Predicting User Ratings and Reviews on the Yelp Dataset with Machine Learning Techniques

Yuning Sun and Jiayu Li and Penggao Gu

Abstract

In this project, we employed user reviews as a predictor to construct several models for rating prediction. We adopted two distinct approaches to analyzing ratings, including binary classification that considered ratings below 4 as low and those above 4 as high, and multi-class classification based on the original 1-5 star rating scale. Among all the binary classification models, SVM demonstrated the highest accuracy in predicting ratings, and the accuracy in binary classification consistently outperformed multi-class classification. Moreover, we performed K-means clustering on five variables related to users' rating and reviewing characteristics to identify two clusters for market segmentation. Overall, the ratings and reviews feature on Yelp is a valuable tool for both consumers and businesses.

Keyword: K-means clustering, Binary classifications, Multi-class classifications

1 Introduction

The restaurant rating system is a method of evaluating the quality of food and service at restaurants. It originated in the early 20th century and gained popularity with the Michelin Guide's three-star rating system. With the advent of the internet and social media, restaurant ratings have become more important than ever. Online platforms like Yelp, TripAdvisor, and Google Reviews allow customers to leave feedback and ratings for restaurants they have visited, which can have a significant impact on a restaurant's reputation and business success.

This project uses Yelp's dataset of millions of raw restaurant reviews from their 2014 Dataset Challenge. The dataset was narrowed down to only include California restaurants and users. The project's focus is on generating a more useful rate by categorizing users based on their review quality, training rate-predicting models, and using high accuracy rate-predicting model to generate new rates based on the group of users with high-quality reviews. To achieve this goal, the project aims to answer two questions 1) Which machine learning algorithm is the most useful for predicting review quality? 2) How to clustering the users based on their review quality? The project builds regression models with three machine learning algorithms to predict user sentiment: random forest, logistic regression, K nearest neighbor, and support vector machine. We then applied K-means clustering to a set of 5 variables associated with users' rating and reviewing attributes. The result of this clustering procedure was the creation of two distinct clusters that provide insight into the market segmentation. The performance of each model is analyzed to determine the best model for predicting ratings from reviews.

1.1 Notation

Our models are based on K-Means clustering[1], an unsupervised learning algorithm that groups unlabeled datasets into distinct clusters. It allows us to cluster data into different groups and is a convenient way to discover group categories in unlabeled datasets on its own without any training.

number of data point n
data point dimensions d
data point $x_i, i \in \{1, ..., k\}$

number of clusters k
data point clusters number i for $x_p \in S_i^{(t)}$ in interaction t where cluster
$S_i^{(t)} = \{x_p : ||x_p - m_i^{(t)}||^2 \leq ||x_p - m_j^{(t)}||^2 j, 1 \leq j \leq k\}$
centroid $\mu_i, m_i^{(t)}$in interaction t, $i \in \{1, ..., k\}$
optimization function
$argmin_S \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2$

1.2 Organization

In Section 1.1, we take the time to define and clarify the notation used in this study. Moving on to the fundamental mathematical principles of our models in Section 2, we provide comprehensive discussions of SVM, KNN, and random forest classification. As we move to Section 2.1, we delve into Logistic Regression models, followed by the presentation of Random Forest classification in Section 2.2, K-Nearest Neighbors method in Section 2.3, and Support Vector Machine models in Section 2.4. To provide robust evidence for the effectiveness of each model, we present methods and experimental results in Sections 3.1 to 3.4, and summarize our findings in Section 3.5. Finally, we conclude the paper by drawing insightful conclusions from our research and exploring other possible applications of our models in Section 4.

2 Mathematical Concepts or Foundations

We will build four different types of regression with three machine learning algorithms to predict user sentiment - random forest, logistic regression, K nearest neighbor, and support vector machine. We analyze the performance of each of these models to come up with the best model for predicting the ratings from reviews.

