Predicting Ratings in Hawaii's Google Local Reviews

Zheng Zeng and Runpeng Jian

University of California, San Diego

Abstract

This study focuses on the enhancement of recommendation systems and guiding user choices among massive amounts of information. As discussed in our class, traditional recommendation models often fail to accurately capture user preferences due to their reliance on conventional methodologies. To address this, we propose a novel multi-feature linear regression model. This model integrates user's rating behavior, business's general reputation and sentiment analysis into the predictive model. By combining these features, the proposed model aims to capture a comprehensive picture of user feedback, blending both quantitative and qualitative insights. Our approach allows for a more nuanced understanding of what drives user ratings, going beyond simple numerical analysis to include the subtleties of user sentiment and opinion. Code can be found here.

1 Introduction

Making decisions online is not always quick and simple, users are often inundated with an overwhelming array of information, including ratings, reviews, images, and more. While this wealth of data can be beneficial, it often complicates the decision-making process rather than streamlining it. Therefore, the primary challenge is to effectively utilize this information to assist users make satisfactory and informed decisions. This project seeks to address this challenge by exploring an advanced recommendation system for today's digital decision-making landscape.

The consumer frequently turns to platforms like Google Maps and Yelp to guide their choices, yet the process of sifting through these resources can be time-consuming. This project is motivated by the question of how to effectively utilize the extensive available online information to enhance user decision-making experiences. Our focus is on predicting the preferences of users regarding various businesses by analyzing the Google Local Review dataset, which encompasses users' past reviews and detailed business information. Our objective is to forecast the businesses that will resonate most with users by accurately predicting the ratings these users would assign to them. By doing so, we aim to substantially enhance the efficacy of the decision-making process.

Our approach, building upon models discussed in class, introduces an innovative method that uses a Gradient Boosting Regressor and XGBoost (Extreme Gradient Boosting) Regressor. This method uniquely integrates the results of sentiment analysis with user review features, marrying qualitative data with quantitative insights. This fusion is instrumental in augmenting the recommendation system's ability to mirror user preferences and opinions accurately.

The efficacy of our proposed model is demonstrated through its comparative performance against several established models. It shows a marked improvement over traditional models such as the baseline linear regression, latent factors model, standalone sentiment analysis, random forest model. This significant enhancement emphasizes the advantages of synergizing sentiment analysis with a Gradient Boost regressor.

This report is organized to offer a thorough insight into our project. It starts with a background overview and a review of related work, then delves

deeply into our approach, covering data prepossessing, the modeling process, and how we evaluate our models. Following this, we explore the outcomes and constraints of the model we've proposed, and conclude by suggesting directions for future exploration in this field.

2 Background

The evolution of recommendation systems is began with the innovative concepts of collaborative and content-based filtering. The early groundwork in collaborative filtering was laid by Goldberg et al.[3], who introduced the novel idea of using user-item interaction patterns to predict preferences. However, this method primarily focused on user behavior, overlooking the potential insights from the content or attributes of the items themselves.

Complementing collaborative filtering, Pazzani and Billsus explored content-based filtering[7]. This approach differs by recommending items similar to those a user has previously shown interest in, thereby creating a more personalized experience based on individual tastes and preferences.

A notable stride forward in this field was the development of the Latent Factor Model (LFM), a significant enhancement in collaborative filtering techniques[2]. LFM brought a new dimension to prediction accuracy by uncovering and utilizing the hidden patterns in user preferences and item characteristics.

Yet, like all technologies, collaborative filtering isn't without its challenges. A prominent issue is its scalability. As the number of users and items in the system grows, the computational complexity skyrockets. This poses a significant hurdle, especially in scenarios requiring swift, real-time recommendations.

Interestingly, the scope of recommendation systems isn't limited to user-item interactions alone. There's an emerging trend of incorporating diverse data sources to enrich the recommendation process. This includes leveraging social interactions (as noted by Guy in 2022)[4], demographic information about users, and contextual data. These methods are versatile and adaptable across various domains since they do not presuppose the nature of the items being recommended.[1]

In response to the complexities associated with traditional collaborative filtering, our study proposes a fresh approach. We aim to predict ratings using user review metadata and have chosen Linear Regression (LR) as our baseline model due to its lower computational demands and proven effectiveness in various settings. LR's simplicity in handling large datasets makes it an ideal point of comparison for more complex models.

