### Abstract

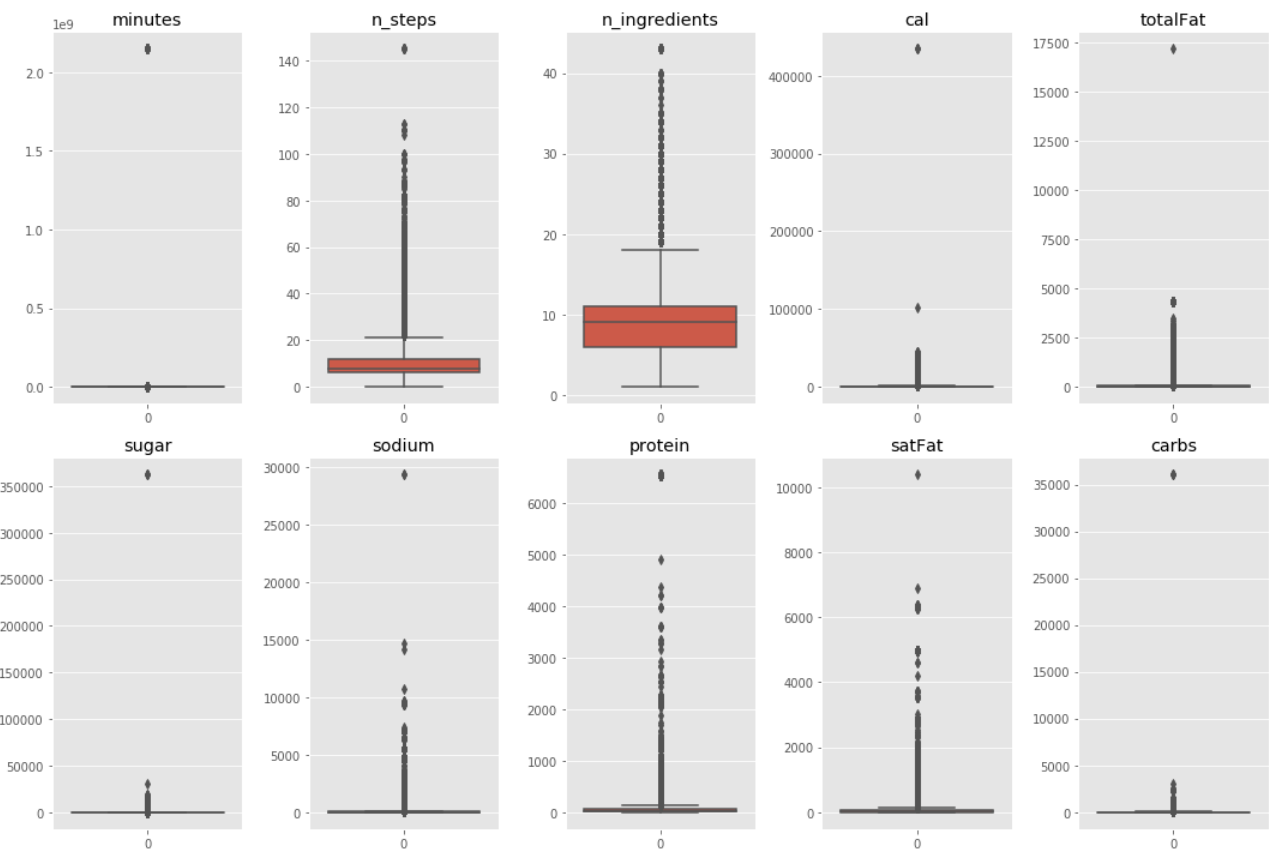
### Our Work focuses on recommending recipes to users using the "Food.com Recipes and User Interactions" dataset. Through exploratory analysis, we identified key features influencing recipe preferences. Our approach included preprocessing the data for enhanced feature integration and designing a modified Bayesian Personalized Ranking (BPR) model. Incorporating elements like nutritional content, tags, and cooking time into the BPR framework, we achieved an accuracy of around 77%. This model demonstrates improved performance in personalized recipe recommendations by leveraging a more comprehensive dataset analysis.

