PREPARED BY TEAM TRACE3B

- Vy Nguyen
- Kento Morita
- Qui Nguyen
- Aria Zhou
- Sahil Chennadi
- Yuheshwar Kamakkapalayam Subramani

Customer & Social Analytics Project

ANALYZING RECIPE SUCCESS

A Deep Dive into User Ratings

March 17, 2025

Table of Contents

1. Introduction

1.1 Project Background

Recipe-sharing platforms have increasingly become an important tool for home cooks and food enthusiasts worldwide. These platforms not only provide access to a vast array of recipes but also allow users to rate and review their experiences. Therefore, understanding what drives user satisfaction and high ratings is crucial for recipe creators and platforms who aim to optimize their offerings. This project delves into the factors that influence recipe success by deploying machine learning models in two main areas: recipe attributes and user sentiment. By answering two questions, "What recipe attributes influence ratings?" and "How do user sentiments impact ratings?, the analysis provides actionable insights for recipe creators like food bloggers, meal kit services, and recipe-sharing websites to optimize their design for higher user satisfaction.

1.2 Executive Summary

The project examines recipe data from a popular recipe-sharing platform to identify the key factors that drive user ratings. Using two main datasets, recipe attributes and user sentiments, a comprehensive analysis was conducted that included data cleaning, exploratory data analysis, feature engineering, and machine learning modeling. The models revealed that taste, ease of preparation, and user experience are the most significant drivers of high ratings, outweighing factors like recipe complexity or nutritional content. Based on these findings, actionable recommendations for recipe creators centered on recipe ingredients, steps, and tags.

2. Data Description

2.1 Overview of Data

The analysis is based on two primary datasets sourced from Kaggle:

- **RAW_recipes.csv:** This dataset contains 231,637 rows and 12 columns detailing various recipe attributes in addition to unique identifiers like recipe ID and contributor ID, as summarized in the table below:

Attribute	Description
name	The name of the recipe.
minutes	Minutes required to prepare the recipe.
tags	Food.com tags associated with the recipe (e.g. 'course')
nutrition	Nutritional information, including calories (#), total fat (PDV), sugar (PDV), sodium (PDV), protein (PDV), saturated fat
n_steps	Number of steps required to prepare the recipe.
steps	The text description of each step in the recipe.
description	User-provided description for the recipe.
ingredients	A list of ingredients required for the recipe.
n_ingredients	The total number of ingredients in the recipe.

- **RAW_interactions.csv:** This dataset contains 1,158,039 rows and 5 columns, capturing user interactions with recipes besides recipe ID and user ID, including:
    - **Rating**: User ratings on a scale of 0 to 5.
    - **Review**: User-provided feedback in text form.

2.2 Exploratory Data Analysis

Running Exploratory Data Analysis on the two datasets, we discovered that the data is is heavily skewed towards higher ratings (4 and 5 stars). This indicates a strong positive bias, suggesting that users are more likely to rate recipes highly.

