Predicting User Video Game Ratings from Amazon Reviews
CSE 190 – Final Assignment

Alexander Ishikawa
A09464040
aishikaw@ucsd.edu

Abstract

Everyday millions of users leave reviews on products using Amazon. These reviews in turn can help other potential buyers figure out whether or not they should purchase the specific item. Figuring out a relationship between a user review and their rating can provide insightful information. This paper sets out to discover a predictive model to accurately predict the user rating given the content of the review and other product information.

I. Introduction

The purpose of this paper is to explore multiple different models to predict the user rating from the review and choose the most accurate among them. The dataset is derived from Amazon and focuses on reviews from video games. The models will be testing different combinations of the dataset's elements to come up with most accurate prediction. Some of the models that are tested include: Naïve Bayes, Linear Regression and Latent Dirichlet. The model with the best accuracy with the most simplistic model will be chosen as the final model.

II. Dataset

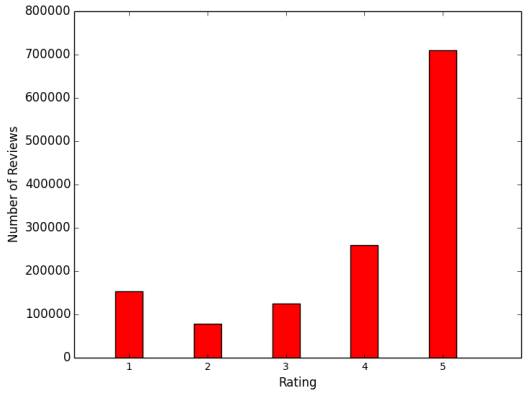
The dataset was derived from Amazon reviews that pertain to the category video games. Each review consists of the following labels:

(1) reviewerID – the ID of the reviewer.
(2) asin – the product ID of the item being reviewed.
(3) reviewerName – the name of the reviewer.
(4) Helpful – the first number is the amount of people who voted the review as being helpful and the second number is amount of people who voted on the review.
(5) reviewText – the entire review in text form
(6) overall – the rating out of 5 that the reviewer gave the product.
(7) summary – a shortened version of the review.
(8) unixReviewTime – time of the review
(9) reviewTime – time of the review in dd/mm/yyyy.

The dataset also includes metadata for the products which includes the labels:

(1) asin – the product ID.
(2) description – description of the item.
(3) imURL – the URL of an image for the item.
(4) related
    (a) also_viewed – items that were viewed by other users who viewed this item
    (b) buy_after_viewing – items that were bought buy users that also bought this item
(5) salesRank – the rank of an item (by number of purchases) in different categories
(6) categories – list of categories that the item fits into

The dataset consists of 1,324,759 reviews from 826,773 users about 50,210 products.
The distribution of the ratings is shown below:



Graph 1: Distribution of Ratings

The data was divided into three sets: training, validations and testing. The first 500,000 reviews were for training the next 200,000 were for validation and the last 624,759 reviews were for testing.

III. Predictive Task

The predictive task is to successfully predict what a user would rate an item based on the review and product information.

1. Accuracy

The final model will be chosen on the accuracy of the performance on the predictive task. This will be measured by the formula for Mean Squared Error:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

Where n is the number of reviews, $\hat{y}$ is the predicted rating and y is the actual rating given by the user.

2. Baseline

The models that are developed will be compared to a basic baseline to determine how well they perform. The baseline for this paper was just using the average rating for all reviews as the predicted value. The MSE for the baseline was:

$$Train\ MSE = 1.8977238448$$

$$Test\ MSE = 1.9271636764$$

3. Data Pre-processing

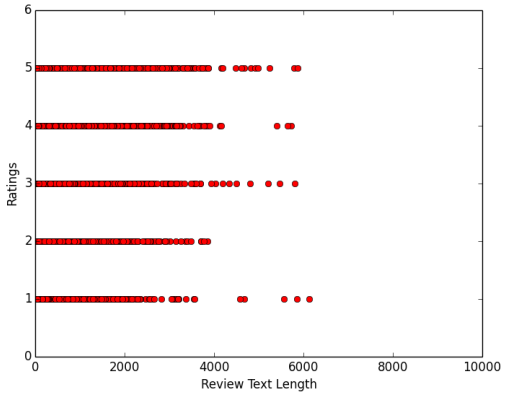
Before evaluating data, the data must be processed in order to obtain a proper result. For reviews without any helpful votes the review is given the average helpful ratio of all reviews. When using the review text as a unigram set all punctuation and stop words must be removed in order to get the correct results.