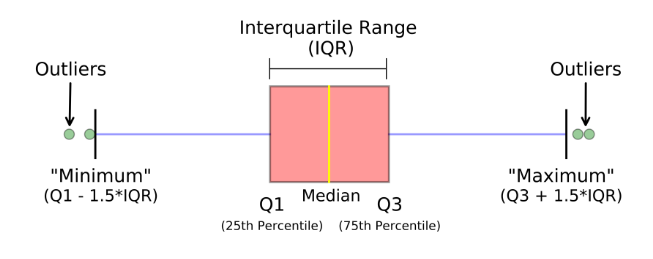
### Exploratory analysis of the data

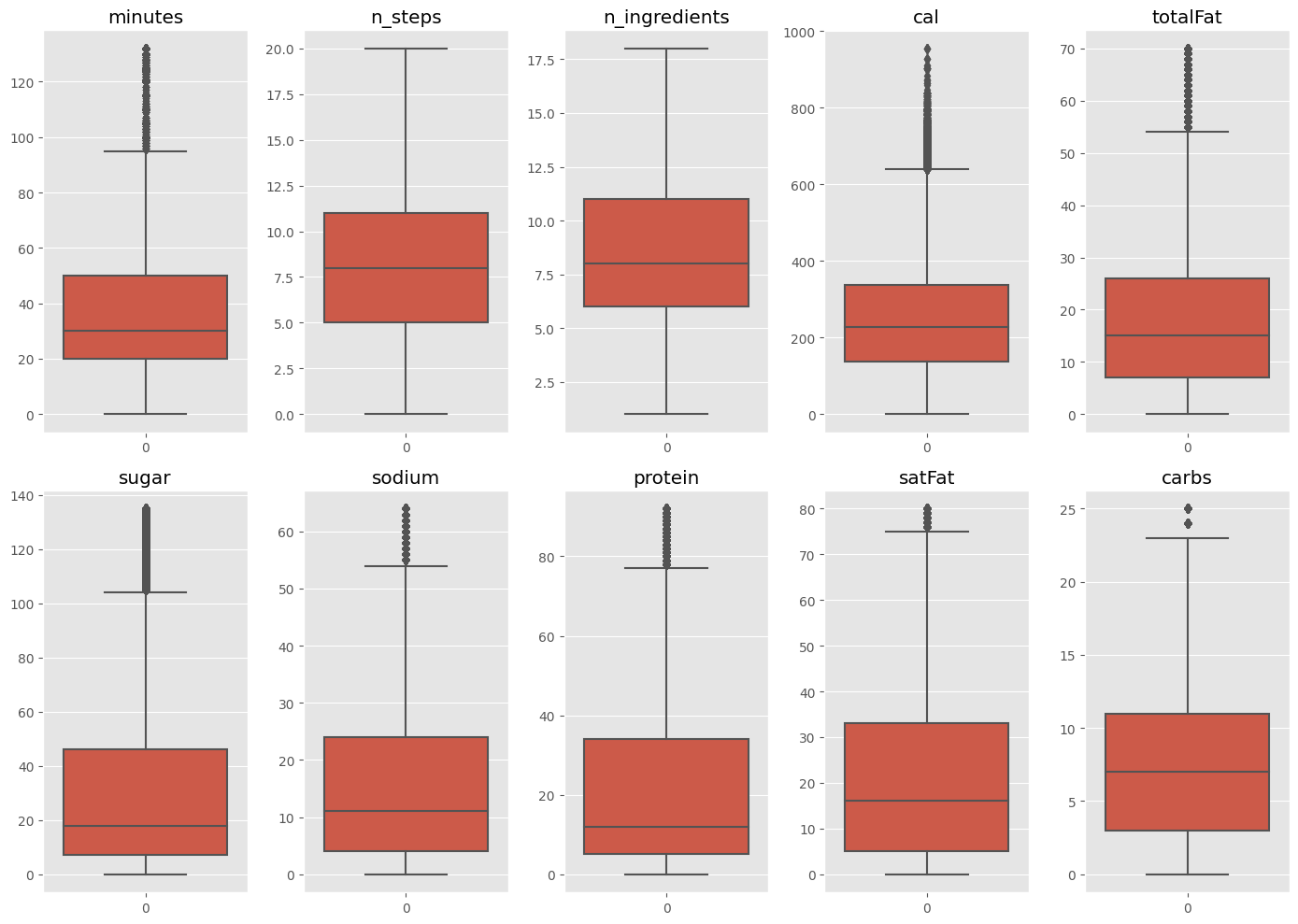
**Food.com**, a digital gastronomic repository and social platform, with over **180,000 recipes** and **700,000+ reviews,** covering **18 years of user interactions** and uploads on Food.com. It functions as an expansive database of recipes, accompanied by a wealth of user-generated content in the form of reviews and interactions.

(I) **Box plots :** The box plots( before & after removing outliers) provide a visualization of the distribution of various numerical attributes of a dataset concerning recipes and user interactions. These attributes include preparation time (minutes), complexity (n\_steps), diversity (n\_ingridients), and nutritional content (cal, totalFat, sugar, sodium, protein, satFat, carbs).

The **Lower bound** was set to *Q*1−1.5×*IQR* and the **Upper bound** was determined as *Q*3+1.5×*IQR*, where *Q*1 and *Q*3 represent the first and third quartiles, respectively, and *IQR* is the interquartile range.