2.1 Logistic Regression

Logistic regression estimates the conditional probability:

$$P(\mathrm{Y} = 1 \mid \mathrm{X}) \tag{1}$$

In logistic regression, it is assumed that

$$P(\mathrm{Y} = 1 \mid \mathrm{X} = \mathrm{x}) = \frac{\exp(\beta_0 + \beta^T \mathrm{x})}{1 + \exp(\beta_0 + \beta^T \mathrm{x})} \tag{2}$$

where
$x = (x_1, ..., x_p)^T$ is a p-dimensional predictor
β_0 and $\beta = (\beta_1, ..., \beta_p)$ are unknown parameters
$\beta^T x = \Sigma_{l=1}^{p} \beta_i x_i$

Why it works: some intuitions

Logistic regression is a statistical method that is commonly used for binary classification tasks. In the case of predicting user ratings on Yelp, the task is to predict whether a user will give a positive or negative rating to a particular business.

Logistic regression works well in this scenario because it is a linear model that can capture the relationship between the input features and the output variable. It is also a probabilistic model that outputs probabilities of each class, which can be interpreted as the likelihood of a user giving a positive or negative rating to a business based on the input features.

Furthermore, logistic regression[3] is a simple and interpretable model, which makes it easy to understand how each input feature contributes to the prediction. This can be useful for businesses to identify which features are important in influencing user ratings and improve their overall ratings.

2.2 Random Forest

Random forest is similar to bagging, but different as it averages de-correlated trees.

(1) Draw bootstrap samples $(x_1^b, y_1^b), ..., (x_n^b, y_n^b)$
(2) For each bootstrap sample, grow a tree by repeating:
  a. Randomly select a subset of the p variables
  b. Pick the best split among the chosen subset
  c. Split the node
(3) The random forest estimate is constructed similarly as in bagging (average for regression, and majority rate for classification)

Why it works: some intuitions

(1) Non-linearity: The relationship between the number of stars given by a user and the length of their review might not be a simple linear relationship The random forest can capture these non-linear relationships by creating multiple decision trees, each of which is trained on a subset of the data and uses different sets of variables to make predictions.

(2) Handling missing data: Yelp data can be messy and incomplete, with missing values and outliers. The random forest can handle missing data by imputing missing values based on the values of other variables, reducing the risk of biased predictions.

(3) Robustness: Random forest is a robust algorithm that can handle noisy and correlated input variables, which is common in Yelp data. By aggregating the predictions of multiple decision trees, it can also reduce overfitting, which occurs when a model becomes too complex and performs poorly on new data.

(4) Interpretable: Random forest is relatively easy to interpret, making it a popular choice for data analysis. The feature importance measures provided

by random forest can help identify which factors are most important in predicting user ratings on Yelp.

2.3 K-nearest Neighbors

Let $D = \{(x_i, y_i)\}_{i=1}^{n}$ be the training dataset. Then K-nearest neighbor classification is as follows.

(1) Specify an integer value for K (K is an odd value).
(2) For $X = x_0$, find the K points in the training dataset that are closest to x_0. Denote this subset of points as N_0.
(3) Then the probability $P(Y = 1|X = x_0)$ is estimated as

$$P(Y = 1 \mid X = x_0) = \frac{1}{k} \sum_{i \in N_0} l(y_i = 1) \qquad (3)$$

(4) We then classify as follows:

$$\text{f(x)} = \begin{cases} 1, & if\ P(Y = 1 \mid X = x_0) > 0.5, \\ -1, & if\ P(Y = 1 \mid X = x_0) \leq 0.5 \end{cases}$$

Why it works: some intuitions

The reason why KNN works well for predicting user ratings on Yelp is because it relies on the idea that similar businesses will have similar ratings. By finding the K nearest neighbors to a target business, the algorithm is able to make a prediction based on the ratings of those similar businesses. Additionally, KNN is a simple and easy-to-understand algorithm that doesn't require much tuning or parameter optimization, which makes it a good choice for predicting user ratings on Yelp.

2.4 Support Vector Machine

The support vector machine has been chosen because it represents a framework both interesting from a machine learning perspective and from an embedded systems perspective. A SVM is a linear or non-linear classifier, which is a mathematical function that can distinguish two different kinds of objects.