4. Features

The following are features that can be used in the model and are tested for their relevance using Linear Regression.

(1) Review Text Length

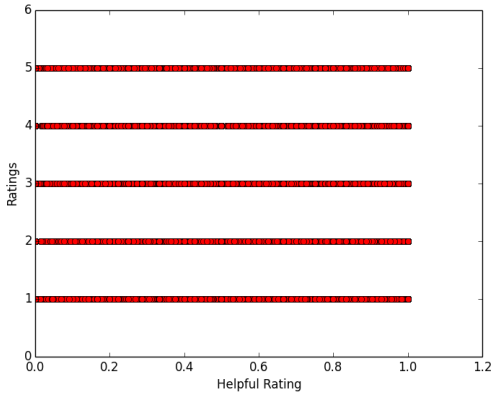
The review text length was found to have a correlation of -3.52602234e-04 which is a very low correlation.



Graph 2: Review Text Length vs Rating

(2) Helpful Ratio

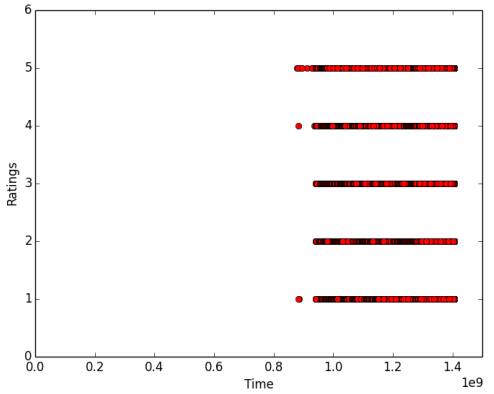
The helpful ratio, which is the number of helpful ratings divided by the total number of votes, was found to have a correlation of 0.72685085, which is a decent correlation.



Graph 3: Helpful Ratio vs Rating

(3) Time

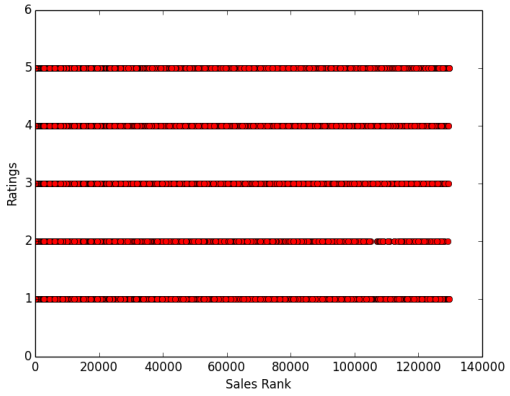
The time of the review was found to have a correlation of 1.33545522e-10 which is a very low correlation.



Graph 4: Time vs Ratings

(4) Sales Ranking

The ranking of an item based on how many times it was purchased. The correlation was found to be -9.42181021e-06 which is very low.



Graph 5: Sales Rank vs Rating

(5) Unigram Set of the Most Popular Words

The unigram set is a set of one word phrases that are used in the review. The most popular words are then used on the model. Some of the more negatively and positively weighted words are listed below:

Positive

1. loves : 0.324274926795
2. perfectly : 0.306762587825
3. condition : 0.289667244041
4. glad : 0.239908943249
5. awesome : 0.23236205006
6. addictive : 0.231401347131
7. love : 0.220792356984
8. excellent : 0.220203731424
9. happy : 0.210360306085
10. works : 0.208344068381

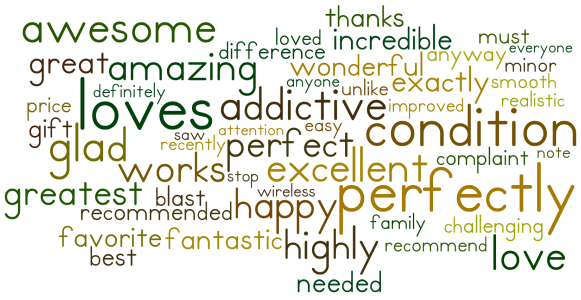


Figure 1: Word Cloud of Top 50 Positively Weighted Words

Negative

1. waste : -0.806060629341
2. worst : -0.629422579678
3. terrible : -0.529016474544
4. horrible : -0.510735687269
5. return : -0.405300621918
6. poor : -0.395943328936
7. disappointed : -0.392176679869
8. boring : -0.343870349693
9. worse : -0.290208796795
10. rent : -0.28831565009

Figure 2: Word Cloud of Top 50 Negatively Weighted Words

IV. Model

1. Relevant Features

Most of the features explored were found to have a very low correlation to the user rating which means that they most likely will not help when determining a useful predictive model. However, the unigram set of the most popular words had a strong correlation meaning that the best model will likely be a variant of the unigram. In addition to the unigram the helpful ratio also showed a correlation. It is possible that the combination of both features could produce a more accurate model.

