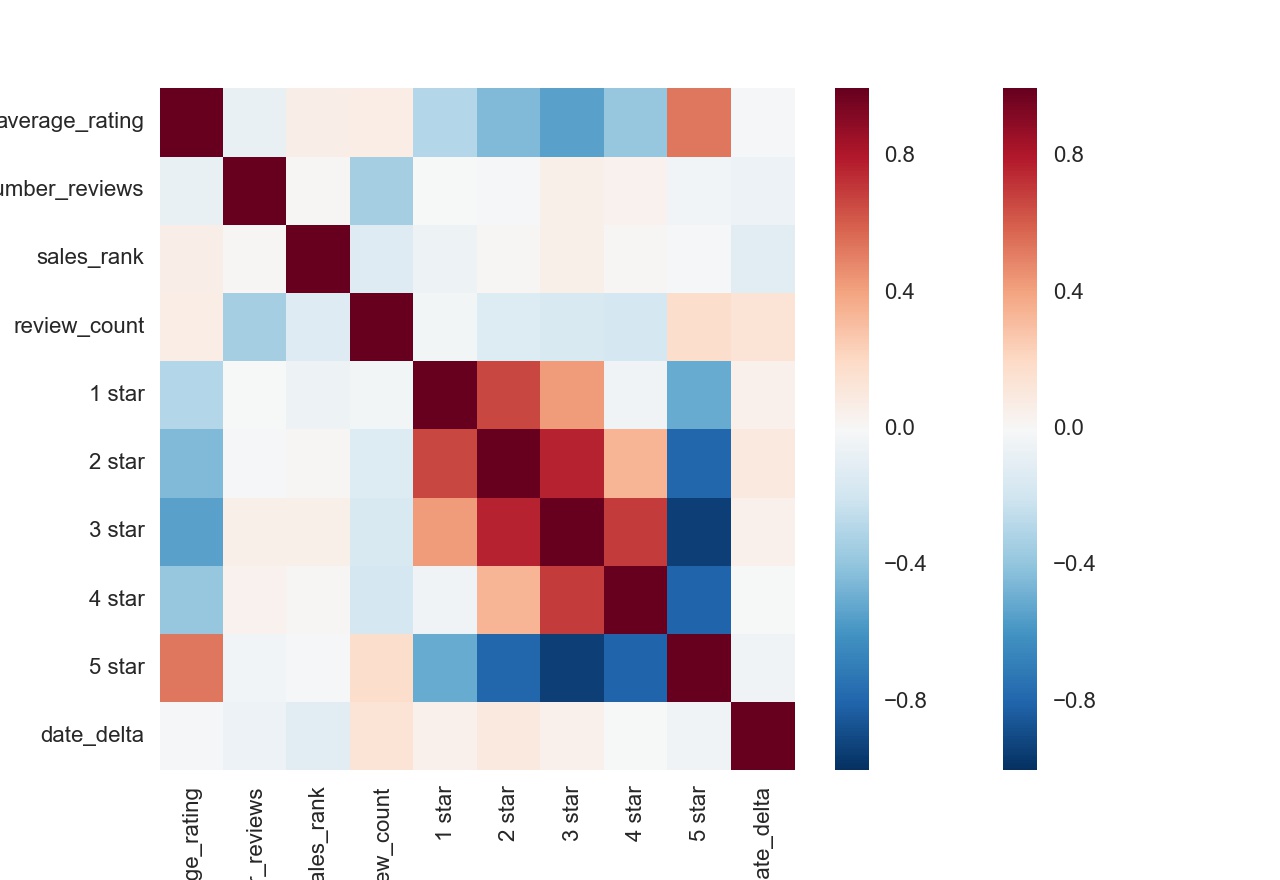
The percentage of each star among all review ratings was first formatted as a dictionary per book, so this variable was expanded into five separate variables with values of their corresponding percentages. For instance, a book may have 0% 1 star reviews, 5% 2 star reviews, 50% 3 star reviews, 35% 4 star reviews and 10% 5 star reviews. All the percentages of each book add up to 100%.