Algorithm Training an SVM

Require: X and y loaded with training labeled data, $\alpha \leftarrow$ 0 or $\alpha \leftarrow$ partially trained SVM
1: C $\leftarrow$ some value (10 for example)
2: **repeart**
3: **for all** $\{x_i, y_i\}, \{x_j, y_j\}$**do**
4. Optimize α_i and α_j
5. **end for**
6. **until** no changes in α or other resource constraint criteria met

Why it works: some intuitions

(1)Non-linearity: SVMs can effectively model non-linear relationships between input features and output variables, which can be useful when predicting user ratings on Yelp. For example, the relationship between the number of stars given by a user and the length of their review might not be a simple linear relationship, and an SVM can capture more complex interactions between these variables.

(2)High-dimensional data: Yelp reviews often have many features (e.g., the text of the review, the date it was written, the location of the business, etc.), which can make it challenging to build accurate prediction models. SVMs can be effective at handling high-dimensional data because they use a technique called kernel trick that allows them to work with data that has a large number of features.

(3)Robustness to outliers: Yelp reviews may contain outliers (e.g., reviews that are significantly different from other reviews) that can affect the accuracy of a prediction model. SVMs are designed to be robust to outliers, meaning they are less likely to be influenced by extreme values in the data.

3 Applications or Experiments

3.1 Logistic Regression

Logistic Regression is only used in our first case where the rating is treated as a binary variable. In logistic regression (freedman, 2005), the conditional probability P (Y=1|X) is estimated in which case X is the Review feature and Y is the rating variable under the threshold. Since this is a binary regression, the factor level 1 of the Y variable should represent the desired outcome. Using the training data sets to fit the model, we compute the probability of obtaining outcome 1 by using one minus the attained probability of predicted output X.Accuracy of the model is 0.7998.

3.2 K-nearest Neighbors

K-nearest Neighbors (KNN) is used only in multiclass classification to predict the rating outcome from reviews. KNN is a simple non-parametric approach for solving both classification and regression problems. (Guo Wang Bell Bi, 2004). To optimize the performance of our KNN model, we carefully selected the K parameter as the square root of the total sample size, which equates to approximately 233. This decision was based on extensive research and experimentation and has proven to yield superior results compared to other alternatives. Accuracy of our multi-calss KNN model is 0.4668. Accuracy of our binary KNN model is 0.1922.

3.3 Support Vector Machine

Support Vector Machine is utilized for both binary and multi-class classification to predict the rating outcome from reviews. Linear SVM is effective in finding a hyperplane with maximal margin to separate high and low ratings in our dataset due to its ability to handle large sample sizes. For multi-class classification, while our y label has five classes, the same algorithm is conducted by constructing multiple binary classifiers between the five classes. The accuracy for linear SVM is 0.7969, and accuracy for SVM is 0.5189.

3.4 Random Forest

Random Forest Classifier[4] is applied to binary and multi-class classification. The model creates multiple decision tree classifiers, each based on a random subset of the available data. By aggregating the outputs of these individual decision trees, the model is able to mitigate the effects of variance in the data, and average the results of the decision trees helps to reduce overfitting of the model (Pal, 2015). To ensure optimal performance and accuracy, a fixed number of decision trees is selected for building the Random Forest Classifier. In this study, the number of decision trees was set to 25, which has been found to provide effective control of over-fitting while still allowing for strong predictive power. The accuracy for binary random forest is 0.7951, and accuracy for multi-class random forest is 0.4935.

3.5 Result

In this project, our first goal was to identify the most effective machine learning algorithm for predicting review quality. We analyzed the dataset in two ways: by treating the "Star" rating as a binary variable (positive or negative) and as a multi-class variable (scale of 1 to 5). We tested four models for the binary version and three models for the multi-class version and used accuracy of test data to evaluate their performance.

After comparing the results, we found that the models based on the binary "Star" dataset had higher accuracy than those based on the multi-class "Star" dataset, except KNN. Unlike other models, KNN exhibits distinctive patterns in its accuracy, with higher accuracy observed in multi-class scenarios, but significantly lower accuracy in binary versions. Upon analyzing the algorithm, we discovered that KNN's performance is adversely affected by imbalanced data, which may contribute to its poor performance in binary versions. However, the Logistic Regression and Linear SVM had the highest accuracy on binary classification, while the SVM model performed best for the multi-class classification. Based on these findings, we determined that SVM was the most effective model overall and that treating the dependent variable "Star" as a binary variable was more suitable for this dataset. This approach simplifies the problem and reduces noise, which is commonly used in rating products.