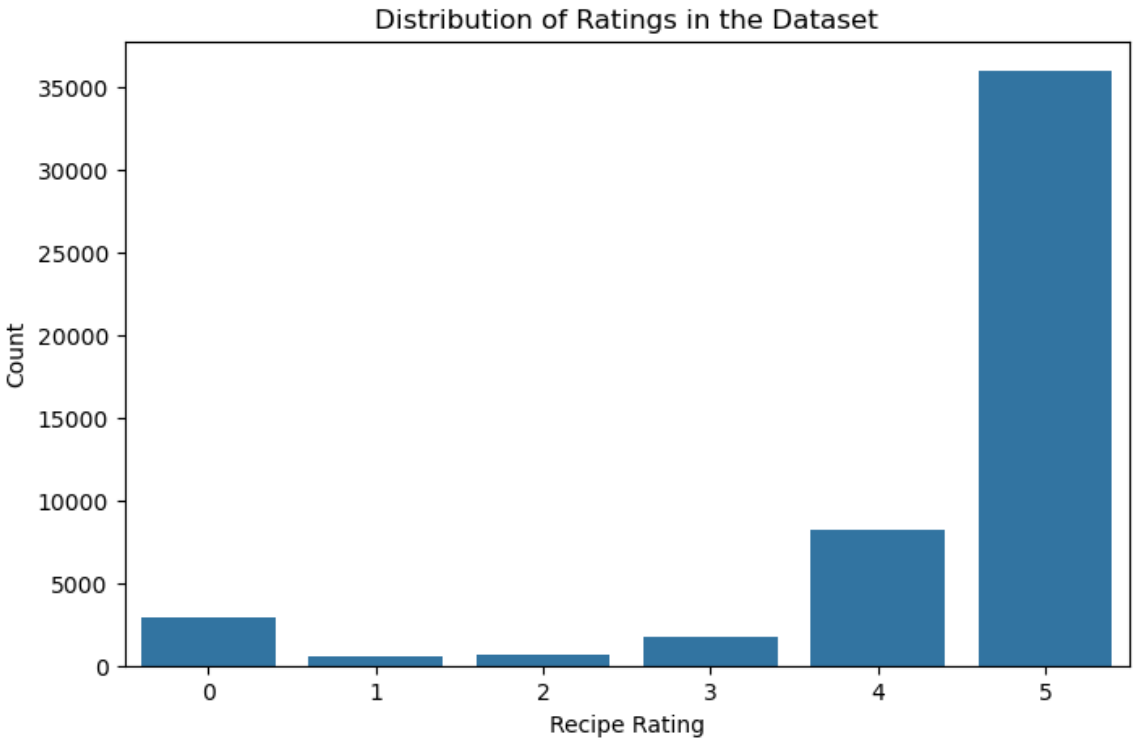

Figure: Distribution of Ratings in the RAW_recipes.csv dataset.

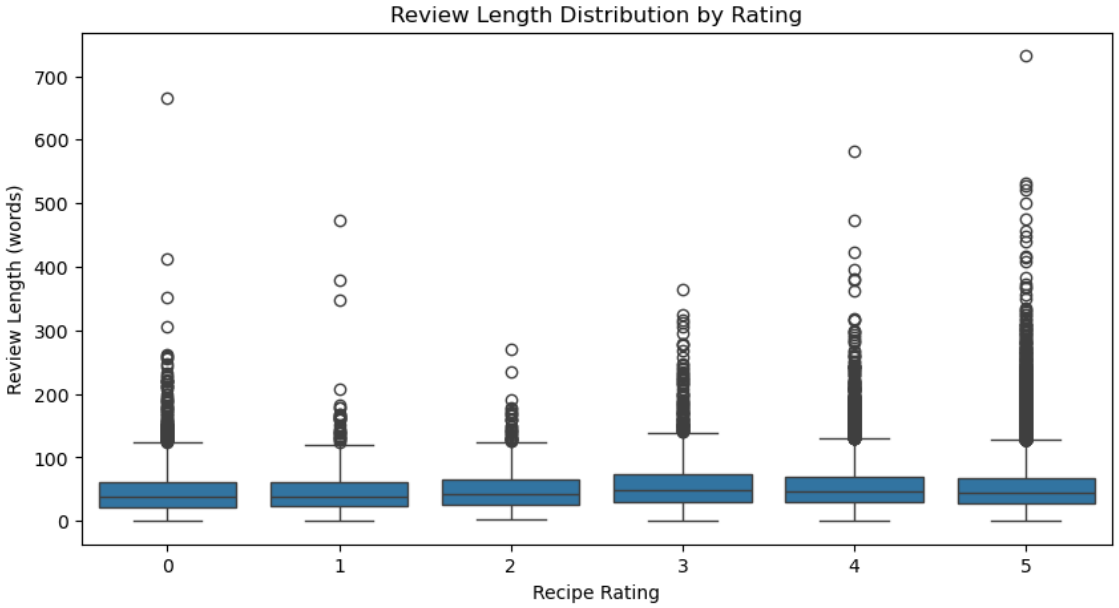

Figure: Review Length Distribution by Ratings in the RAW_recipes.csv dataset.