In addition, the review titles and review texts were expanded into arrays of word counts using the CountVectorizer function in sklearn library.

There were 10 predictor variables to sales rank that were not from the reviews: book list name, publisher, publish date delta (the number of days difference from the oldest book in this sample), total number of reviews, average review rating, and percentage of reviews with a rating of 1 star to 5 star. As described in the table below, sales rank had a wide range, because its standard deviation was about 50% higher than the mean. The total number of reviews also had a wide range from 21 to 15,755. In contrast, the average rating had a relatively small range, and the standard deviation was less than 3% of the mean. We probably could not expect much variation in sales rank explained by the average review ratings. However, the percentage of 1 star to 5 star among review ratings may provide more information.

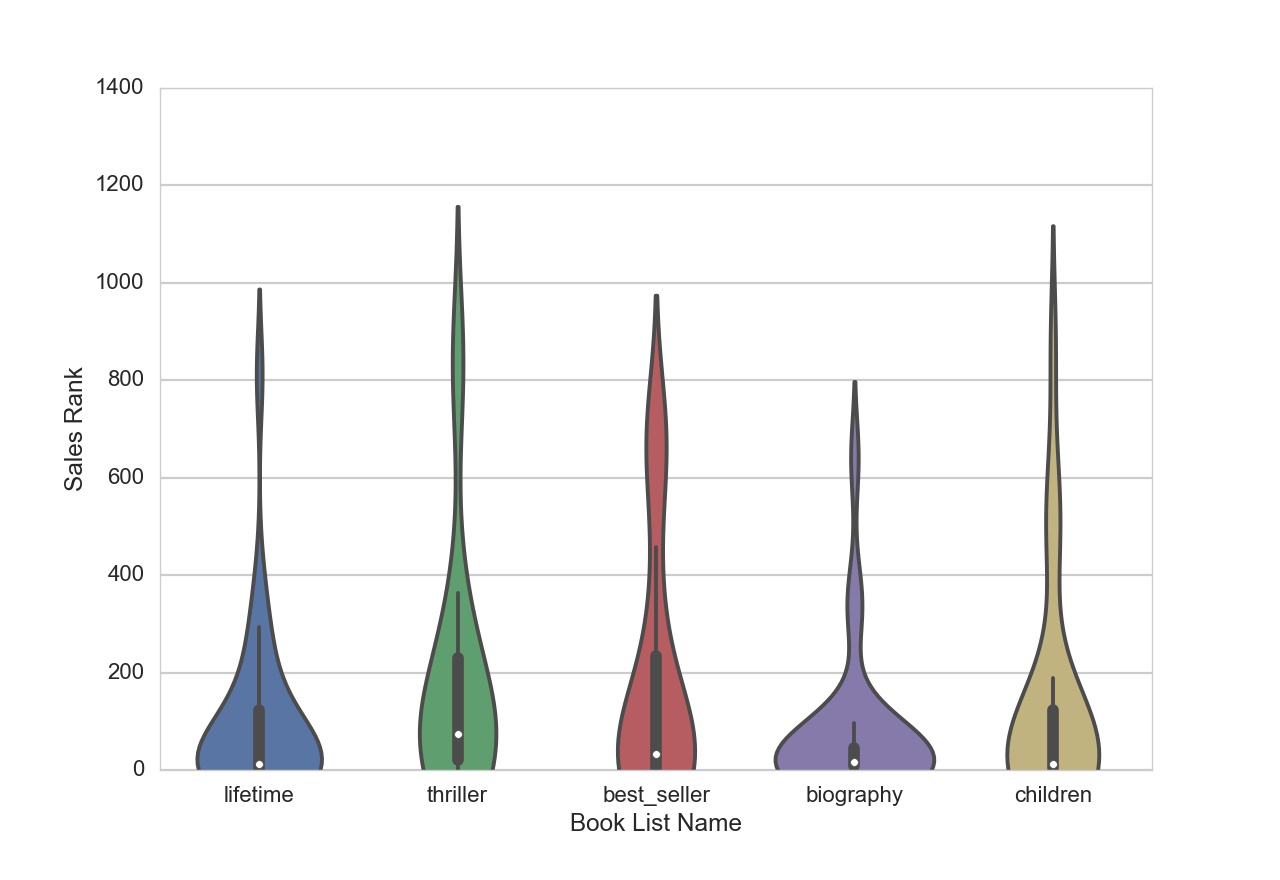
Table Sumary statistics of numeric variables