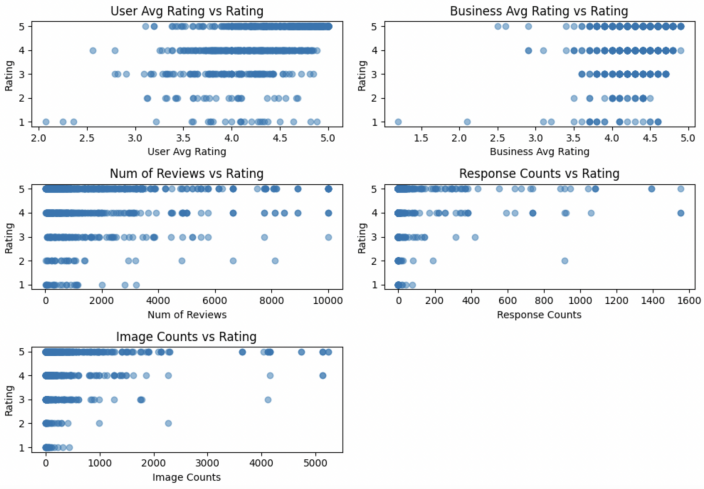
We then elevate our analysis by employing the Gradient Boosting Regressor (GBR) as our primary tool for prediction. GBR excels in modeling intricate, nonlinear relationships and thrives in high-dimensional data scenarios, making it a perfect fit for our objective of predicting user ratings from review metadata. This decision is supported by the work of He and Chua in their 2017 paper 'Neural Factorization Machines for Sparse Predictive Analytics'[5], which showcases the strength of gradient boosting in recommendation systems. By marrying the straightforward approach of Linear Regression with the sophisticated capabilities of Gradient Boosting Regressor, we aim to develop a nuanced and efficient model capable of extracting deep insights from user review metadata.

3 Methodology

For our study, we utilized two distinct datasets provided by Google Local Data: the Users' Reviews dataset and the Businesses' Metadata dataset in Hawaii. From the Users' Reviews dataset, we extracted a range of features that appeared to have potential utility then merged with corresponding data from the Businesses' Metadata dataset. To understand how each feature influenced user ratings, we employed data visualization techniques. Following this exploratory phase, we moved on to the core of our research - modeling the prediction of ratings and compare with each other. The subsequent sections of this paper will delve into the specifics of our methodology and findings.

3.1 Datasets Overview

In our research, we employed two key datasets: the Users' Reviews (10-core) dataset and the Businesses' Metadata dataset, both specific to Hawaii, as referenced in the works of Li (2022) [6] and Yan (2023) [8]. The Users' Reviews includes data from 64,336 users across 1,504,347 entries. Each entry in this dataset encompasses eight distinct fields: user_id, name, time, rating, text, pics, and resp (business's response to the review, including the time of response in Unix format and the response text), along with gmap_id. The Businesses' Metadata dataset contains 11,686 local businesses that cover 2,311 business categories. It comprises 21,507 entries, each detailed across 15 fields: name, address, gmap_id, description, latitude, longitude, category, avg_rating, num_of_reviews, price, hours of operation, MISC, the current state of the business (state), relative_results as related businesses recommended by Google, and url.



The scatter plots suggest that both users and businesses with higher average ratings are correlated with giving or receiving higher individual ratings and the image count may have a bit to do with it when the review counts exceed 300 or so. However, no clear trends are observed between the number of reviews and response counts and propensity to give higher or lower ratings.

3.2 Feature Engineering

In our analytical approach, we strategically selected features that we hypothesized would impact user ratings, drawing upon patterns observed in our data explorations and logical deductions. We sourced essential attributes from the Users' Reviews dataset, including user IDs, business identifiers (gmap_id), and the textual content of the reviews. Additionally, we quantified the number of images per business (image_counts) and the frequency of business responses to user reviews (response_count). We modify response_count(if have more than 1 response mark as 1, else 0) based on the scatter plot and our modified response_mark feature.

Turning to the Businesses' Metadata dataset, we further enriched our feature set. We incorporated the average business rating (bus_avg_rating) and, in place of the raw count of reviews, we implemented the 'rating_product', a derived metric combining the average rating and the number of reviews. Moreover, we calculated the average rating each user provided (user_avg_rating), thus fostering a holistic dataset that offers a multifaceted perspective for our subsequent analysis and predictive modeling endeavors.

In our original approach, we shuffled the entire dataset right after extracting it. However, this method resulted in a surprisingly high MSE. We hypothesized that this could be due to the distribution of features and labels becoming less representative of the underlying patterns in the data when shuffled too early in the process.