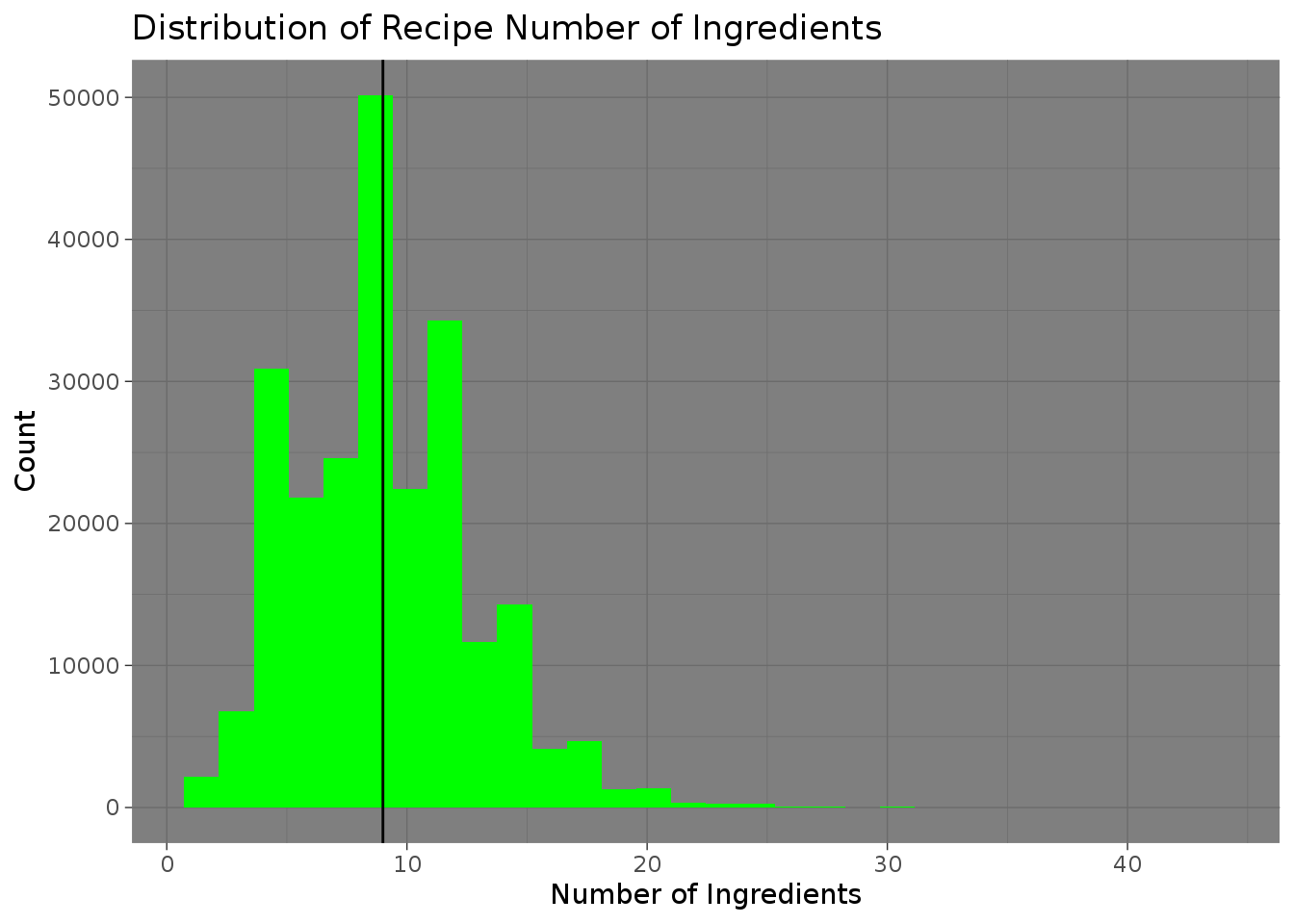




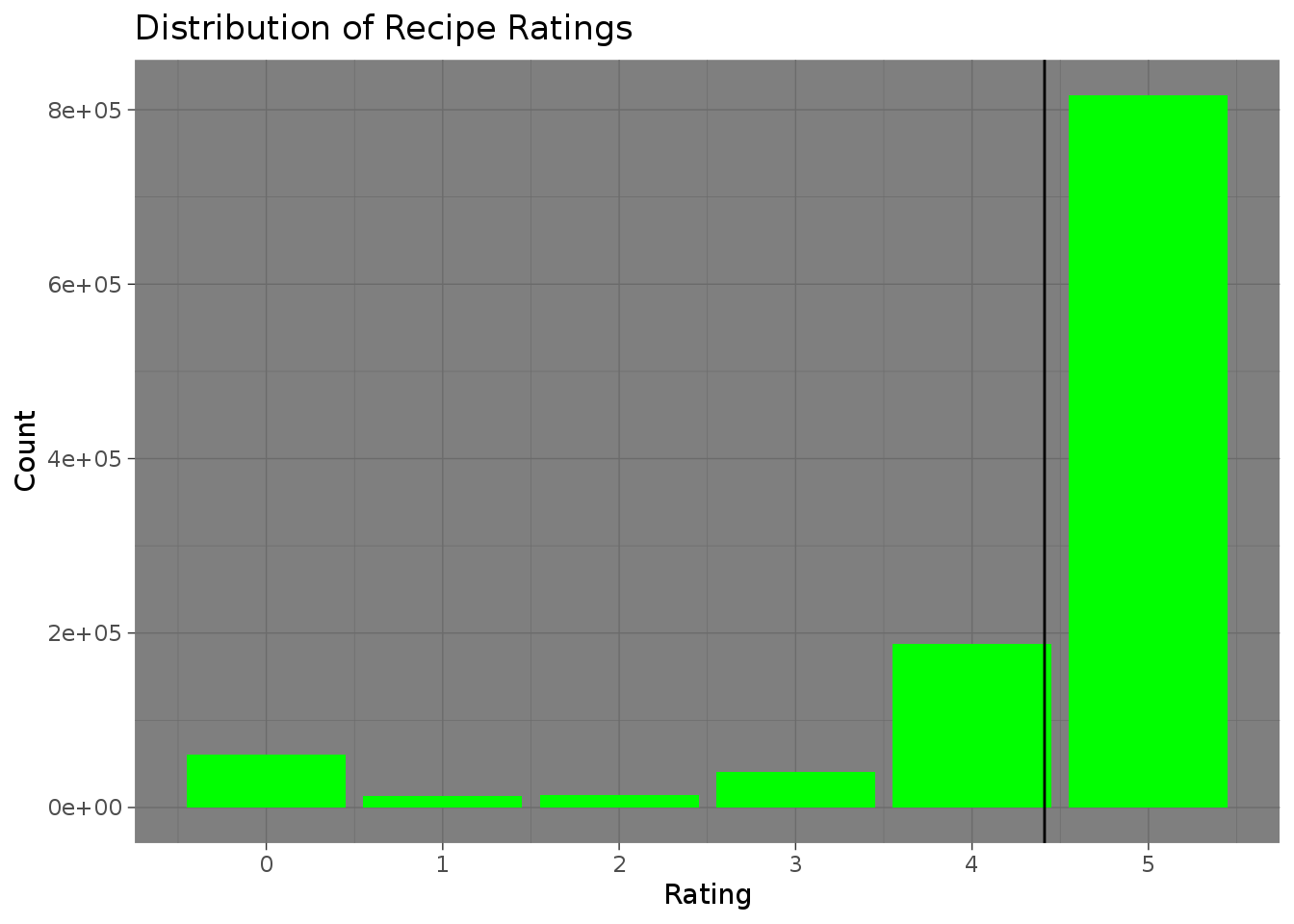


(II) **Dataset Distributions:**

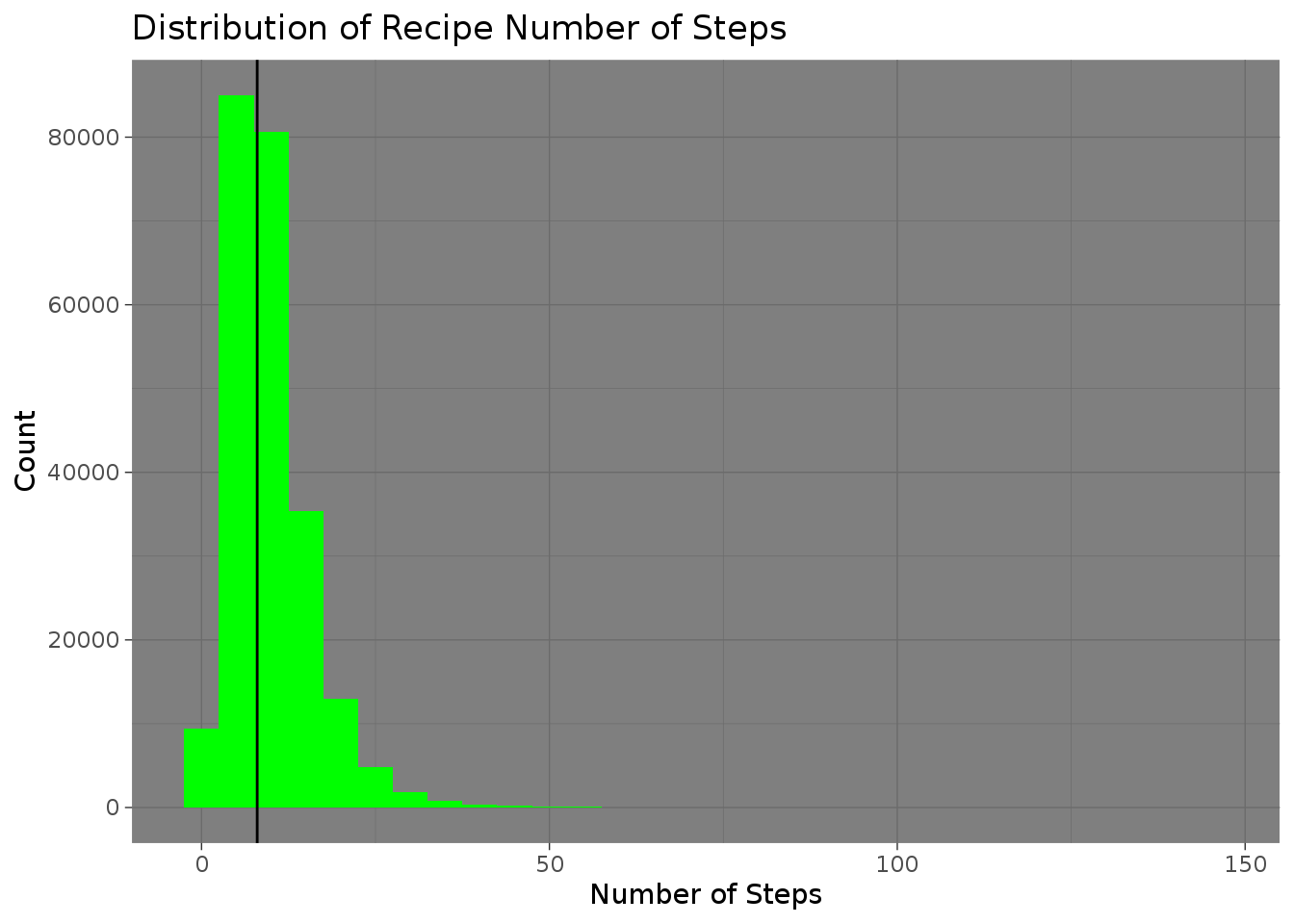
This analysis examines the distributions of various characteristics within a large dataset of recipes and user interactions from Food.com. By analyzing the number of ingredients, recipe ratings, number of preparation steps, cooking time, and temporal submission patterns.