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Average rating** | **Sales rank** | **Review count** | **1 star** | **2 star** | **3 star** | **4 star** | **5 star** | **Date delta** |
| **count** | 126 | 126 | 126 | 126 | 126 | 126 | 126 | 126 | 117 |
| **mean** | 4.50 | 130 | 1796 | 0.04 | 0.04 | 0.08 | 0.18 | 0.66 | 16492 |
| **Standard deviation** | 0.13 | 215 | 2830 | 0.04 | 0.03 | 0.04 | 0.07 | 0.14 | 4109 |
| **Minimum** | 4.30 | 1 | 21 | 0.00 | 0.00 | 0.01 | 0.04 | 0.32 | 0 |
| **25th percentile** | 4.40 | 4 | 292 | 0.01 | 0.02 | 0.04 | 0.13 | 0.55 |  |
| **50th percentile** | 4.50 | 21 | 802 | 0.03 | 0.03 | 0.08 | 0.18 | 0.66 |  |
| **75th percentile** | 4.60 | 141 | 2057 | 0.05 | 0.05 | 0.10 | 0.23 | 0.78 |  |
| **Maximum** | 4.70 | 876 | 15755 | 0.21 | 0.14 | 0.20 | 0.40 | 0.94 | 20701 |

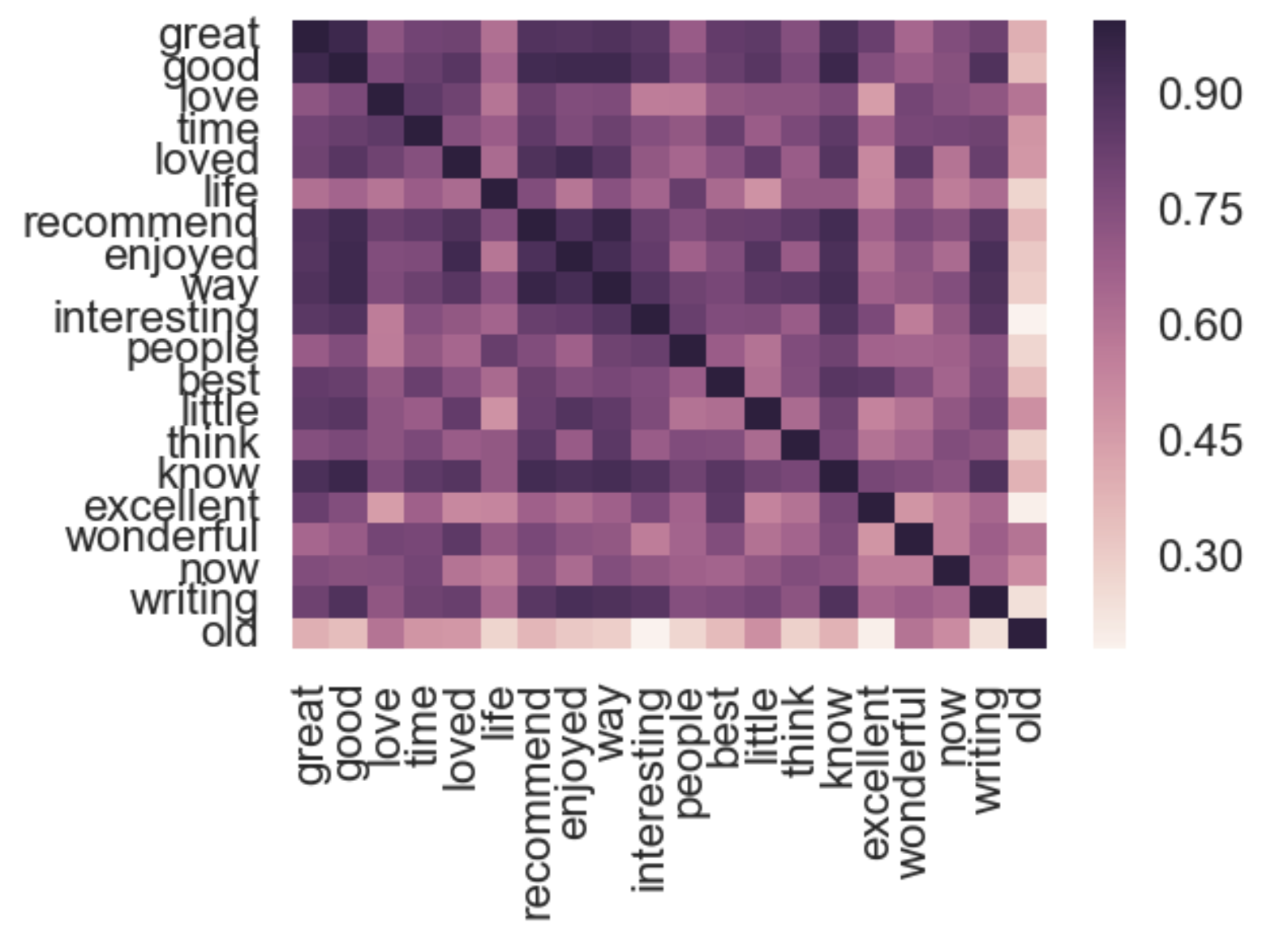
Figure Correlation matrix among non-review text variables