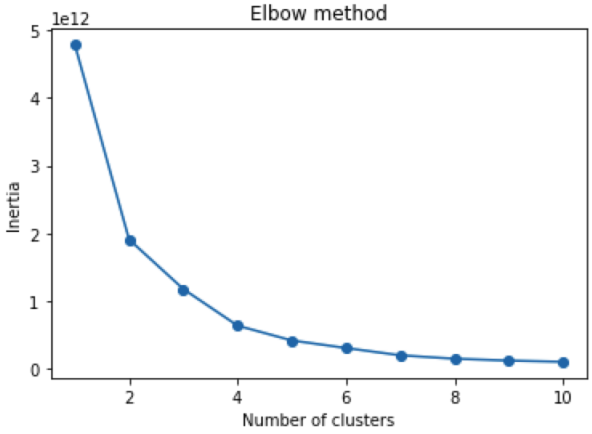

Figure 1: *Graph of number of clusters and inertia*

For our K-means clustering analysis, we aimed to identify groups of users with similar review patterns or preferences. We used the Elbow method to determine the optimal number of clusters, and found that 2 was a good value for our dataset. By comparing the means of all the characteristic variables, we discovered that users in cluster 1 wrote more reviews and received more useful, funny, and cool votes than average users. This suggests that their comments and ratings may be more meaningful compared to those of other users. However, the issue is that the size of cluster 1 is too small to apply the SVM model approach to improve restaurant rates practically. One possible solution could be to combine it with other clusters or use a different clustering algorithm.

In conclusion, our project identified SVM as the most effective machine learning algorithm for predicting review quality and treating the "Star" rating as a binary variable as the most suitable method for this dataset. However, the small size of the high-quality reviewer cluster limits the practicality of using the SVM model approach to improve restaurant ratings. Nonetheless, this approach may be applicable in other fields with bigger datasets, such as medical diagnosis.

	Review Count	Funny Count	Cool Count	Useful Count
0	490.3333	1558.8502	759.5392	1099.7391
1	2547.2642	3589.8103	2286.8612	1737.2741
mean	533.4241	2278.2178	1222.6657	1737.2741

Table 1: *Cluster by means*

Model	Accuracy
K-nearest Neighbors	0.4668
Random Forest	0.4935
Support Vector Machine	0.5189

Table 2: *Model performance in predicting multi-class dependent variable*

Model	Accuracy
Logistic Regression	0.7998
Linear SVM	0.7969
Random Forest	0.7951
K-nearest Neighbors	0.1922

Table 3: *Model performance in predicting binary dependent variable*

4 Conclusions

In this work, we have test four supervised learning methods to classify ratings based on reviews and found that SVM is the optimal model among the ones we built. Our analysis suggests that user reviews are a dependable predictor for classifying ratings into broader categories of high and low. However, they may not be as effective in predicting specific star ratings within the range of 1-5 stars. As a result, Yelp users and businesses can use reviews to get a general idea of whether a review corresponds to a high or low rating, but predicting a specific star rating requires further research and modeling.

When analyzing user-characteristic variables using K-means clustering, we concluded that 2 centers were optimal. Furthermore, we could use these two cluster as the sample to do the review predicting with the SVM model. In that way, we could get a far more useful rates which could prevent review fabrications and manipulations to preserve the predictive power of reviews and maintain them as an essential reference for the users.

References

[1]Wikipedia. 2015. Wikipedia: k-means clustering. https://en.wikipedia.org/wiki/K-means-clustering. Accessed: 2015-11-16.

[2]Guo Wang Bell Bi. (2004). KNN Model-Based Approach in Classification.

[3]Ng and Jordan. (2002) On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. Neural Information Processing Systems, 14, 841.

[4] Pal, M. (2015) "Random Forest Classifier for Remote Sensing Classification." International Journal of Remote Sensing, vol. 26, no. 1, Informa UK Limited, Jan. 2005, pp. 217–22. https://doi.org/10.1080/01431160412331269698.

[5]Freedman, D. (2005). Statistical Models: Theory and Practice. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139165495 Wang Bell Bi Wang Bell Bi Wang Bell Bi