Further observation of review length showed that longer reviews were often associated with extreme ratings, either highly positive (5-star) or highly negative (1-star). This suggests that users with strong opinions tend to provide more detailed feedback, while shorter reviews were more common for neutral ratings (3-star). Finally, a correlation analysis between recipe attributes and ratings revealed weak linear relationships, with most attributes showing near-zero correlation with the response variable.

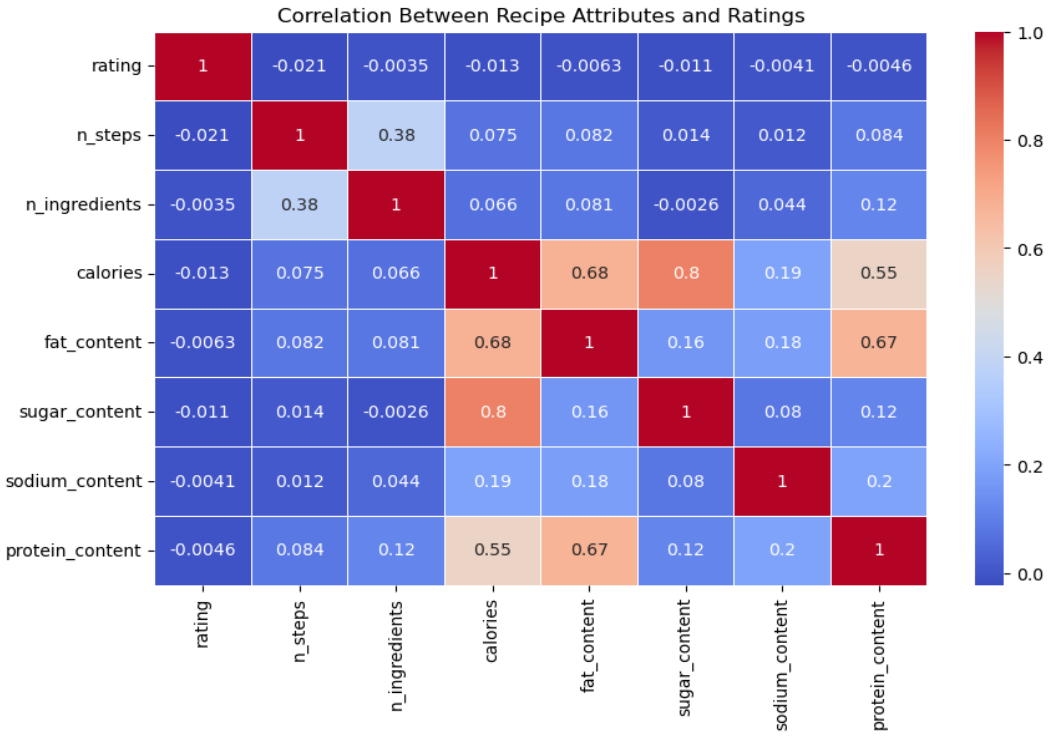

Figure: Correlation Matrix between Rating and its Attributes.

Figure: Word Clouds of Highly-Rated Recipes and Reviews.

The two word cloud above reveals a preference for moderate ingredient counts (7-9 ingredients) and manageable steps (5-8 steps) for highly-rated recipes, indicating users favor simplicity and flavor. Terms like "high-calorie" and "low-calorie" suggest dietary preferences also play a role. Besides, they highlights positive descriptors like "delicious," "easy," and "flavorful" in high-rated reviews, emphasizing the importance of taste and ease of preparation. Meanwhile, lower-rated reviews often lacked descriptive language, correlating with dissatisfaction.

3. Recipe Attribute Analysis

3.1 Data Preprocessing

The primary dataset used to train the machine learning models is RAW_recipes.csv. However, we extracted the rating variable from the RAW_interations.csv to the RAW_recipes.csv dataset. The merging process involved aggregating the average rating for each recipe ID, making it the dependent variable in our models.

