Personalized Recipe Recommender System

- Author: Mounika Reddy Kummetha
- For: UMBC Data Science Master Degree Capstone by Dr. Chaojie (Jay) Wang

Background

- Understanding Personalized Recipe Recommender Systems:
  - Tailoring recipe suggestions to individual tastes and dietary needs.
  - Need for Personalization in Culinary Experiences:
  - Emphasizing technology's role in enhancing user culinary experiences.
- Project Goal:
  - Develop a system using data-driven techniques and machine learning.
  - Offer recipe recommendations based on user interactions and preferences.

Objective

Focusing on User Preferences:

- Catering to diverse culinary tastes and dietary restrictions.
- Ensuring an easy-to-use interface for personalized suggestions.

Anticipated Outcomes:

- Dynamic system learning from user interactions.
- Helping users discover recipes aligned with personal preferences.

Relevance in Today's World:

- Aligning with trends in health consciousness and digital solutions.
- Enhancing the daily life experience through personalized nutrition.

Data Overview

Dataset Overview:

Sourced from "Food.com Recipes and User Interactions" on Kaggle.

Comprehensive data on recipes, user profiles, and interactions.

Data Sources and Size:

URL: Food.com Recipes and User Interactions on Kaggle.

Total size: 104.59 MB.

Dataset

Data Structure:

- Collective datasets contain over (total number of rows) rows.
- RAW_recipes.csv: 231637 rows, 12 columns.
- RAW_interactions.csv: 1132367 rows, 5 columns.
- PP_recipes.csv: 178265 rows, 8 columns.
- PP_users.csv: 25076 rows, 6 columns.
- Interactions datasets: 23176 rows each, 5 columns.
- Ingr_map.pkl: 11659 rows, 3 columns.

About Dataset

Time Span Covered:

Data from 01/02/2010 to 01/02/2020.

Reflecting a decade of culinary trends and interactions.

Row Representation:

RAW_recipes.csv & PP_recipes.csv: Each row is a unique recipe.

RAW_interactions.csv, Interactions datasets: Each row is a user interaction (rating/review).

PP_users.csv: Each row is a user profile with preferences and dietary info.

ingr_map.pkl: Each row is an ingredient in its standardized form.

Data Preprocessing

Initial Data Assessment:

- Overview of datasets' initial state.
- Identification of missing values, duplicates, and inconsistencies.

Cleaning and Standardization:

- Removal of duplicates and handling of missing values.
- Standardization of data formats for consistency.

Data Transformation:

- Conversion of data into a usable format for analysis.
- Example: Transformation of categorical data into numerical values.

Preprocessed Data Overview

Feature Engineering:

- Identification and creation of new relevant features.
- Example: Derivation of a 'popularity score' from user interactions.

Data Reduction Techniques (if applied):

- Usage of techniques like PCA for dimensionality reduction.
- Explanation of the rationale behind data reduction.

Preprocessed Data Overview:

- Presentation of datasets after preprocessing.
- Highlight the improvements and changes compared to the initial data.

Ensuring Data Quality:

- Measures taken to ensure data reliability and quality.
- Addressing specific challenges encountered during preprocessing.