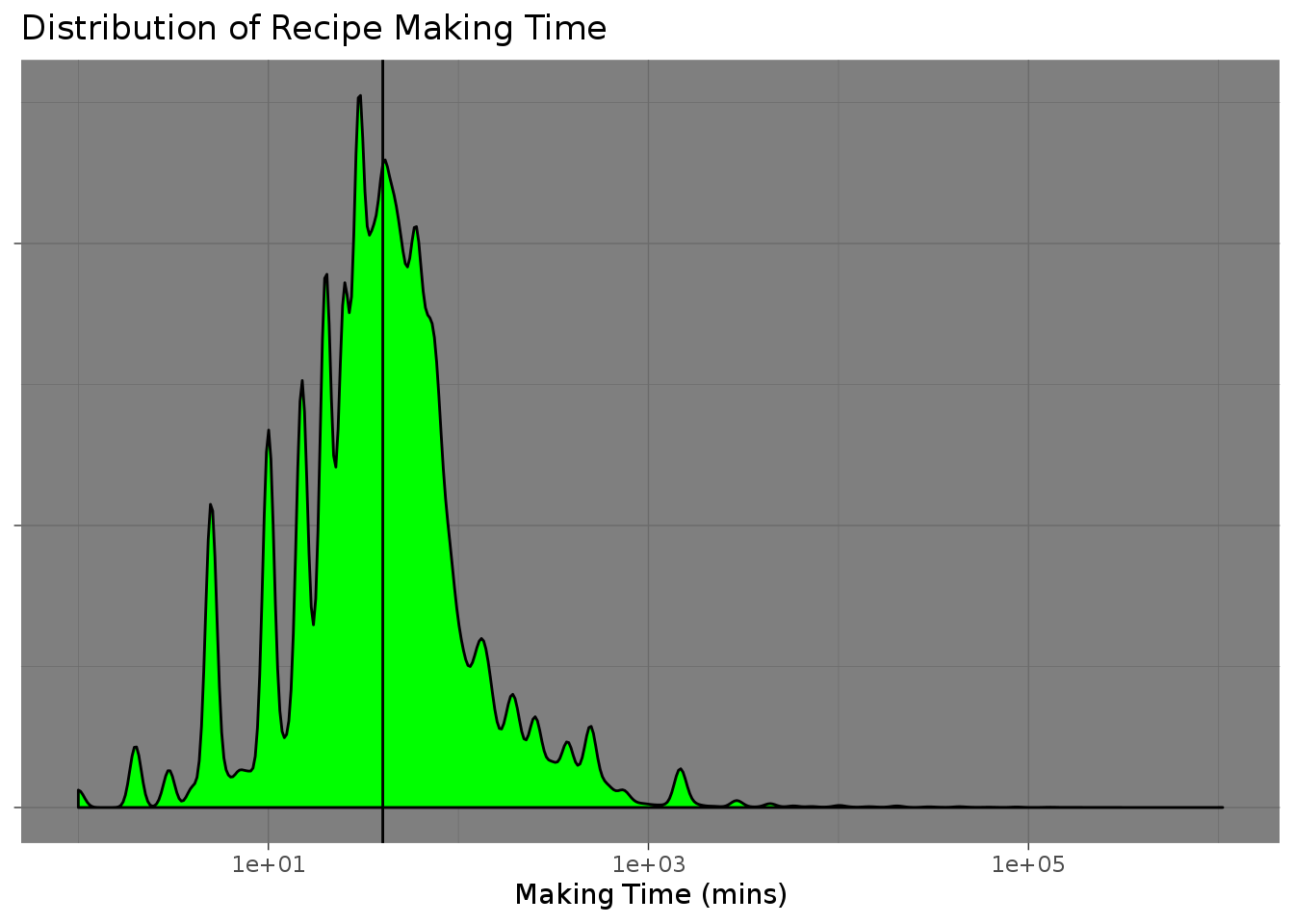
* **Distribution of Recipe Number of Ingredients:** This histogram suggests that most recipes contain between 5 to 15 ingredients, with two prominent peaks around 7 and 10. The distribution is right-skewed, indicating that while most recipes keep ingredients to a moderate number, there are recipes that require significantly more.