To address this, we modified our approach and decided to shuffle the dataset only after extracting all the relevant features. This alteration meant that the shuffling occurred once the dataset was fully formed, with all the intended features from both the Users' Reviews and Businesses' Metadata datasets in place. This change had a significant impact: we observed a noticeable decrease in the MSE. This result suggested that shuffling the data after feature extraction led to a more representative distribution of the data for training and testing purposes, thereby enhancing the model's ability to learn and generalize from the training data to unseen data.

3.3 Rating Prediction Modeling

Our task is to predict ratings given users and businesses information. In the following sections, we will talk about the models we explored respectively. Some of them were discussed in lecture, so we will explain more on models not demonstrated before.

3.3.1 Linear Regression (Baseline)

Linear Regression is known for its simplicity and low time complexity, making it highly suitable for large datasets. This efficiency is crucial in handling the extensive user review and business metadata we have compiled. Additionally, Linear Regression provides ease in integrating new features into the model. This adaptability is particularly valuable given our data's multidimensional nature, where we may need to experiment with different combinations of user and business attributes. Moreover, as a baseline model, Linear Regression offers a clear, interpretable benchmark.

3.4.1.1 Model_1

bus_avg_rating only

$\hat{y} = \beta_0 + \beta_1 \times \text{bus\_avg\_rating}$

Model_1 using only the business's average rating as a feature as a very baseline. It prepares the training and testing datasets by isolating the average business ratings and assigning them as the sole input feature, along with a constant term (intercept set to false). it turns out the test rmse is 0.88596.

3.4.1.2 Model_2

user_avg_rating only

$\hat{y} = \beta_0 + \beta_1 \times \text{user\_avg\_rating}$

Model_2 using only the business's average rating as a feature and test rmse significantly improved to 0.78773.

3.4.1.2 Model_3

user_avg_rating + bus_avg_rating

$\hat{y} = \beta_0 + \beta_1 \times \text{user\_avg\_rating} + \beta_2 \times \text{bus\_avg\_rating}$

Model_3 using only the business's average rating and user's average rating as the feature and test rmse improved to 0.76021.

3.4.1.2 Model_4

all features

$\hat{y} = \beta_0 + \beta_1 \times \text{user\_avg\_rating} + \beta_2 \times \text{bus\_avg\_rating} + \beta_3 \times \text{num\_of\_reviews} + \beta_4 \times \text{response\_mark} + \beta_5 \times \text{image\_counts}$

Model_4 includes all the features and test rmse slightly improved to 0.76020. This suggests that the newly added feature may not exhibit a strong linear regression relationship with the target variable, we might consider other models.

3.3.2 Bag of Words (BOW)

The Bag of Words (BoW) model is utilized to translate textual data from user reviews into numerical vectors, which act as input features for a machine learning model. Initially, the text from each review is tokenized, splitting it into individual words. The BoW model then processes these tokens, forming vectors where the frequency of each word is treated as a distinct feature. These vectors are then inputted into a Ridge Regression model. The Ridge Regression technique, characterized by its ability to apply regularization, is specifically chosen for its effectiveness in handling overfitting and managing the complexities associated with large feature sets typical of BoW models. This method provides a solid basis for our recommendation system, leveraging the simplicity of BoW for data transformation and the robustness of Ridge Regression for prediction. The synergy of these two techniques facilitates a more nuanced and accurate prediction of user ratings, drawing on the strengths of both methods. The effectiveness of this method is demonstrated by the RMSE metrics obtained from our model evaluations. The training set yielded an RMSE of 0.8335, while the testing set showed an RMSE of 0.8342.

3.3.3 Latent Factor Model (LFM)

To uncover implicit relationships between users and business, we utilize a Latent Factor Model (LFM) to enhance the recommendation system. LFM works by discovering latent features underlying the interactions between users and items. Initially, user-item interactions in our study, i.e. user ratings are encoded into a matrix. The LFM then decomposes this matrix into lower-dimensional user and item matri-

ces, revealing latent factors that represent underlying characteristics of items and preferences of users. These factors are learned through iterative optimization techniques, with the goal of minimizing the difference between predicted and actual interactions. The resulting model is capable of predicting user preferences for items, even those not previously interacted with, by leveraging these latent factors. This approach is particularly powerful for capturing complex patterns and relationships in the data, leading to more accurate and personalized recommendations in our system. The effectiveness of this method is demonstrated by the RMSE metrics obtained from our model evaluations. The testing set showed an RMSE of 0.82194 without tuning.

