# Predicting user rating on Amazon Video Game Dataset

CSE190A Assignment2

Hongyu Li
UC San Diego
A98080960
holi@ucsd.edu

Wei He
UC San Diego
A12095047
whe@ucsd.edu

## ABSTRACT

Nowadays, accurate recommendation system becomes more and more important. Based on the product and user review information, it is possible to learn from these reviews and predict the rating these users will give to the new product. This paper mainly describes the dataset we study and perform analysis, the predictive task that we study in this assignment, the model we use to approach this task, the literature related to this problem such as how to train the model, and our results as well as the conclusion.

## 1. INTRODUCTION

The data of reviews and user information has become one of the most important data for companies to expand their business. Meanwhile, personalization of production information has a large impact on customers' production purchase decision and satisfaction in today's competitive market. When a customer visits a product on Amazon or other Online shopping website, the companies always want to know how likely this customer will like this product and purchase it at the end. Therefore, a recommendation system, which involves predicting user responses to options, will become a significant part of Online shopping websites. In addition, relative products will also be recommended to customers in order to increase sales of these products. An accurate recommendation system is more likely persuade customers to purchase more, stimulate the shopping desire of its users and keep users staying on the websites longer. Therefore, we decide to use some models to build up a rating predictor that can be used in a recommendation system using the knowledge we learned from data mining course.

## 2. DATASET ANALYSIS

### 2.1 Dataset

The datasets we choose to use for this project come from the Amazon product data[1]. We will use the per-category data files of Videos Game for this project.

**Table 1: Basic statistics of the Video Games Review dataset**

| | Amount |
|---|---|
| Number of review | 1,324,759 |
| Number of user | 826,773 |
| Number of product | 50,210 |

The Video Games dataset is consisted of two parts: the product review data which contains the reviews each user writes for his/her purchase of the product, and the product metadata which includes the information of each product.

In the Video Games review dataset(Tbl. 1), each review contains the information of the reviewer(user) id, the product id, the reviewer name, the helpfulness rating of the review, the text of the review, the overall rating of the review, the summary of the review, and the time of the review.

In the Video Games metadata dataset, for each product, there are the product id, the product name, the product price, the list of related products, the product sales rank, the brand and the categories the product belongs to.

### 2.2 Analysis

We first analyzed the distribution of the overall ratings of the reviews(Tbl. 2). It shows that more than half of the reviews have an overall rating of 5, and only around 25% of the reviews have a score less than or equal to 3. To our surprise, it turned out that most reviewers, in fact, were satisfied with the game they purchased.

**Table 2: Overall distribution of ratings from 1 to 5**

| Score(1-5) | Percentage |
|---|---|
| Score of 5 | 53.58% |
| Score of 4 | 19.65% |
| Score of 3 | 9.39% |
| Score of 2 | 5.85% |
| Score of 1 | 11.54% |

We then analyzed the relationships between each possible feature available in the dataset and the average overall ratings of the reviews:

(A) ***Review time and average rating:***
Figure 1 describes the relationship between review time

and the average overall rating of the review during that time. The x axis is the review time from Nov. 1997 to July 2014. The y axis is the average overall rating. We can conclude from this figure that the average overall rating does not change too much when the review time changes.

(B) **length of description and average rating:**
Figure 2 describes the relationship between length of description of the product and the average overall rating of the reviews of the product. The x axis is the length of the product description, while the y axis is the average overall rating. It seems that the longer description of the product, the higher rating this product will have. So the length of product description will be a possible good feature to predict user rating.

(C) ***price and average rating:***
Figure 3 describes the relationship between the price of products and the average overall rating of these products. The x axis is the price of the product and y axis is the average overall rating. From the figure, we can see that the higher price of the product, the higher rating it will received. It seems that price will be a possible good feature to predict the rating of new product give to a user.

(D) ***ranking of product and average rating:***
Figure 4 describes the relationship between the the product sales ranking and the average overall rating of it. The x axis is the product sales ranking and y axis is the average overall rating of the product. From the figure, we can see that the higher ranking of the product, the higher rating it will received. So it will be a possible good feature.

(E) ***product count and average rating:***
Figure 5 describe the relationship between the number of reviews of the product and the average overall rating of it. The x axis is the number of reviews of the product, and y axis is the average overall rating of it. From the figure we can see that the average rating seems to increase when the product has more reviews.