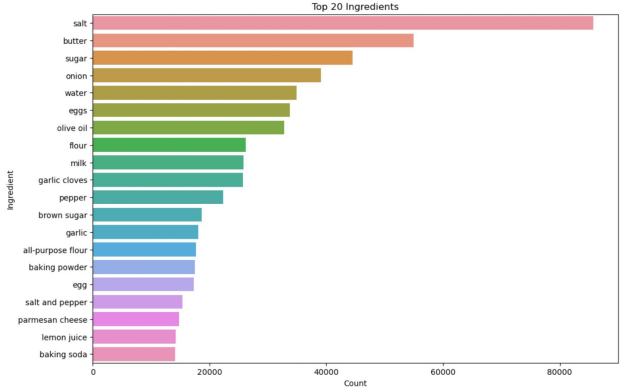
EDA Visualization

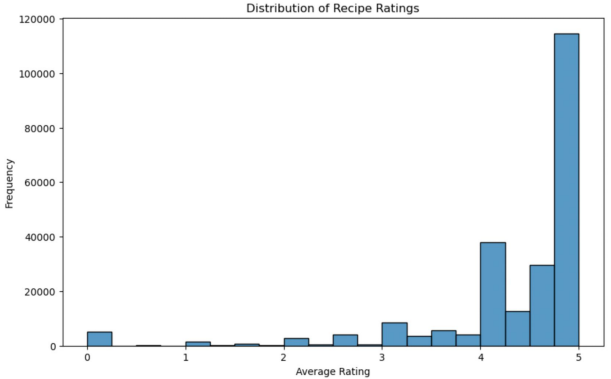
Distribution of Recipe Ratings

- The majority of recipes contain between 5 to 10 ingredients.
- There is a significant decrease in frequency as the number of ingredients increases, with v few recipes containing more than 20 ingredients.
- This distribution suggests that simpler recipes with fewer ingredients are more common o preferred.

Top 20 ingredients

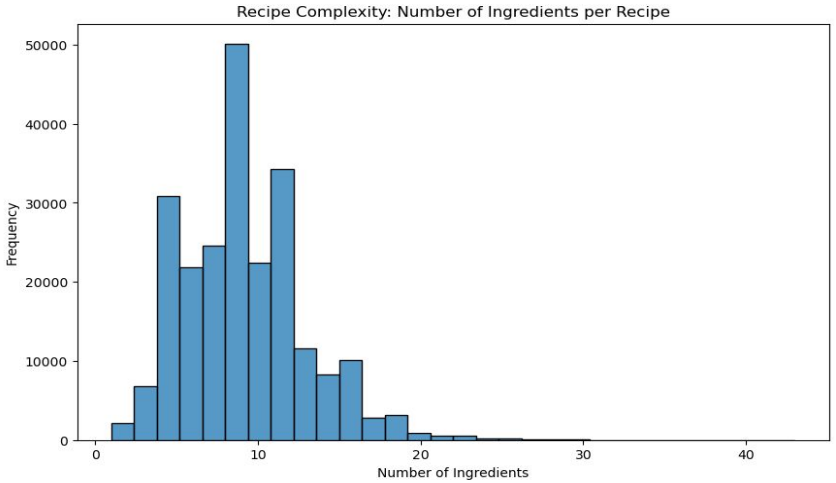
- Salt is the most commonly used ingredient, followed by butter and sugar.
- Basic baking ingredients like flour, eggs, and milk are very common, indicating a prevalence of baking recipes.
- Olive oil and garlic are the most common ingredients used in savory dishes.
- There is a significant drop in frequency from common ingredients like salt to less common ones like lemon juice and baking soda.





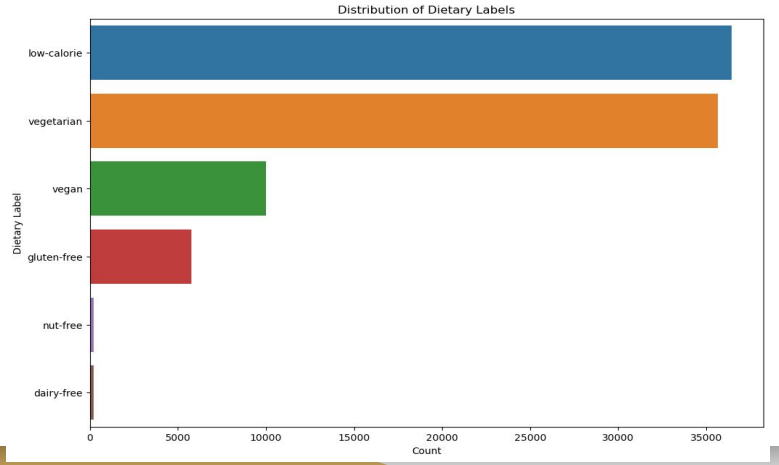
EDA

- The majority of recipes contain between 5 to 10 ingredients.
- There is a significant decrease in frequency as the number of ingredients increases, with very few recipes containing more than 20 ingredients.
- This distribution suggests that simpler recipes with fewer ingredients are more common or preferred.