Additionally we modified three key attributes to enhance our feature set, including tags, nutrition, and submitted. The tags attribute represents a collection of categorical labels associated with each recipe. Since the number of tags could indicate the level of categorization and, potentially, the popularity of a recipe, we extracted the count of tags for each recipe and created a new numerical feature named n_tags. The nutrition attribute consists of a list of nutritional values, such as calories, fat content, and protein. We computed the total sum of these values to generate a feature labeled nutrition_sum, which reflects the overall nutritional richness of the recipe. Finally, the submitted attribute contains the date on which a recipe was uploaded. To analyze potential temporal trends, we calculated the total number of years from submission and stored it as total_year_from_submission, transforming it into a numerical feature.

Since the dataset was sourced from Kaggle, it contains no missing values. Thus, no imputation was necessary, and all records were utilized as provided. The dataset was then divided into training and testing subsets to facilitate model evaluation. We ensured a stratified split to maintain balanced distributions of ratings across both sets, ensuring that the training process was not biased by an uneven distribution of ratings. Additionally, since the rating

distribution was heavily skewed toward 4 and 5, we performed oversampling by SMOTE to balance the dataset and improve model performance.

3.2 Modeling Methods

After preparing the dataset, we implemented multiple machine learning models to predict recipe ratings. The first model used was a standard Linear Regression model, which served as a baseline by modeling the relationship between recipe attributes and ratings in a linear fashion. However, since ratings are distributed between 0 and 5, we suspected that Linear Regression might not be well-suited for this problem. Recognizing these limitations, we also applied a Generalized Linear Regression approach, which allows for modeling distributions beyond the standard normal distribution.

To improve predictive accuracy, we experimented with a Neural Network as the third model, leveraging its multi-layered architecture to capture intricate relationships in the data. Following this, we implemented Random Forest as the fourth model as it is particularly effective in handling non-linearity and feature interactions. Feature selection was conducted before training to determine the most relevant predictors for each model, ensuring an optimal balance between performance and computational efficiency. The used independent attributes are minutes, n_steps, n_ingredients, n_tags, nutrition_sum, and total_year_from_submission, whereas the dependent variable is rating.

3.3 Modeling Results and Performance

To assess model performance, we employed 10-fold cross-validation to determine optimal parameters and evaluated predictions using Mean Squared Error (MSE) as the primary comparison metric between models, as illustrated in the table below:

Model	Average MSE
Linear Regression	127.94
Generalized Linear Model	0.98
Neural Network (3 layers)	0.98
Random Forest	0.15

The results showed that Linear Regression performed the worst, with an average MSE of 127.94, indicating that it struggled to capture the underlying patterns in the data. Meanwhile, the Generalized Linear Regression and Neural Network models improved considerably, achieving an MSE of 0.98. The Random Forest model performed the best, achieving the lowest MSE of 0.145, indicating a strong ability to generalize across the dataset. Despite the promising performance of the Neural Network, its increased computational complexity and training time were notable trade-offs compared to the Random Forest model, which provided superior accuracy with lower computational costs.