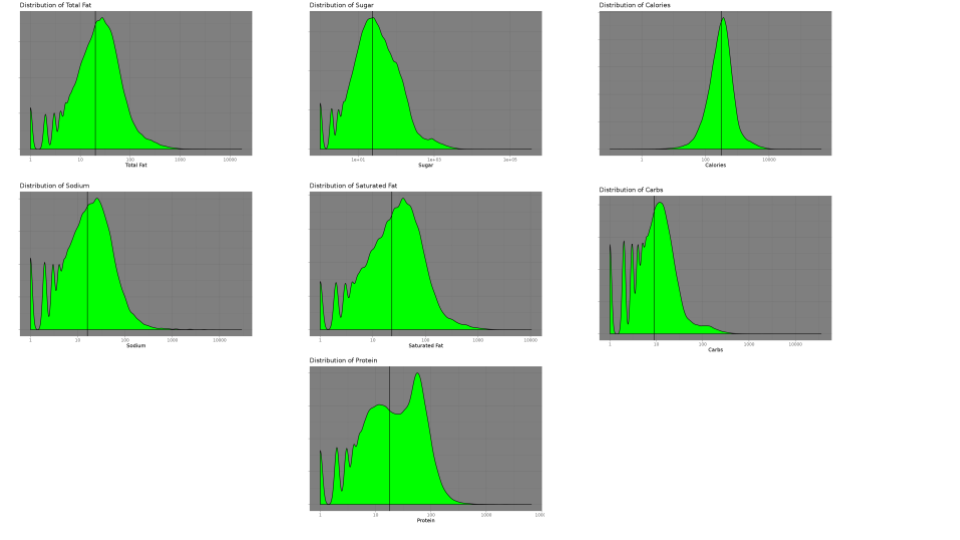
**Distribution of Recipe Ratings:** The distribution is highly left-skewed, showing that the vast majority of recipes have high ratings, with a significant peak at the rating of 5. This could indicate a tendency for users to rate recipes favorably or that only well-received recipes are commonly rated.



**Distribution of Recipe Number of Steps:** Recipes typically have a small number of steps, most often less than 10, as evidenced by the sharp peak at the lower end of the distribution. The long tail to the right suggests that there are few recipes with many steps, which could be more complex or involved dishes.

* 
* **Distribution of Recipe Making Time:** This plot shows that the majority of recipes require a relatively short preparation time, with a sharp peak at lower time values. The distribution has several peaks and is heavily right-skewed, implying that while most recipes are quick to make, there are some that take considerably longer.

(III) **Nutritional Analysis**: Each graph shows the frequency of recipes containing certain levels of calories, carbohydrates, proteins, fats, and other nutritional metrics.