(F) ***user count and average rating:***
Figure 6 describes the relationship between the number of reviews of the user and the average overall rating. The x axis is the number of reviews of the user and y axis is the average overall rating. From the figure, we can see that while the average overall rating vary because of the different number of reviews of the user, there is no a clear tendency of that.

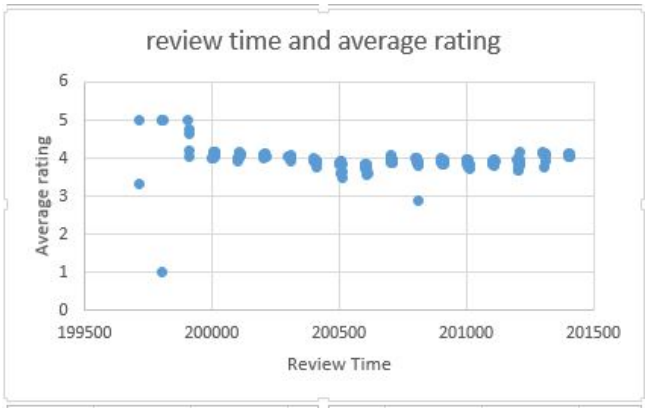


**Figure 1.** Relationship between review time and average overall rating score

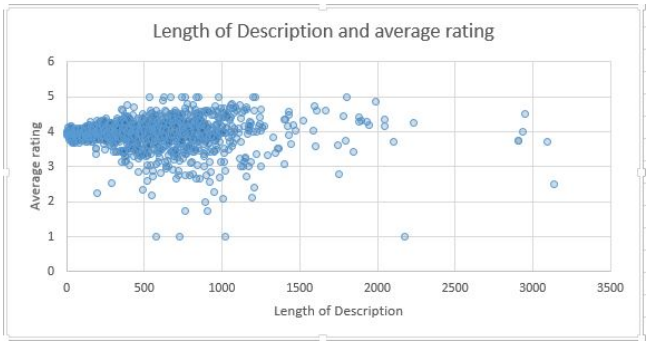


**Figure 2.** Relationship between length of description and average overall rating score


**Figure 3.** Relationship between price of product and average overall rating score

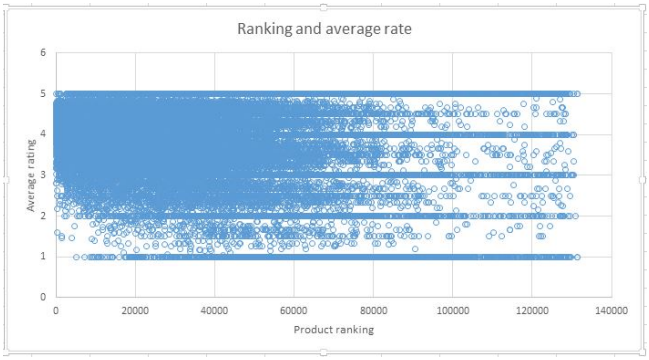


**Figure 4.** Relationship between product sales ranking and average overall rating score

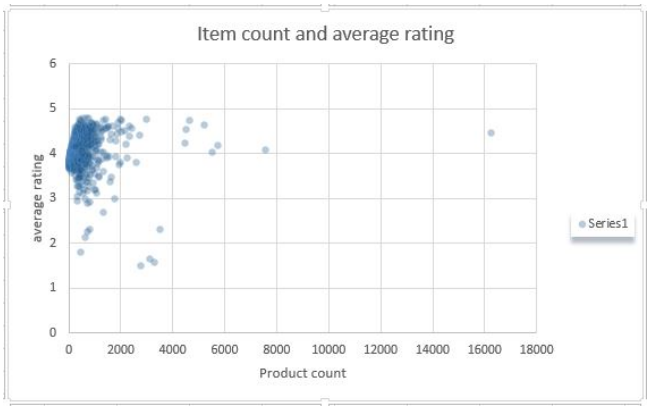


**Figure 5.** Relationship between product reviews amount and average overall rating score

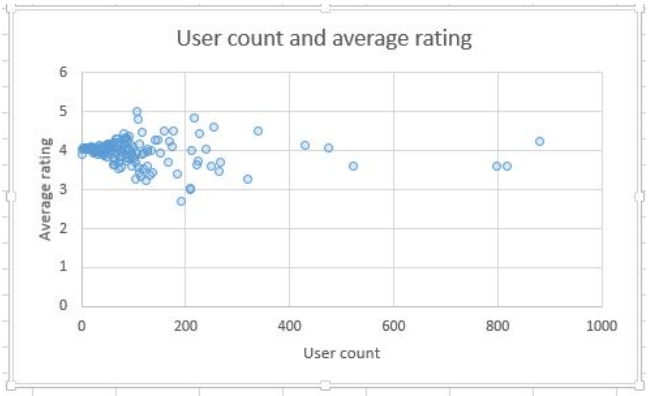


**Figure 6.** Relationship between user reviews amount and average overall rating score

## 3. PREDICTIVE TASK

In this project, we will try to predict the overall rating a user may give to a game he/she purchase based on the previous reviews data, and the product data. If we can come up with a accurate model, it could be used in a recommendation system to predict whether or not a user will enjoy the recommended products.

In order to evaluate our models' accuracy, we will be using the Mean Squared Error (MSE) between the actual rating and our predicted rating.

To test our models, we will use the every fifth data from the Video Game review set as the test set, and use the rest of the review set and the Video Game metadata set as the train set. In such a way, we will test the our models with 20% of the dataset in a more random way than just using the last 20% of the dataset.

Since our goal is to predict a user's satisfaction of a possible future purchase, when doing the prediction, the predictor will only have the access to the user id and the product id of the test set. No other review information of the testing pair of user and product will be available to the predictor since it is treated as a possible future purchase.

The baseline model calculates the mean rating of each user's reviews. It will then use the mean rating as the predicted rating. If the user has never been seen in the train set, the predictor will use the global review rating instead.

## 4. PREDICTION MODEL

After evaluate the relationship between every feature in the dataset and the overall rating score, we decide to use a regression model to predict the overall rating. Such regression model will estimate the relationship between these features and the overall rating. Because our task is a predictive task instead of a classified task, a regression model will work better than other classification models.