The Pearson correlation coefficients among those variables showed that the sales rank is slightly positively related with the average review rating (Fig 1.). This is counterintuitive in that a larger sales rank means a worse sale. Sales rank is slightly negatively related with the sales rank, which is sensible in that the more people reviewed a book, it is more likely that there have been more people that read this book or more volumes of this book have been sold through Amazon. The percentage of 5 star reviews is negatively correlated with 1 star to 4 star reviews, while the percentage of 1 star to 4 star reviews are highly positively correlated with each other. The publish date delta was a numeric variable computed from taking the difference of the publish date of each book from the earliest publishing date of all books in our dataset. This variable is slightly negatively related to the sales rank, which means that the earlier a book is published, the more likely that is has a better ranking in sales among books of those selected booklists.

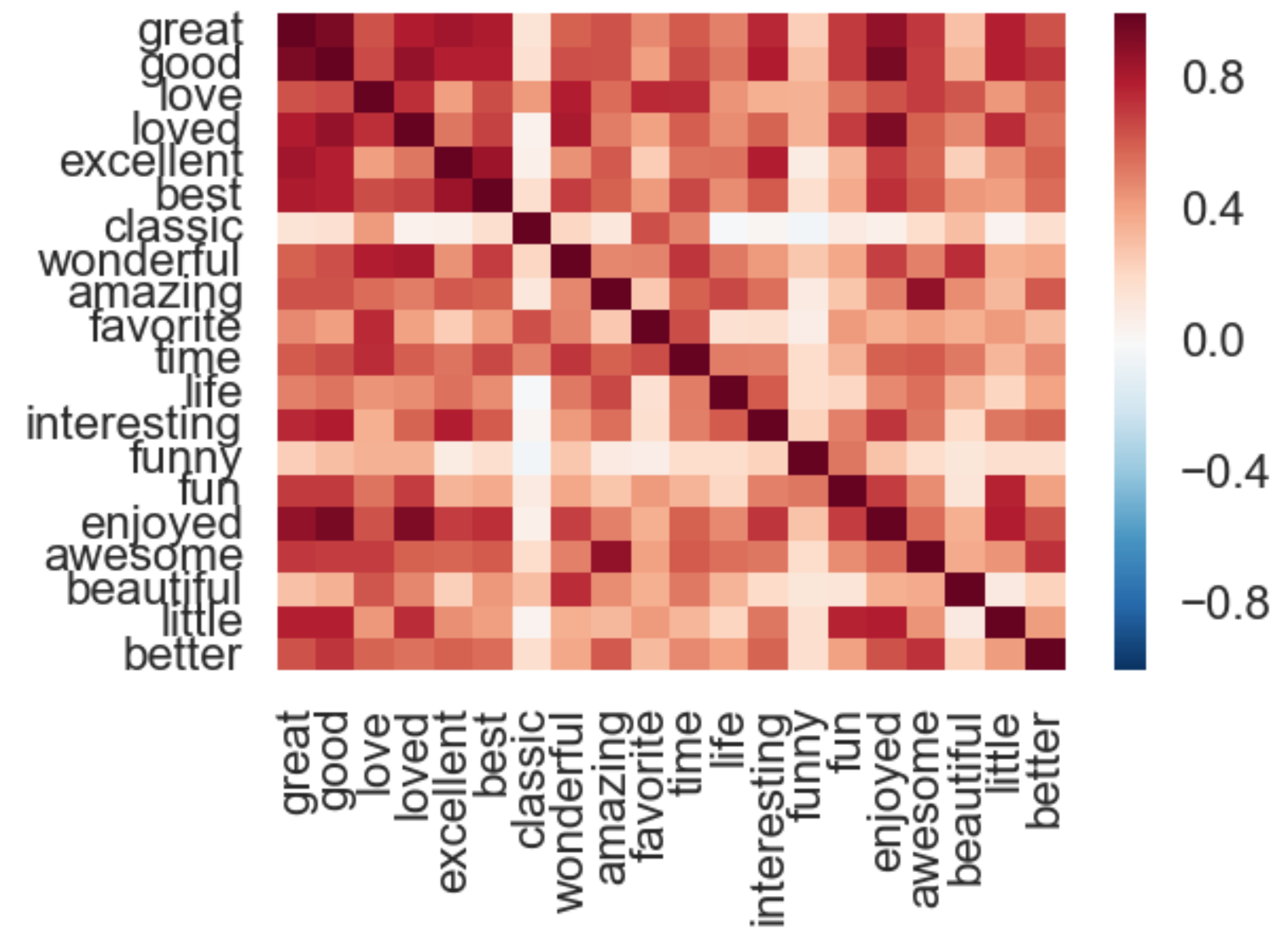
When I ran a simple linear regression of sales rank on the rest of the variables, only number of reviews and the publish date delta were statistically significant at 99% confidence level, and about 10% of the variation in log sales rank was explained according to the adjusted R2 value. Because the sales rank values are highly skewed (Fig 2.), the log transformation was taken for the regression.