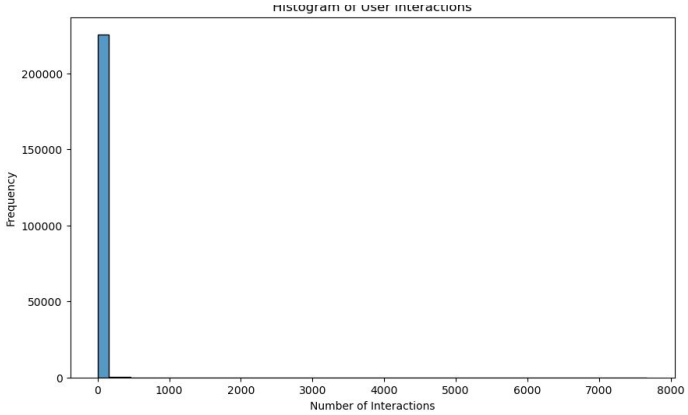
- Low-calorie recipes are the most frequent, followed by vegetarian recipes, indicating a possible trend or preference for healthier eating options.
- Vegan recipes are also quite common, which could suggest a significant vegan demographic or interest in vegan cooking.
- Gluten-free and nut-free recipes are less common, but still significant, suggesting awareness and accommodation for these dietary restrictions.

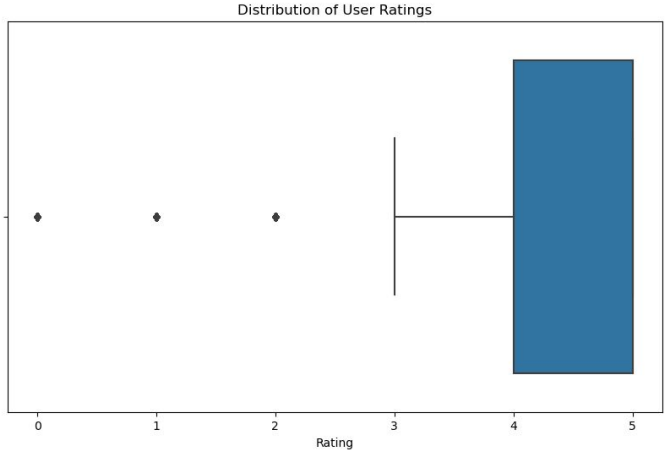
  Dairy-free recipes are the least common among the die



EDA part 2

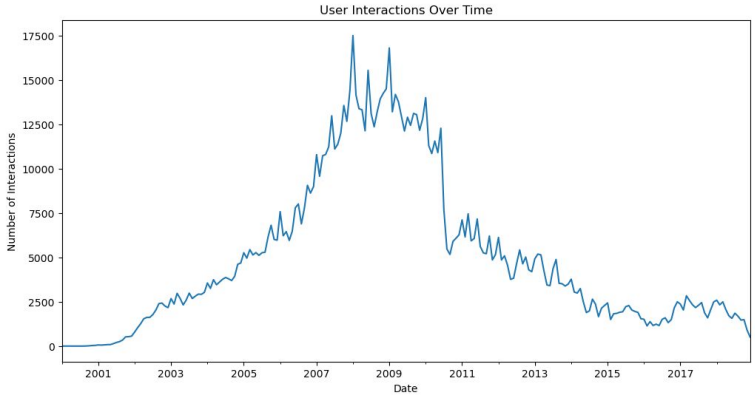
- A significant proportion of ratings are at the highest score, which is 5. This could indicate a tendency for users to rate only when they have a positive experience or are satisfied.
- Ratings 1 to 4 are much less frequent, with very few users giving these scores compared to the rating of 5.
- The data may suggest a possible issue with rating scale usage or could imply that the user base generally experiences positive outcomes with whatever is being rated.
- The vast majority of users have a very low number of interactions, close to 0, indicating a highly skewed distribution of interactions.
- There are very few users with a high number of interactions, as evidenced by the frequency dropping to near zero past the first bin of the histogram.
- This distribution suggests that most users are only minimally engaged.

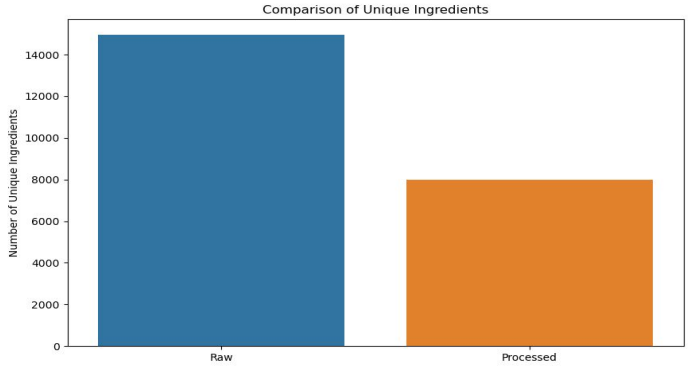




EDA part 2

- The graph compares the number of unique ingredients between raw and processed categories.
- Raw ingredients significantly outnumber processed ones, with over 14,000 unique raw ingredients compared to around 8,000 processed ingredients.
- This suggests that there is a greater variety of raw ingredients used than processed ones, which could indicate a broader diversity in raw food items or possibly a tendency for processed foods to utilize a more limited set of common ingredients.