Regression analysis is a process to learn the relationships between features and parameters to predict real valued outputs. There are several regression models we have tried to predict the product rating in this assignment: the simplest Linear Regression, the Ridge Regression, and the Random Forest Regression.

1. **Linear Regression:**
   $rating(u, p) = \theta_0 + \theta_1 \times F_1 + ... + \theta_n \times F_n$
   where $u$ is the reviewer ID, $p$ is the product ID, and F represents each feature we use in this linear regression. In this model, we have used each single feature in the lists above and decide which features we used in the combination one based on the Mean Square Error of each.

   For user rating and product rating features, we change the model to become as following:

$$r(u,p) = \begin{cases} \text{PR}(p), \text{ if } u \notin \text{users and } p \in \text{products} \\ \text{avgRating, if } u \notin \text{users and } p \notin \text{products} \\ \theta_0 + \theta_1 \times UR(\text{u}) + \theta_2 \times PR(\text{p}), \\ \text{if } u \in \text{users and } p \in \text{products} \end{cases}$$

   where PR is average rating of particular games $p$;
   avgRating is the average rating of all video games;
   UR is the average of all rating this user $u$ gives.

   The Mean Square Error of this model is 1.8730921582, which is slightly better than the baseline. However, we want to improve the performance more.

   In order to optimize it, we change the model to include more features and hope to make more improvement:

$$r(u,p) = \begin{cases} \text{PR}(p), \text{ if } u \notin \text{users and } p \in \text{products} \\ \text{avgRating, if } u \notin \text{users and } p \notin \text{products} \\ \text{w,} \\ \text{if } u \in \text{users and } p \in \text{products} \end{cases}$$

where $w = \theta_0 + \theta_1 \times PR(\text{p}) + \theta_2 \times PC(\text{p}) + \theta_3 \times Price(\text{p})$
PC is number of review of product $p$

The Mean Square Error of this model is 1.868723698, which is better than the first model.

For first model, it has the issue of overfitting the training set. Therefore, for its MSE on test set will higher than the second model.

The third model we try is:

$$r(u,p) = \begin{cases} \text{PR}(p), \text{ if } u \notin \text{users and } p \in \text{products} \\ \text{avgRating, if } u \notin \text{users and } p \notin \text{products} \\ \text{w,} \\ \text{if } u \in \text{users and } p \in \text{products} \end{cases}$$

where $w = \theta_0 + \theta_1 \times UR(\text{p}) + \theta_2 \times PR(\text{p})$, UR represents user average rating, PR represents product average rating.

Surprisingly, the mean square error of this model is 2.008625895, which is worse than the baseline. The reason of it may be user average rating and product average rating are not independent variable, while the variable user average rating has worse Mean Square Error as shown in table 3.