Figure The violin plot of sales rank distribution categorized by book list names

The top 20 most frequently occurring words in the review texts were ‘great’, ‘good’, ‘love’, ‘time’, ‘loved’, ‘life’, ‘recommend’, ‘enjoyed’, ‘way’, ‘interesting’, ‘people’, ‘best’, ‘little’, ‘think’, ‘know’, ‘excellent’, ‘wonderful’, ‘now’, ‘writing’, ‘old’. The number of occurrences of those words were highly correlated with each other (Fig 3.)

Figure The correlation matrix of word counts in review texts of books

Similar to the review texts, the top 20 most frequent words in review titles are: ‘great’, ‘good’, ‘love’, ‘loved’, ‘excellent’, ‘best’, ‘classic’, ‘wonderful’, ‘amazing’, ‘favorite’, ‘time’, ‘life’, ‘interesting’, ‘funny’, ‘fun’, ‘enjoyed’, ‘awesome’, ‘beautiful’, ‘little’, ‘better’. Most of those words are positive adjectives, common in both review titles and review texts. The review texts have more neutral words such as ‘people’, ‘know’, ‘now’, etc. The review titles are more succinct in expressing a reviewer’s overall positive or negative view on a book. The high frequency of positive adjectives in review titles is likely because all of the books in our sample came from the Amazon editor chosen book lists, so they predominately have positive reviews.