3.3.4 Gradient Boosting Regressor

This method builds trees one at a time, where each new tree helps to correct errors made by the previously trained tree. This sequential tree building technique can lead to deeper insights since each tree is learning from the mistakes of its predecessors. It is highly effective for datasets with complex relationships and is known for its high accuracy. When all features were incorporated into the Gradient Boosting Regressor, there was a notable improvement in the model's performance. Specifically, the RMSE for the test set decreased to 0.754119 without tuning, reflecting a more accurate prediction capability.

3.3.5 XGBoosting Regressor

XGBoost(Extreme Gradient Boosting) stands out for its efficiency and performance in handling large and complex datasets. It is a scalable and optimized version of gradient boosting, known for its speed and accuracy. XGBoost employs an advanced regularization (L1 and L2), which improves model generalization capabilities and helps to reduce overfitting, making it superior to traditional gradient boosting. It also offers flexibility to define custom optimization objectives and evaluation criteria, adding to its versatility. XGBoost is often the go-to choice for many machine learning competitions due to its robust performance across a wide range of datasets. The testing set showed an RMSE of 0.77095 without tuning.

3.3.6 Sentiment-Guided XGBoost and Gradient Boosting Regressor

Based on the outcomes of our RMSE result for now, it has become evident that XGBoost and Gradient Boosting Regressor emerge as the top-performing models for our dataset. In light of their impressive performance, we have decided to further enhance our feature set by incorporating the Bag of Words (BoW) methodology. By adding BoW as a feature, we aim to leverage its textual analysis strengths, enriching our models with more detailed and representative data inputs. For comparison, we also added TextBlob and VADER(two popular tools used in NLP for sentiment analysis). TextBlob's sentiment analysis is more straightforward and works well for general texts. In contrast, VADER is particularly powerful for texts with informal language, such as social media content, due to its sensitivity to modern internet language and its use of a lexicon that includes slang and emoticon. result shows in evaluation.

4 Evaluation

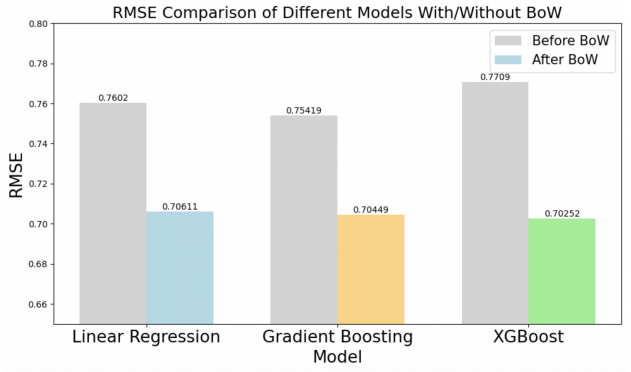


Figure 1: Adding BoW without tuning

The data presented in Figure 1 clearly illustrates a significant improvement in model performance upon incorporating Bag of Words (BoW) sentiment analysis into our feature set.

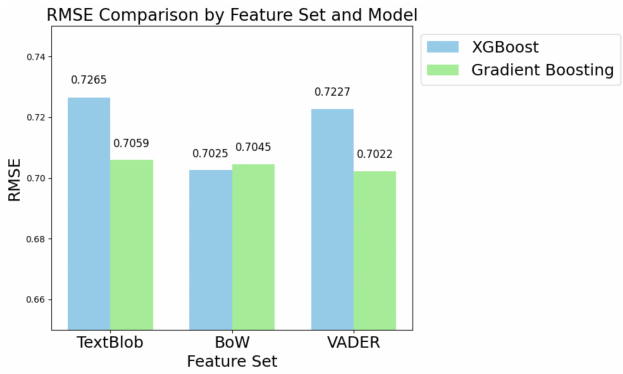


Figure 2: Different Language Tools Before Tuning

The comparison of three different language processing tools aimed to determine if any particular tool would significantly enhance our model's performance. However, the results indicated only marginal improvements. Given this outcome, we have opted to simultaneously fine-tune all three tools. This approach allows us to leverage the unique strengths of each tool, potentially leading to a more substantial cumulative improvement in our model's accuracy and effectiveness.