Compared to other two models, the strengths of linear regression are this model is the simplest to use, and its widespread availability. It always be the first model we use in a predictor.

The weakness of linear regression are it presumes that a linear model is the appropriate theoretical model to represent the behavior we analyze, and it only looks at linear relationships. If the useful feature does not have linear relationship with output, it will not be a good predictor. In addition, it assumes all data are independent, which shown in the data set analysis that it is not possible. Some data have closer relationship with others.

2. **Ridge Regression:**
The motivation to use ridge regression are that we have seen the model included all useful features in linear regression does not have too much better performance than baseline, and the number of input variables exceeds the number of observations. So we choose ridge regression, which is able to abandon the requirement of an unbiased estimator.

$\hat{\beta}_{ridge} = (X'X)^{-1}X'Y$
which is chosen to minimize the penalized sum of square $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$

In ridge regression, we use two models the same as linear regression to test its performances.
First model:

$$r(u,p) = \begin{cases} \text{PR}(p), \text{ if } u \notin \text{users and } p \in \text{products} \\ \text{avgRating, if } u \notin \text{users and } p \notin \text{products} \\ \theta_0 + \theta_1 \times UR(\text{u}) + \theta_2 \times PR(\text{p}), \\ \text{if } u \in \text{users and } p \in \text{products} \end{cases}$$

where PR is average rating of particular games $p$;
avgRating is the average rating of all video games;
UR is the average of all rating this user $u$ gives.

The Mean Square Error of this model is 1.8623981275, which is better than the baseline. However, we want to improve the performance more.

Second model:

$$r(u,p) = \begin{cases} \text{PR}(p), \text{ if } u \notin \text{users and } p \in \text{products} \\ \text{avgRating, if } u \notin \text{users and } p \notin \text{products} \\ \text{w,} \\ \text{if } u \in \text{users and } p \in \text{products} \end{cases}$$

where $w = \theta_0 + \theta_1 \times PR(\text{p}) + \theta_2 \times PC(\text{p}) + \theta_3 \times Price(\text{p})$
PC is number of review of product $p$

The Mean Square Error of this model is 1.857289318, which is better than the baseline (1.89718136279) and the first model of ridge regression.

The third model we try is:

$$r(u,p) = \begin{cases} \text{PR}(p), \text{ if } u \notin \text{users and } p \in \text{products} \\ \text{avgRating, if } u \notin \text{users and } p \notin \text{products} \\ \text{w,} \\ \text{if } u \in \text{users and } p \in \text{products} \end{cases}$$

where $w = \theta_0 + \theta_1 \times UserC(\text{p}) + \theta_2 \times ProdC(\text{p}) + \theta_3 \times PR(\text{p})$, UserC stands for number of review of user u, ProdC stands for number of review of product p.

The Mean Square Error of this model is 1.889612502, which is worse than the baseline. The reason of it may be user number of review is not a correlative feature to the product rating.

Compared to linear regression, the strengths of ridge regression is that it puts further constraints on the parameters $\beta_j$, in the linear model.

The weakness of ridge regression is that it penalizes the size of the regression coefficients and it is not able to zero out coefficients. So we have to either end up including all the coefficients in the model, or none of them in the model.

3. **Random Forest Regression:**
Based on the general performance of previous two models, linear regression and ridge regression, we decide to select another model to compare with these two in order to find out which model has the best performance. We choose random forest regression, which is model that is applicable to handle categorical predictors naturally.

In random forest regression model, we only use one model because of time limit issue.

$$r(u,p) = \begin{cases} \text{PR}(p), \text{ if } u \notin \text{users and } p \in \text{products} \\ \text{avgRating, if } u \notin \text{users and } p \notin \text{products} \\ \text{w,} \\ \text{if } u \in \text{users and } p \in \text{products} \end{cases}$$

where $w = \theta_0 + \theta_1 \times PR(\text{p}) + \theta_2 \times PC(\text{p}) + \theta_3 \times Price(\text{p})$
PC is number of review of product $p$

The mean square error of this model is 1.86747235, which is worse than the baseline (1.896615128) and the same model using ridge regression.

The advantages of random forest regression model are accuracy and instability. For accuracy, it is competitive with the best known machine learning methods. For instability, if we change the data a little, the individual trees may change but the forest is relatively stable because it is a combination of many trees.

The disadvantages of random forest regression model are overfitting and groups favors. it will overfit the dataset with noisy regression tasks. In addition, it will more favor smaller groups over large groups if dataset contains groups of correlated features of similar relevance for the output.