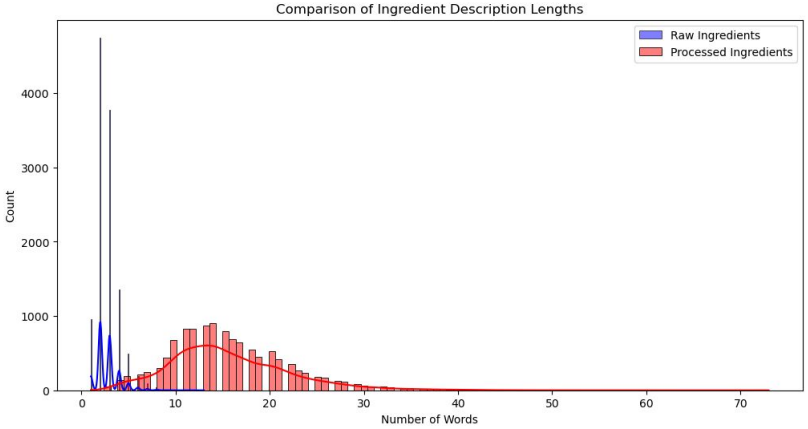


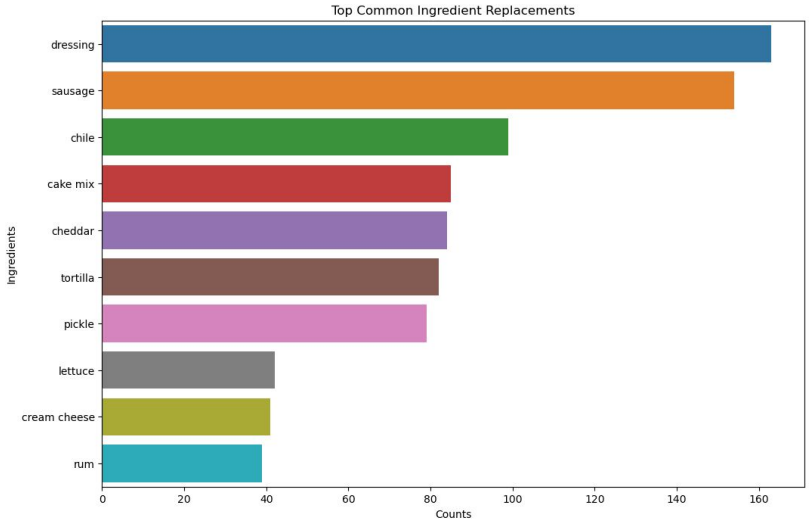
- User interactions have fluctuated over time with several peaks and troughs.
- There was a significant increase in interactions leading up to around 2009, after which a major peak is observed.
- After the 2009 peak, there is a general decline, with some variability, suggesting a decrease in user engagement over time or changes in the user base.
- The reasons behind the fluctuations would require further contextual information to understand the driving factors behind these patterns

EDA part 2

- The second graph illustrates the top common ingredient replacements.
- The most replaced ingredient is dressing, followed by sausage, chile, cake mix, and cheddar.
- Each of these ingredients has a specific count of replacements, with dressing exceeding 160 occurrences.
- This graph might indicate the frequency of ingredient substitutions in recipes, which could reflect dietary preferences, availability, or trends in cooking.

- The third graph compares the length of ingredient descriptions between raw and processed ingredients.
- Raw ingredients tend to have descriptions that are shorter in word count, with a high frequency of descriptions that are only 1-2 words long.
- Processed ingredients have a more spread out distribution with a peak around 10 words, suggesting that processed ingredients often require longer descriptions, possibly due to more complex names or additional qualifiers (like brand names or preparation styles).
-





Using RandomForestClassifier

Involves setting up a RandomForestClassifier, a popular machine learning model suitable for handling both classification and regression tasks.