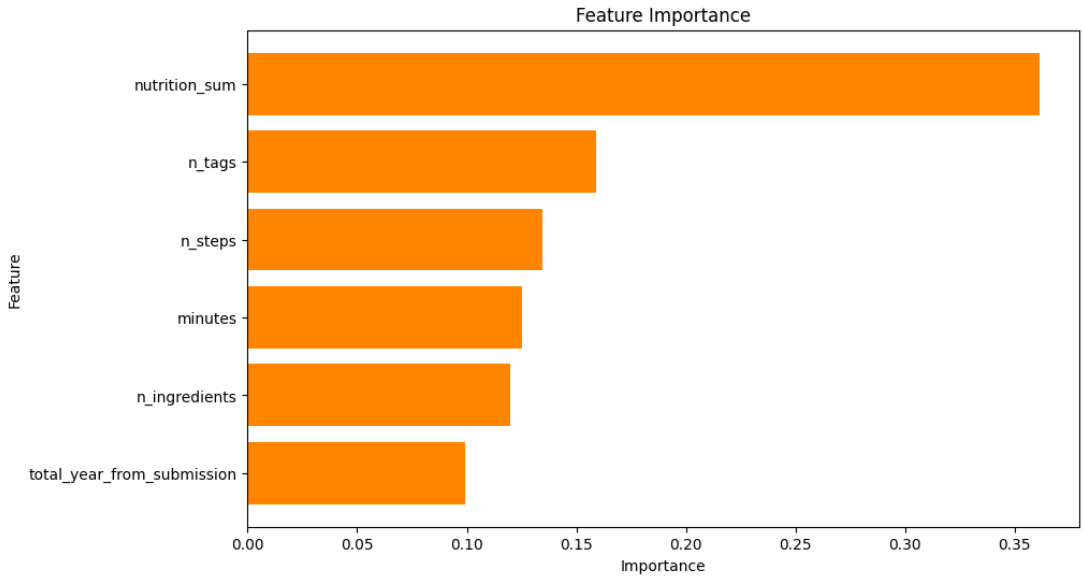

Figure: Feature Importance from Random Forest Model.

From the Random Forest model, we conducted a feature importance analysis to understand which attributes had the most significant impact on recipe ratings. The analysis revealed that *nutrition_sum, n_tags, and n_steps* were among the top contributors to the model's predictions. These findings indicate that nutritional content, the number of associated tags, and the year of submission play crucial roles in determining a recipe's rating.

3.4 Key Insights

Several key insights emerged from our analysis. Firstly, the number of tags associated with a recipe was positively correlated with its rating. Recipes with a greater number of tags tended to receive higher ratings, suggesting that well-categorized recipes are more appealing to users.

Secondly, nutritional attributes also played a role in rating predictions. While recipes with balanced nutritional values received favorable ratings, those with extreme values, either too high or too low, were generally rated lower.

Besides, temporal trends also revealed interesting patterns. More recent recipes exhibited higher average ratings, possibly due to changing user preferences or improvements in recipe quality over time. Finally, the feature importance analysis above confirmed that attributes such as nutrition_sum, n_tags, and submitted_year were among the most influential predictors of recipe ratings.

4. Review Sentiment Analysis

4.1 Data Preprocessing

The RAW_interactions.csv dataset was utilized for this analysis, with several preprocessing steps applied to prepare the data. Similar to recipe attribute analysis, feature engineering was performed to derive new features, including:

- Review Length: Number of words per review.
- Word Count: Total number of characters in the review.
- Sentiment Score: A heuristic score based on the occurrence of positive and negative words.
- TF-IDF Vectorization: Text reviews were converted into numerical feature representations.
- Nutritional Features: The nutrition column was parsed to extract values such as calories, fat, sugar, sodium, and protein content.
- Rating Categorization: A Binary classification target was created where ratings ≥4 were labeled as high-rated (1) and <4 as low-rated (0).

Furthermore, as we incorporated NLP models to analyze user reviews, we preprocessed the text review by:

- Tokenization and stopword removal were applied to clean the text data.
- Lemmatization was performed to standardize word forms.

- All reviews were lowercased to maintain consistency.

Additionally, missing review text was replaced with empty strings to prevent data loss. Meanwhile, outliers in numerical columns, such as calories and fat content, were identified using interquartile range (IQR) analysis and removed if they fell beyond acceptable thresholds.. Finally, before splitting into 80% training and 20% testing sets, we addressed the imbalanced distribution of ratings (with most reviews rated 4 or 5 stars) by applying SMOTE oversampling method again to generate samples for underrepresented classes (1-star and 2-star ratings).