# 5. FEATURE SELECTION

From looking at the dataset, we find several possible features that may related to the overall rating prediction: user's average rating, product's average rating, number of reviews of user, number of reviews of product, review time, product description length, product price, and product sales ranking.

After analyze the relationship of these possible features and the overall rating, we find out that some of these features seem to be not as good as other, such as the review time and the number of review of user. These "bad" features do not show an apparent relationship with the overall rating.

In order to identify the "good" features and the "bad" features, we then run a Linear Regression model with each feature to see how the MSE changes (Tbl. 3).

**Table 3: Feature testing with Linear Regression**

| Feature | MSE |
|---|---|
| No feature | 1.89718136276 |
| User average overall rating | 2.0118332215 |
| Product average overall rating | 1.62454190501 |
| Number of reviews of user | 1.89720275375 |
| Number of reviews of product | 1.89318761661 |
| Product description length | 1.89608503372 |
| Product price | 1.89027608356 |
| Product sales ranking | 1.88811920473 |

From the feature testing, we can see that using features like user average overall rating, and number of reviews of user result worse MSE values than using a constant prediction rating. Such features appear to be "bad" features that will make the predictor less accurate.

Hence, the features we are going to use for our models are: product average overall rating, number of reviews of product, product description length, product price, and product sales ranking.

# 6. RESULTS

**Table 4: MSE on different models**

| Model | MSE |
|---|---|
| Baseline Mean Predictor | 2.03909764788 |
| Linear Regression | 1.62454190501 |
| Ridge Regression | 1.62454156635 |
| Random Forest Regression | 1.62944738429 |

Table 4 shows the Mean Squared Error achieved by each model with using all the features we select. It turns out that all Regression models are having a better result than the Baseline model. Among these Regression models, Random Forest Regression model has a worse performance than the other two. The Ridge Regression model appears to be our best model that is a little better than the Linear Regression model.

As mentioned in prediction models, each model has its own advantages and disadvantages. The reason for Ridge Regression model has the best performance among these three models is that it puts further constraints on the parameters and it mainly focuses on the $X'X$ predictor correlation matrix.

However, for Random Forests Regression, the reason for it has the worst performance among these three models is that it overfits the training set with noisy regression task.

# 7. RELATED LITERATURE

The dataset we choose to study is available on the Amazon product data page hold by Prof. Julian McAuley. There is no paper only using this Video Games dataset, but there are several papers of Prof. Julian that build models with using all the Amazon product data including this Video Games one.

One model builds a network of substitutable and complementary products for the recommender system[3]. This is a different approach from our Regression models. This is a network model that in built on the relationship between each products. After building the network, the model is able to identify the products related to the one the users have purchased. As a result, related products will be recommended to the users.

Another model builds a predictor to predict how a user will response to a product[2]. This is a similar predictive task as we have. However, it is considering different factors of the product and user. While our model uses price, description, and average rating as features to predict the overall rating. This model evaluate the age of the product, the age of the user, and the state of the society in order to predict how a user will like a product. It considers that people's tastes are changing overtime.

These models are using the state-of-art technologies to predict people's preference. They take people's needs and people's bias over time into account. While our model may work well on a past dataset, these models are predicting

the future. They consider the most recent purchase in the database, and predict as an up-to-date recommender system. It concludes that prediction based on the old data may not meet people's today's preference.

# 8. CONCLUSION

Among the three Regression models we have tried, Ridge Regression model performs the best prediction over the Video Games dataset. The reason is that Ridge Regression has strict constraints on the parameters in its model, which will reduce the mean square error on testing set.

Among the features we have tried, product average overall rating, number of review of product, product price, product description length, and product sales ranking are having positive effect to the testing results. The parameters of these features show that: product average overall rating has the largest effect to the predicted rating, which indicated that most users' preferences on Video Games are similar.

With the model and features we select, there is a significant improvement comparing with the baseline model, which means our features and model selections are on the right track for our predictive task. It shows that most users are having a common way to evaluate the game they play.

# 9. REFERENCES


[1] Amazon product data, http://jmcauley.ucsd.edu/data/amazon/links.html.

[2] J. McAuley and J. Leskovec. From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews. *WWW*, 2013.

[3] J. McAuley, R. Pandey, and J. Leskovec. Inferring Networks of Substitutable and Complementary Products. *Knowledge Discovery and Data Mining*, 2015.