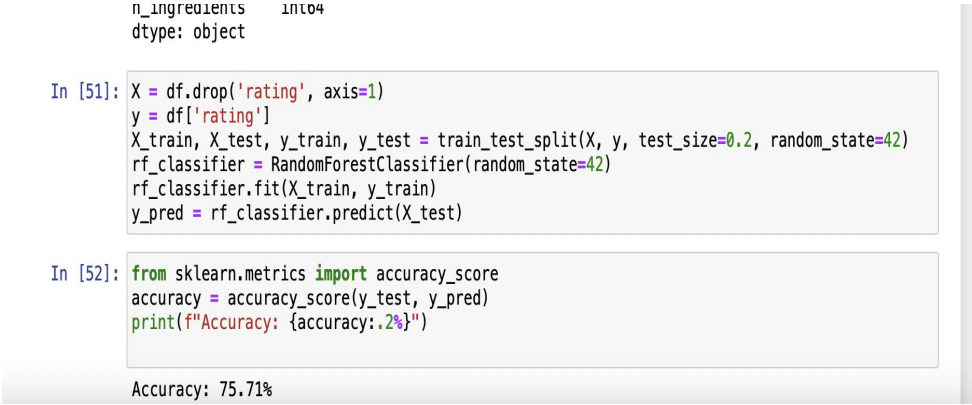
The dataset is split into training and testing subsets to validate the model's performance.

After training the model with the training data, predictions are made on the test set.

Model Training

In this phase, we are attempting to predict the ratings using the number of minutes it takes to prepare the dish, the number of steps, and the number of ingredients. I have employed a RandomForest Classifier to predict the ratings for each dish.

Accuracy: 75.71%



Conclusion

Utilization of User Data for Effective Recipe Recommendations

Engaging Casual Users
- Most users have limited interactions, indicating a large casual user base.
- Tailoring recommendations to these users could increase engagement.

Leveraging High Ratings
- User ratings skew towards higher values, showing satisfaction with chosen recipes.
- Recommend high-rated recipes more frequently to attract user interest.

Seasonal and Time-Sensitive Recommendations
- Analysis of user interactions over time reveals patterns in recipe popularity.
- Use this data to suggest seasonal or trending recipes.

Conclusion

Efficient Machine Learning Models, Features, and Understanding User Preferences

Identifying Complex Data Relationships
- Scatter plots suggest complex, non-linear relationships in recipe data.
- More sophisticated models may be required to capture these nuances.

Non-Linear Models for Better Feature Interaction
- Lack of strong correlations between key features like 'minutes', 'n_steps', and 'n_ingredients'.
- Indicates the need for non-linear models or advanced feature engineering.

Exploring Underutilized Recipes and User Preferences
- Analysis of popular ingredients and simplified ingredient lists can inform recommendations.
- Correlation analysis highlights key predictive features for user engagement, like the number of interactions and average ratings.

Challenges and Solutions

Challenges Faced:

- During the course of this project, we encountered several challenges that tested our problem-solving skills and determination. Some of the major challenges included:
    - Data Quality: Dealing with missing values, duplicates, and inconsistencies.
    - Model Complexity: Managing the complexity of machine learning models.
    - User Diversity: Catering to a wide range of user profiles and preferences.

Solutions Implemented:

- We adopted a proactive approach to address these challenges, implementing the following solutions:
    - Data Quality Assurance: Rigorous data cleaning and curation techniques to ensure data reliability.
    - Model Complexity Management: Employing model simplification techniques to enhance model interpretability.
    - User Diversity Handling: Developing user-centric strategies to provide personalized recommendations for diverse user groups.

Future Enhancements

Continuous Improvement:

- Our commitment to excellence extends beyond the current project. We are dedicated to continuously enhancing our recipe recommender system to provide users with the best culinary experiences.

Enhancement Ideas:

- In the future, we plan to implement several enhancements, including:
  - User Interface Refinement: Creating a more intuitive and user-friendly interface.
  - Algorithm Refinement: Fine-tuning our recommendation algorithms for improved accuracy.
  - Expanding Data Sources: Incorporating additional data sources to enrich recommendations.

User Feedback Integration:

- User feedback is invaluable to us. We will integrate user suggestions and preferences to further personalize recommendations and enhance the user experience.