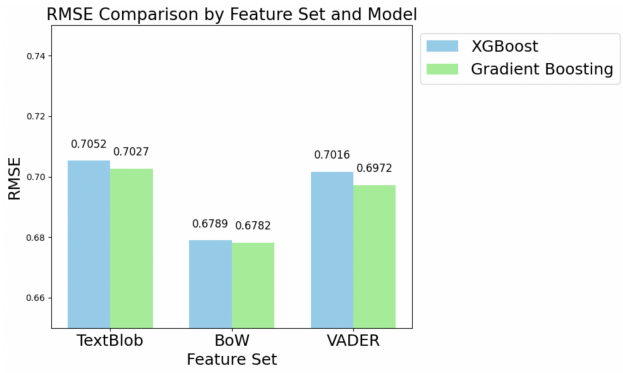


Figure 3: different language tools after tuning

Following the fine-tuning process with the three language processing tools, it became evident that the Bag of Words (BoW) approach outperformed the others, achieving a notably lower RMSE of 0.6782. This result underscores the effectiveness of BoW in enhancing the predictive accuracy of our model compared to the alternative methods tested.

User ID	Predicted Rating (GradientBoost)	Actual Rating
35249	3.83	5
72627	4.78	5
125988	4.42	5
56147	1.26	1
32345	4.96	5
100570	4.51	5
97701	4.63	5
142778	4.53	4
122301	5.00	5
14162	3.57	1

Figure 4: Rating prediction of 10 random users

Ultimately, I chose 10 users at random from the test set to predict their ratings and compared these predictions with their actual ratings

5 Conclusion

In our project, the Gradient Boosting Regressor distinguished itself by delivering outstanding performance and managing complex datasets. The strength of this model lies in its iterative refinement process, where it progressively corrects the errors from preceding steps, thereby enhancing its predictive accuracy. Furthermore, sentiment analysis played a more significant role in predictions than initially anticipated. Despite its simplicity, the Bag of Words (BoW) approach yielded impressive results in analyzing specific datasets, such as the one we worked with. Following the development of this predictive model, we can rank merchant ratings from highest to lowest. When users choose a merchant category, we can offer recommendations based on these ratings, forming the foundation of a basic recommendation system.

6 Limitation

Limitations we encountered include data skewness favoring high ratings and the inherent constraints of the Bag of Words (BoW) methodology. The disproportionate representation of high ratings in the dataset

leads to a systemic bias, impairing the model's precision in predicting lower ratings. Additionally, the BoW approach, while effective in basic text analysis, proves inadequate in discerning the complexities of ironic or sarcastic reviews, limiting its interpretative accuracy. Addressing these challenges is imperative to enhance the robustness and accuracy of the recommendation system, ensuring equitable and comprehensive predictive performance across diverse user feedback.

7 Future Work

In future studies, addressing the limitations of the current model will involve two primary strategies. First, rectifying the data skewness issue by enriching the dataset with a greater proportion of lower ratings, potentially through enhanced data collection or synthetic data generation. This adjustment aims to balance the dataset, improving the model's proficiency in predicting lower ratings. Second, incorporating advanced Natural Language Processing (NLP) techniques, such as context-aware algorithms like BERT or GPT, will be pivotal. These sophisticated models, known for their effectiveness in understanding nuanced language, including irony and sarcasm, could substantially refine the system's text analysis capabilities. Additionally, exploring hybrid approaches that combine machine learning with advanced NLP may offer a comprehensive solution, enhancing both the predictive accuracy and linguistic interpretability of the recommendation system. This future direction aims to bolster the system's overall robustness and reliability, ensuring more accurate and user-centric recommendations.

References

[1] Pablo Castells and Dietmar Jannach. Recommender systems: A primer, 2023.

[2] Jiansheng Fang, Xiaoqing Zhang, Yan Hu, Yanwu Xu, Ming Yang, and Jiang Liu. Probabilistic latent factor model for collaborative filtering with bayesian inference. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, January 2021.

[3] David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12), 1992.

[4] Ido Guy. *Social Recommender Systems*, pages 835–870. Springer US, United States, January 2022. Publisher Copyright: © Springer Science+Business Media, LLC, part of Springer Nature 2011, 2015, 2022.

[5] Xiangnan He and Tat-Seng Chua. Neural factorization machines for sparse predictive analytics, 2017.

[6] Jiacheng Li, Jingbo Shang, and Julian McAuley. Uctopic: Unsupervised contrastive learning for phrase representations and topic mining, 2022.

[7] Michael J Pazzani and Daniel Billsus. Content-based recommendation systems. In *The Adaptive Web*, pages 325–341. Springer Berlin Heidelberg, 2007.

[8] An Yan, Zhankui He, Jiacheng Li, Tianyang Zhang, and Julian McAuley. Personalized showcases: Generating multi-modal explanations for recommendations, 2023.