Figure Correlation matrix of the most frequent 20 words from review titles word counts

# Feature selection and analysis

As shown in the previous section, the sales rank is highly skewed. A log transformation was applied to generate a normal distribution of this dependent variable for the analyses.

There were two types of models used: Lasso linear model and regression tree based non-linear ensemble methods, including Random Forest, adaboost, and XGBoost. Ordinary linear regression was not considered, because there were 126 observations with more than 20K predictor features.

Lasso and Random Forest algorithms have built in feature selection functionality. Those two models were attempted first. However, the resulting average training error (RMSE) was less than 15% lower than the testing error, which means our models were significantly over-fitting. Thus, further feature selections that are separate from predictive model fitting were needed.

I tried three feature selection methods, including F-regression, Mutual Information (MI), and Principal Component Analysis (PCA). In addition, the outputs of those feature selection methods were input to the predictive models in order to assess whether there was any improvement in testing error from the same predictive models but different input features.

The top 50 features selected from the F-regression lead to a lower testing error in Random Forest model compared to using the top 10 features in PCA or top 50 features in MI. The total number of features to include from those feature selection methods was chosen to minimize testing error.

The best performing model in minimizing testing error was XGBoost with the maximum depth of each decision tree to be 1. Compared to random forest and adaboost, XGBoost has the advantage of pruning each regression tree and thus could significantly help with preventing overfitting. The optimum depth parameter was determined using the grid search cross-validated method. Although Random Forest also has a parameter to specify the minimum amount of improvement in explanation of variation from each split to curtail over fitting, the grid search optimized Random Forest model still appeared to have a testing error more than five times as high as the training error. The XGBoost model, when combined with outputs from F-regression, the testing error became 3.2 compared to the training error being 2.1. Therefore, the chosen model is taking the top 50 features from F-regression as input to the XGBoost model using a depth 1 of each regression tree.

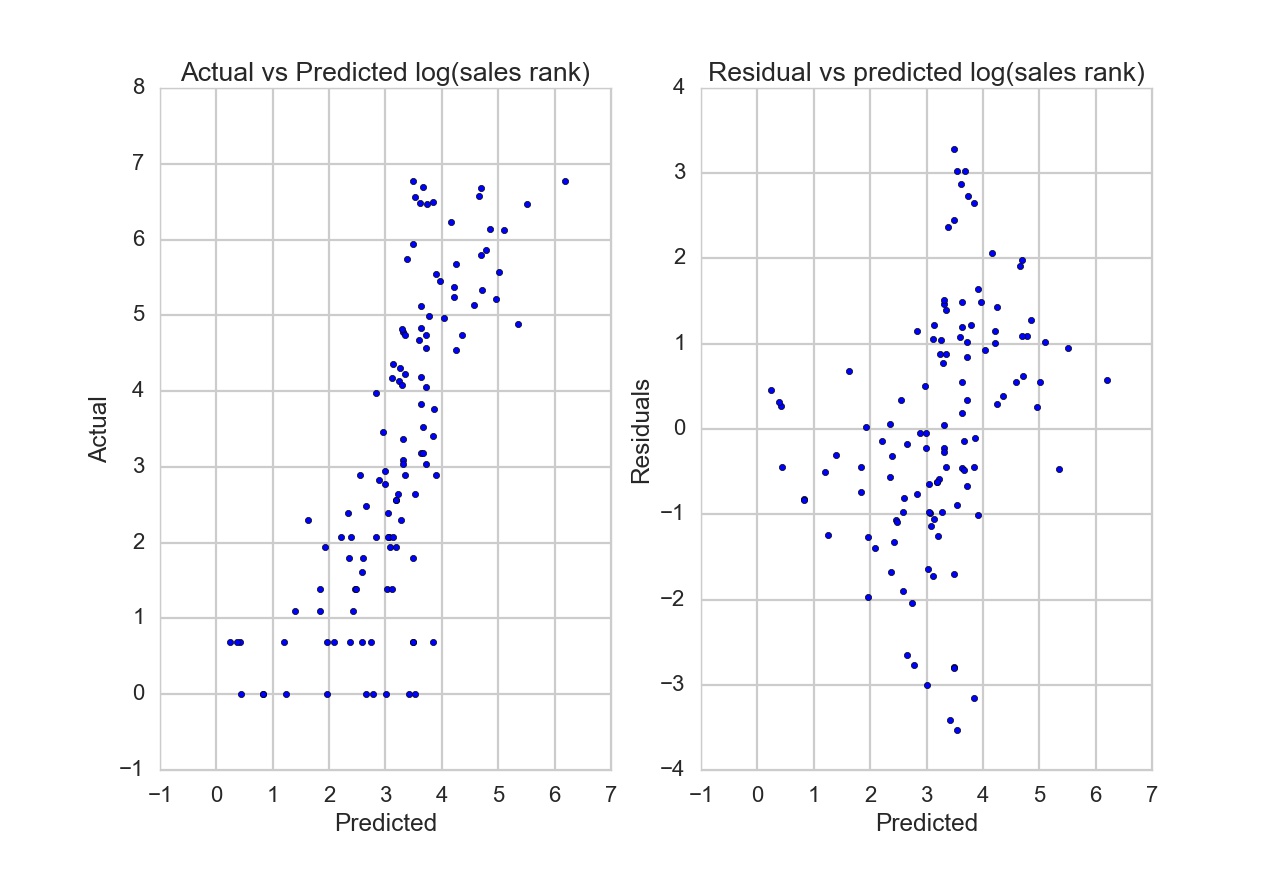
In addition to using word counts as predictive features, I also tried topic analysis with Latent Dirichlet Allocation (LDA). However, the resulting test set R2 was worse when using topic features chosen from LDA, still applying XGBoost model (Fig. 5).