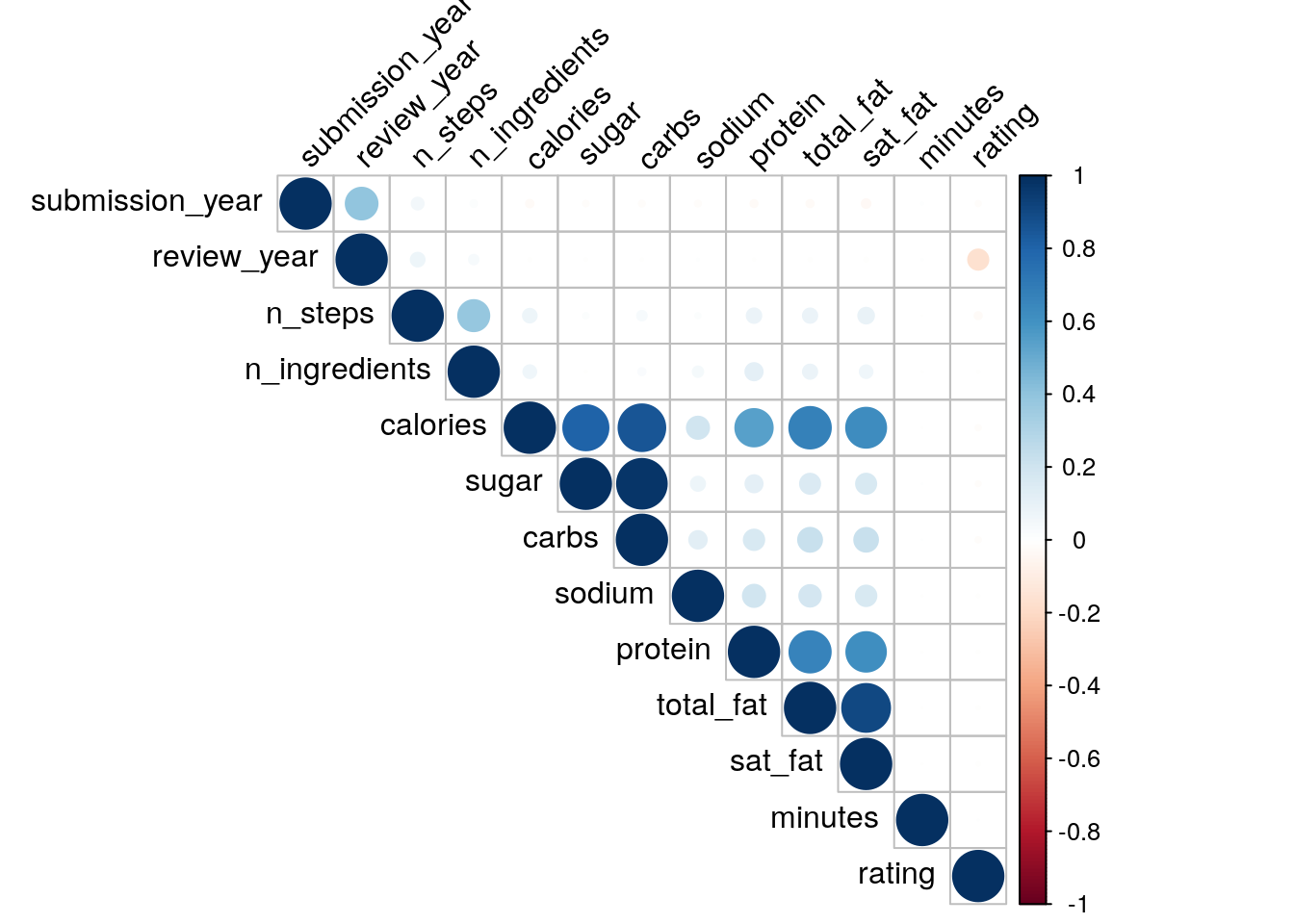
* **Distribution of Calories:** The calorie distribution is log-normal, peaking at a lower calorie count. This indicates that most recipes contain a moderate number of calories, aligning with general eating habits and possibly dietary guidelines. Median Calories: 312.7
* **Distribution of Carbohydrates:** The carbohydrate distribution shows multiple peaks and a right skew, which may suggest the presence of different categories of dishes, such as low-carb vs. high-carb recipes, and a tendency towards recipes with fewer. carbohydrates.Median Carbs: 9
* **Distribution of Protein:** Protein content also appears to be log-normally distributed with a right skew. The distribution suggests that recipes typically do not have a high protein content, with a few exceptions, possibly due to the inclusion of meat-heavy or high-protein dishes. Median Protein: 18
* **Distribution of Saturated Fat:** The distribution of saturated fat shows a right-skewed pattern. Most recipes contain lower levels of saturated fat, with fewer recipes having very high saturated fat content, potentially reflecting health-conscious recipe formulation. Median Saturated Fat: 23

**Distribution of Sodium:** Sodium levels show a multimodal distribution, indicating a wide variation in sodium content across recipes. This might reflect the varying use of salt and sodium-containing ingredients in different types of dishes. Median Sodium: 16

**Distribution of Sugar:** Sugar content shows a significant right skew, with most recipes containing lower amounts of sugar and a long tail towards the higher sugar content. This spread could include everything from savory dishes with minimal sugar to desserts and baked goods. Median Sugar: 24

**Distribution of Total Fat:** Similar to saturated fat, the total fat content in recipes is right-skewed. A majority of recipes appear to have a moderate amount of total fat, with a long tail suggesting the existence of richer, more indulgent recipes.

(IV) **Correlation Plot:** This correlation matrix provides insights into the relationships between different variables related to recipes and user interactions. A coefficient close to 1 indicates a strong positive correlation, meaning as one variable increases, the other tends to increase as well. A coefficient close to -1 indicates a strong negative correlation, meaning as one variable increases, the other tends to decrease. A coefficient around 0 suggests no linear correlation between the variables.



* **Number of Steps and Ingredients:** A moderate positive correlation between the number of steps and ingredients suggests that recipes with more ingredients tend to be more complex and require more steps to prepare.
* **Nutritional Content (Calories, Sugar, Carbs, Sodium, Protein, Total Fat, Saturated Fat):** There are strong positive correlations within nutritional content variables. For example, calories show a strong positive correlation with total fat and sugar, indicating recipes with higher caloric content tend to have more sugar and fat.
* **Minutes and Nutritional Content:** There is a noticeable negative correlation between minutes and rating, implying that recipes with longer preparation times might have lower ratings, potentially due to the preference for quicker recipes.
* **Rating:** Interestingly, there's a significant negative correlation between rating and saturated fat, suggesting that recipes with lower saturated fat content tend to have higher ratings, which could reflect health-conscious preferences among users.

**(V) Motivation**: Given dataset from Food.com is a comprehensive collection featuring over **180,000 recipes and 700,000 user reviews**, spanning a diverse range of culinary tastes and preferences. It's enriched with multifaceted features, including **detailed recipe steps, ingredient lists, nutritional information**, and **extensive user interaction data** such as ratings and reviews. This rich dataset not only captures the vast expanse of cooking styles and ingredients but also provides a deep dive into user engagement patterns, making it an invaluable resource for data-driven insights into culinary trends and consumer behavior.

### Predictive task

**(I) Evaluation of model wrt to relevant baselines and validity of our model’s predictions**

Evaluation of Modified BPR Model for Recipe Recommendation

When evaluating our **proposed modified Bayesian Personalized Ranking (BPR) model**, which incorporates features like nutrition data, tags, and cooking time, it's crucial to compare its performance against relevant baselines.