4.2 Modeling Methods

Similar to previous analysis, feature selection was conducted using mutual information and feature importance scores. After that, another Random Forest Model was trained to predict rating but with a new sets of selected features below:

- Engineered text features (TF-IDF, sentiment score)
- Metadata attributes (number of steps, number of ingredients)
- Nutritional attributes (calories, fat, sugar, sodium, protein).

4.3 Modeling Results and Performance

The performance report for the Random Forest model is summarized in the following table. The model achieved a high accuracy and strong performance across all rating categories. It achieved high precision and recall for 5-star ratings, correctly identifying positive reviews with minimal false positives.

Metric	Score
Precision	0.96
Recall	0.90
F1-score	0.93
Metric	Score

Furthermore, our sentiment analysis revealed a strong correlation between user ratings and text sentiment scores. Reviews with higher ratings (4-5 stars) frequently included positive words like "delicious," "easy," "favorite," "amazing," and "perfect," while lower-rated reviews (1-2 stars) often contained negative terms such as "bland," "dry," "too salty," "tasteless," and "disappointed." Additionally, highly rated reviews exhibited greater polarity, reflecting a stronger positive sentiment, whereas lower-rated reviews were more subjective and opinion-based.

4.4 Key Insights

The model and sentiment analysis above revealed that review length significantly impacts ratings, with longer reviews often associated with more extreme ratings (either 1-star or 5-star). Specifically, users who expressed strong opinions, whether highly positive or highly negative, tended to write lengthier reviews, while shorter reviews were more neutral, typically corresponding to 3-star ratings. Additionally, reviews with rich descriptive language (e.g., adjectives and adverbs) were positively correlated with 5-star ratings, whereas minimal or vague descriptions were more common in lower-rated reviews.

5. Recommendations

Based on the two analyses, the following recommendations can help recipe creators and platforms improve user satisfaction and engagement, ultimately driving higher ratings and success:

5.1 Focus on Taste and Ease of Preparations

First, recipe creators should prioritize taste and ease of preparation in their recipes. The attribute analysis reveals that recipes described with positive language such as "delicious," "easy," and "flavorful" are more likely to receive higher ratings. Additionally, simplifying recipes to include 7-9 ingredients and 5-8 steps makes them more accessible to a broader audience, including beginners and time-constrained users. By focusing on these factors, creators can ensure that their recipes are not only enjoyable but also practical for everyday cooking.

5.2 Leverage Nutritional Appeal

Second, nutritional content plays a significant role in user preferences, particularly for health-conscious audiences. Previous findings indicate that highlighting moderate caloric and fat content can positively influence ratings. However, overemphasizing sodium or sugar levels does not significantly impact user satisfaction. Therefore, recipe creators should focus on promoting balanced nutrition in their recipes while avoiding excessive emphasis on less impactful nutritional factors.

5.3 Encourage Detailed User Reviews

Third, the analysis previously shows that longer reviews often indicate stronger opinions, whether very positive or very negative. Platforms should actively encourage users to leave detailed feedback by prompting them with specific questions about their experience. Additionally, leveraging sentiment analysis on reviews can help identify common pain points in low-rated recipes, enabling creators to address these issues and improve future offerings.

5.4 Encourage Detailed User Reviews

Lastly, the way a recipe is presented can significantly impact user experience. Effective use of tags (e.g., "low-carb," "30-minute meals") helps users quickly find recipes that match their preferences, increasing the likelihood of positive experiences. Furthermore, ensuring that recipe steps are clear and easy to follow is critical for user satisfaction. Creators should prioritize clarity and simplicity in their instructions, avoiding overly complex or ambiguous steps.

6. Conclusion

In summary, this project underscores the importance of data-driven decision-making in the culinary world. By applying the aforementioned insights, recipe creators and platforms can better meet user expectations, ensuring higher satisfaction and long-term success. Nevertheless, future work could expand on these findings by incorporating user demographics, integrating advanced NLP techniques for deeper sentiment analysis, and conducting A/B testing to validate the impact of different recipe formats. These steps would further refine the understanding of user preferences and enhance the effectiveness of recipe optimization strategies over time.