Figure XGBoost Model performance comparison with different features

I ran the randomized 5 fold cross-validated model assessment for 1000 times to generate a distribution of the averaged testing set RMSE from cross-validation. The result is a mean of 3.32 and a standard deviation of 0.16. The corresponding R2 from the testing set is 0.17, and 0.49 from the training set. Thus, we have confidence that our model can explain 17% of log(sales rank) variation in a new random set of books.

The residual, computed as the difference between the predicted and actual log(sales rank), is plotted against both the actual and the predicted values.

Figure Residual plot



There was a significant positive linear relationship between residuals and the actual values (Fig 6). In other words, the log(sales ranking) was underestimated at high values and overestimated at low levels. Yet, the residuals and predicted values did appear to be uncorrelated with each other (Fig 6). The histogram of the residuals did appear to be a normal distribution (Fig 7), so our model is reasonable.

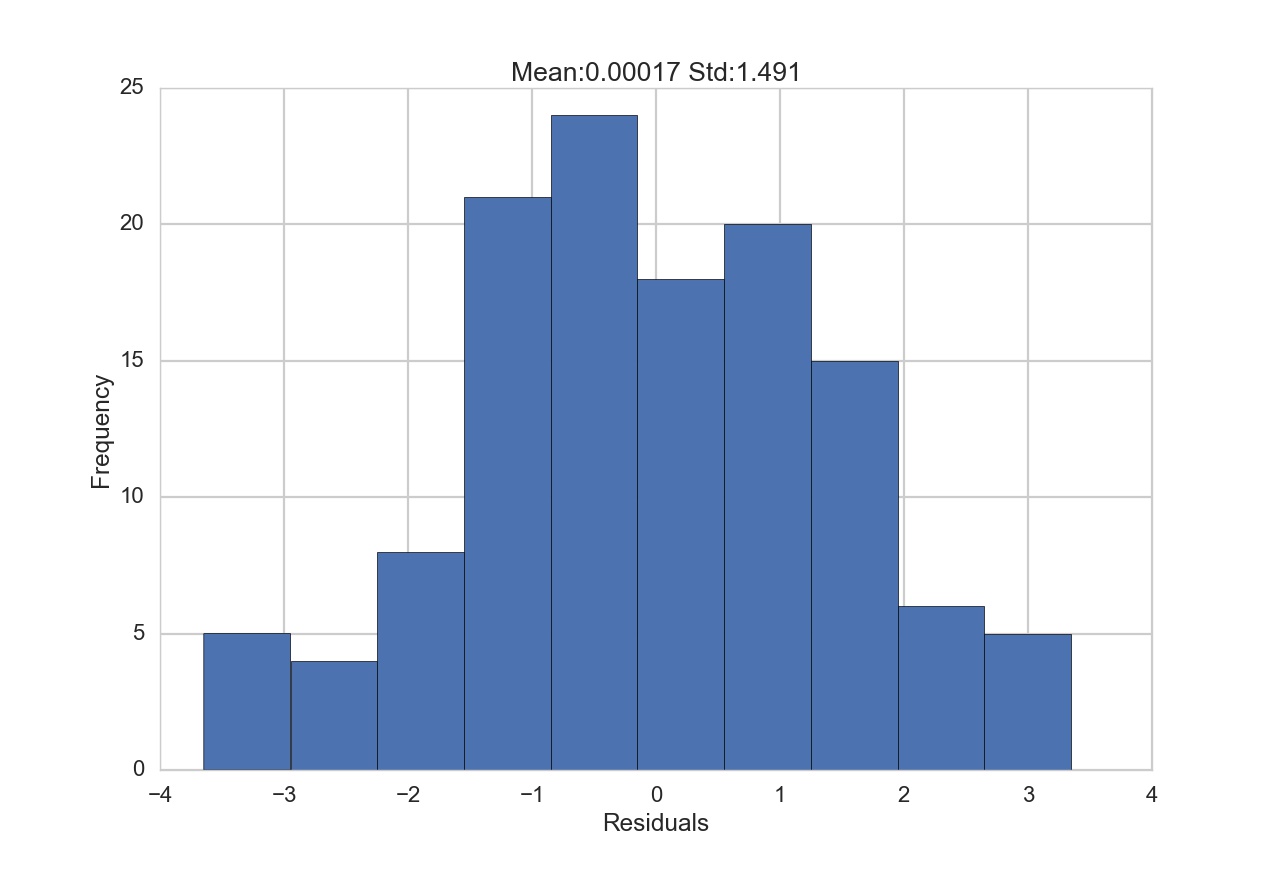


Figure Histogram of residuals

# Conclusion

In this project, the review contents of books on amazon were found to be able to explain 17% of variation in any selection of books’ Amazon sales ranking, which is a relatively weak prediction strength. This was from using a relatively small dataset with 126 books. The features extracted from book reviews are simple word count of each individual word present in the reviews. The total number of features or words is 48,209. Given such a sparse dataset, we were still