**Baseline Comparisons:**

* **Jaccard Similarity-Based Model (Baseline 1):**

This baseline approach utilizes Jaccard similarity, a method that measures similarity between finite sets. In our Jaccard similarity-based model, we identify similar recipes and users within the Food.com dataset. For each recipe, the model finds users who have cooked it and then identifies other recipes cooked by these users, calculating similarity using the Jaccard. Similarly, for each user, it identifies other users with similar cooking patterns by comparing the recipes they've cooked. This method leverages the overlap in user-recipe interactions, using the Jaccard to quantify the degree of similarity. This approach enables personalized recipe recommendations by matching users with recipes reflecting their cooking preferences and patterns, but it might not capture the nuanced preferences of users or the complexity of recipes beyond ingredient similarity.

**Vanilla BPR Model (Baseline 2):**

The Vanilla BPR model offers a more traditional collaborative filtering approach. It creates user-item pairs and utilizes negative sampling during validation to enhance the learning process. This model focuses primarily on the interactions between users and items (recipes), disregarding additional features like nutrition or cooking time. While effective in capturing user preferences based on historical interactions, it may lack the depth provided by incorporating specific recipe features.

Validity of the our **proposed modified Bayesian Personalized Ranking (BPR) model**:

The modified BPR model's validity and superiority over these baselines can be assessed through several metrics commonly used in recommendation systems:

* **Precision and Recall at K:** These metrics measure the model's ability to recommend a list of top-K items that the user will like. Higher precision and recall indicate better performance in capturing user preferences.
* **AUC (Area Under the ROC Curve):** This metric evaluates the model's ability to differentiate between positive (user-interacted) and negative (user-not-interacted) items. A higher AUC suggests better discriminative power.

By comparing these metrics across the baseline models and the modified BPR model, we can establish the effectiveness of incorporating additional features like nutrition, tags, and cooking time into the recommendation process. The modified BPR model outperformed the baselines, especially in scenarios where these additional features play a significant role in user preference. For instance, a user interested in healthy cooking might find better recommendations from a model that considers nutritional content.

In conclusion, the validity of the modified BPR model is evaluated not just on its ability to predict user-item interactions accurately but also on its effectiveness in leveraging the rich, multi-faceted data to deliver more personalized, context-aware recommendations.

**(II) Pre-processing of data and features you will use:**

* **Elimination of Quantifiers:**
  + Step 1: We commence by removing units of measure. These include terms like 'teaspoon', 'gram', and 'clove' which, though important for actual cooking, clutter our data and do not aid in differentiating recipes.

**Standardization of Ingredient Forms:**

* + Step 2: Punctuation is removed to prevent any discrepancies caused by syntactical differences.

**Linguistic Normalization:**

* Step 3: We employ lemmatization to normalize the inflected forms of words. This ensures that different forms of the same word, such as 'pound' and 'pounds', are identified as identical.

**Lowercasing and Accents Removal:**

* + Step 4: All words are converted to lowercase to eliminate any case-sensitive variations. Accents are also stripped to maintain the textual consistency of ingredients across diverse languages.

**Stop Words and Common Words Elimination:**

* + Step 5: We further refine the list by removing 'stop words', which are prevalent in the English language but offer no distinctive value. Additionally, **common cooking words like 'oil' and 'fresh'** that are ubiquitous across recipes are also removed to enhance the distinctiveness of our ingredient features.
* **Proposed word Disassembling**
  + Step 6: We broke down compound ingredients into their simpler forms and removed redundant descriptors. For example, the list ['prepared pizza crust', 'sausage patty', 'eggs', 'milk', **'salt and pepper'**, **'cheddar cheese'**] was transformed into ['prepared pizza crust', 'sausage', 'egg', 'milk', **'salt'**, **'pepper'**, **'cheese'**]. But this approach also led to individual ingredients, such as 'Apple Cider Vinegar', being split into 'Apple', 'Cider', and 'Vinegar', which could cause an issue. Nevertheless, this preprocessing step enhanced the clarity and utility of the dataset.

**Final Ingredient List Compilation:**

* + Step 7: This cleansed and processed words into ingredient lists. These lists represent the essence of each recipe decreasing the unique ingredient list from ~**15,000 to ~2,300**

**Features used to train the Model:** Ingredients, preparation time, Nutrition

**(III) Justifying the features using the results of exploratory analysis**

* **Ingredients:** The ingredient lists have been processed to exclude non-essential information from above steps, providing essence of each recipe. The lemmatization process ensures that the variations of an ingredient are treated uniformly, which is crucial for capturing the essence of recipes. Furthermore, ingredients are the primary factors that determine the uniqueness of a recipe and directly influence a user's decision to make a dish based on dietary preferences, allergies, or available pantry items.
* **Tags:** Tags act as a categorical summarization of a recipe, often including cooking style, regional cuisine, dietary considerations, and occasion. For example, comfort-food, 15-minutes-or-less etc. They serve as a quick reference for users to align recipes with their preferences or needs. Even though tags assisted the model to cluster similar recipes, we removed this to reduce complexity of the model.
* **Minutes (Preparation Time):** The negative correlation between preparation time and recipe ratings suggests that users prefer recipes that require less time to prepare. Incorporating 'minutes' as a feature can help tailor recommendations to fit the user's schedule constraints, increasing the likelihood of a recipe being chosen and made.
* **Nutrition:** Nutrition information is a critical factor for health-conscious users. The exploratory analysis indicated a negative correlation between saturated fat and recipe ratings, implying that users often favor recipes with lower saturated fat content. Including nutritional data as features allows the recommendation model to account for health and dietary preferences, which are increasingly important decision factors for many users.