2. Models Tested

The next step to choosing the best model would be to test the most relevant features using different algorithms and methods to fit the data. The models that are tested in this paper include: linear regression, ridge regression, lasso regression, decision tree regressor and adaboost regressor.

V. Related Literature

(1) Unigram

Given a text or a speech a unigram is a sequence of one word. In this paper a set of unigrams consisting of the most popular words in the given reviews is used as a feature to predict ratings.

(2) Linear Regression

Linear regression assumes a predictor in the form of

$$X\theta = y$$

where X is a set of features from the data, theta is the predictor and y is the real values. [1]

(3) Ridge Regression
Ridge Regression is similar to the previous Linear Regression accept that Linear Regresion aims to minimize:

$$(X\theta - y)^2$$

while Ridge Regression aims to minimize:

$$(X\theta - y)^2 + \alpha(\theta)^2$$

which can minimize some of the overfitting/underfitting that plagues linear regression. [2]

(4) Lasso Regression
Lasso Regression is another regularized form of Linear Regression. Both Lasso and Ridge Regression are very similar. However, one of the main differences between the two is that in Lasso Regression as the penalty is increased some of the parameters will be reduced to zero, while in Ridge Regression all parameters will remain non-zero. This means that Lasso Regression can remove features from regression while Ridge Regression cannot. [3]

(5) Decision Tree Learning
Decision Tree Learning is method that predicts the output by splitting the input feature. [4] A decision tree can automatically detect the most relevant feature. For this paper a Decision Tree Regressor was used to model a predictor to compare to other algorithms.

(6) Adaptive Boosting
Adaptive Boosting or AdaBoost is an algorithm that can be used in conjunction with other types of learning algorithms. AdaBoost trains the new classifier by putting a heavier weight on data that was mislabeled by the previous classifier allowing the current classifier to label the data correctly. [5]

VI. Results

1. Choose Features

Features	**MSE**
Helpful	1.90739421456
Unigram (1000 words)	1.30240768609
Unigram (1500 words)	1.26542039105
Unigram (2000 words)	1.24179610215
Helpful + Unigram (2000 words)	1.23271138567

Since the most useful feature combination has been decided, the next step is to test different models using said features. Since 500,000 reviews for training and 624,759 review for testing takes too much time for my computer the training and testing sets have been cut down to 50,000 each to decide which model performs the best.

2. Linear Regression

$$Test\ MSE = 1.14906563091$$

3. Ridge Regression

alpha	**MSE**
0.0001	1.14906560692
0.001	1.14906539104
0.01	1.14906323239
0.1	1.14904166018
1.0	1.14882735373
10.0	1.14681745378
100.0	1.13521459265

4. Lasso Regression

alpha	**MSE**
0.0001	1.13761841246
0.001	1.1380111236
0.01	1.30615216609
0.1	1.6857342352
1.0	1.74954172755
10.0	1.74954172755
100.0	1.74954172755

5. Decision Tree Regressor

Max Depth	**MSE**
2	1.60471699339
5	1.44320197981
10	1.42967976032
20	1.71485952918

6. AdaBoost Regressor

estimators	**MSE**
5	1.49917957675
10	1.6149280178
50	1.66645045851
100	1.59476507005
200	1.63038914269

VII. Conclusion

Based on the results from the data analysis, the most helpful feature is most definitely the unigram set of the most popular words. The unigram set of words provides a good amount of positively and negatively weighted words which allows a predictive model a good base on which to more accurately predict the real value. The second feature used was the helpful ratio. The helpful ratio also had a decent correlation to the rating therefore adding it to the model saw a small increase in the accuracy of the model. Lastly, the different models were tested. Of all the models tested the Lasso Regression model performed the best with the lowest MSE.

References

[1] Julian McAuley, Lecture 2. pg 4-5. 2015

[2] Ridge Regression. http://scikit-learn.org/stable/modules/linear_model.html#ridge-regression

[3] Zare, Habil (2013). "Scoring relevancy of features based on combinatorial analysis of Lasso with application to lymphoma diagnosis". *BMC genomics* 14: S14

[4] L. Rokach, O. Z. Maimon, Data mining with decision trees: theory and applications World Scientific Pub Co Inc., 2008

[5] Y. Freund, R. Schapire, “A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting”, 1995.

[6] Amazon product data. http://jmcauley.ucsd.edu/data/amazon/.