able to draw important insights from reviewer’s word uses. For instance, words such as soar, emergencies, tequila, grog, devilishly, influences, pronate, mannerisms, dash, eccentric, rocky, journeys, paged, reminisced, humorists, egoistic were among the words whose presence in reviews were the most correlated with the book sales rank. These words may indicate an outstanding characteristic of heroes or plots in a book. In contrast, words such as discounts, chained, concerning, antitheism, nudging, vanity, hallway, erased, strikes, tempest, spade, milky, lifesavers, were the least related with sales ranking. These words are either too common to all books, such as “discounts” and “chained”, or very specific to certain books, such as “strikes”, “spade” are more likely in fights, “milky” is referring to the milky way or the galaxy, and “lifesavers” are referring to star wars, and “vanity” and “antitheism” are likely in philosophical books. Among the most influential words, a higher count of “withdraw”, “paradox”, “dissipated” improved sales rank, and a higher count of “otherworldly”, “cline”, “gait”, “pain” worsened the rank.

Topic analyses using LDA did not improve results. One potential drawback of doing topic analyses using this current dataset is that we have a limited number of observations. Our current feature space is already highly sparse, if adding topics to it, the resulting predictive power of our model would become worse. Ideally, if we have 10K+ books to analyze, a topic analysis may yield much more robust insights.

1. Such as this one: <https://www.amazon.com/b?node=8192263011> [↑](#footnote-ref-1)