### Designing an Appropriate model

**(I) Model Description**

Our enhanced Bayesian Personalized Ranking (BPR) model, tailored for recipe recommendation, combines unique features like **nutritional content** including calories, total fat, sugar, sodium, protein, saturated fat and carbohydrates, **cooking times**, and ingredients from the Food.com dataset. For training, to reduce the complexity, we decided to remove total fat, sugar, sodium and saturated fats from the model. This BPRbatchAdditional model employs multiple embeddings to represent these diverse features, enriching the traditional user-item interaction framework. Training involves generating negative samples alongside user-recipe pairs for validation sets, enhancing the model's discriminative power. This model offers a more contextually aware recommendation system, addressing not just user preferences but also aligning with their dietary needs and cooking interests.

**Hyperparameters** for Training:

K=10, lambda=0.0001, Threshold = -0.22 & Optimizer Adam 0.1 for 100 iterations

**(II) Explanation and justification for using this proposed model:**

Our decision to use the modified Bayesian Personalized Ranking (BPR) model was driven by its ability to integrate diverse and context-rich features such as nutritional content, cooking time, ingredients and categorical tags. This approach allows for more nuanced and personalized recipe recommendations. BPR because of its features like **Personalization** (learns individual user preferences), **Implicit Feedback** (user interactions with recipes) and **Probabilistic Approach**, handles the uncertainty and variability in user preferences.

**(III) Optimization Challenges and Comparative Model Analysis (Strength & Weaknesses)**

We tried multiple models and faced following challenges  
 **B1 - Jaccard Similarity Model Challenges:**

In implementing the Jaccard similarity model for recipe recommendation, we encountered significant scalability issues. The Food.com dataset features a high volume of user-recipe interactions, leading to long processing times for similarity computations. To resolve this, we reduced the dataset size, which resulted in the omission of crucial interaction data. This data reduction compromised the model's ability to capture the full spectrum of user preferences and recipe diversity. Consequently, the Jaccard similarity model became less effective, unable to provide accurate recipe recommendations. This limitation highlighted the model's inadequacy in handling large-scale, complex datasets.

**B2 - Vanilla BPR Model challenges:**

The Vanilla BPR model, focusing solely on user-item pairs, fell short as it didn't consider additional informative features like tags, ingredients, nutrition, and cooking time. This limitation hindered the model's ability to provide contextually rich and personalized recommendations, essential in a diverse dataset.

### Literature Survey

**(I) Origin of existing dataset and Utilization Overview**

The dataset in focus is sourced from Food.com, a prominent culinary website featuring an extensive range of recipes and user interactions. It encapsulates a wealth of data including over 180,000 recipes and 700,000 reviews. This dataset is used in our recommendation model to predict which user will make which recipe. This recommendation task was achieved by analyzing nutritional trends, and user personalizations.

**(II) Other similar datasets studied in the past**

Several similar datasets have been studied in the past in class. The **user game dataset**, provides insights into gaming preferences and behaviors, aiding in the creation of personalized game recommendation systems for each user. Similarly, the **Amazon Musical Instruments reviews** dataset is a resource for e-commerce trend analysis and sentiment assessment, while the **Goodreads reviews** dataset focuses on book recommendation systems and reading habit analytics. Each of these datasets shares a commonality in their use for building recommendation systems and understanding user behavior. However, the **Food.com** dataset's unique combination of recipe content with user interaction data to study food preferences in relation to nutritional information and cooking complexities. It stands out for its application in both health informatics and culinary sciences, bridging the gap between diet, personal taste, and cooking practices.

**(III) Current State-of-the-art methods in Data Analysis: Similarities and Differences with Our Findings**

Current state-of-the-art methods in recipe recommendation often involve complex machine learning algorithms, particularly **deep learning models** like neural collaborative filtering and recurrent neural networks. These models excel in capturing complex patterns in large datasets, such as user preferences and recipe characteristics. Recent trends also include the use of **natural language processing** (whose properties we are already using while preprocessing our data) to better understand recipe content and user reviews. Compared to advanced methods, our findings tend to offer little lesser accuracy, given the current dataset's complexity. However, they require significant computational resources and expertise, highlighting a trade-off between sophistication and practicality in recommendation system design.

### Results and conclusions

**(I) Proposed Model Performance and Significance Relative to Alternative Baselines**

Our proposed BPR model outperforms the baseline models on both validation and test datasets. The Jaccard similarity model (B1) showed the least accuracy, likely due to its simpler heuristic approach that doesn't capture the complex patterns within user interactions. This could also be due to our partial use of the dataset, which results in not all recipes and users being seen while training. The vanilla BPR model (B2) marked an improvement, benefiting from machine learning to discern user preferences. However, our proposed BPR model achieved the highest accuracy, which can be attributed to its incorporation of additional context features like nutrition, tags, and cooking time, providing a more personalized and holistic recommendation system. These results validate our decision to advance the BPR model with feature enrichment, leading to more precise recipe recommendations.

| **Models** | **Validation Accuracy(%)** | **Test Accuracy(%)** |
| --- | --- | --- |
| **B1 :**  **Jaccard** | 63.102 | 62.030 |
| **B2 : BPR** | 72.023 | 73.971 |
| **Proposed**  **BPR** | **76.230** | **76.877** |

**(II) Feature representations which worked well and which did not**

Features such as Date of interaction and Date recipe was submitted had little impact on recommendations. This is likely because these temporal factors don't capture a user's culinary preferences, which are more influenced by content like ingredients, nutrition, and cooking methods. Not all features worked well in nutrition. Calories, protein, and carbohydrates were used as they had the highest effects.

**(III) Interpretation of our model’s parameters**

In a Bayesian Personalized Ranking (BPR) model, the parameter 'K' represents the dimensionality of the latent feature space for users and items. In our model K=10, which is set at a higher value, accommodates the complexity arising from multiple features. This choice ensures that the model can effectively capture the correlations between these numerous features and enhances the overall recommendation quality.

**(IV) Success Factors of the Proposed Model and Reasons for Other Models' Failures**

The proposed BPR model succeeded by effectively integrating contextual features like nutrition, tags, and cooking times, which closely align with users' decision-making processes with **around 77% accuracy** while other baselines failed to account for these nuanced preferences, focusing mainly on interaction history, thus lacking the depth required for personalized, content-rich recipe recommendations.