Recipe Recommender Assignment

November 07, 2023

Overview

As an ML engineer at food.com, our task is to design a recommender system that suggests recipes to users based on their choices and the recipe they are currently viewing. A successful recommender system can increase user engagement and lead to more business opportunities. The performance of the recommendation engine will directly impact the revenue generated by the website. However, building a recommender from scratch is time-consuming. In this assignment, you will explore data and create features to build the recommender.

Steps to approach the problem:

- Create and launch an EMR Cluster on Amazon AWS
- Create and launch a Jupyter Notebook on top of this cluster
- Perform all the necessary tasks provided in task list

Task List

Task 1	Task 2	Task 3
Read the data: Read RAW_recipes.csv from S3 bucket. Ensure each field has the correct data type.	Extract individual features from the nutrition column: Separate the array into seven individual columns to create new columns named calories, total_fat_PDV, sugar_PDV, sodium_PDV, protein_PDV, saturated_fat_PDV, and carbohydrates_PDV.	Standardize the nutrition values: Convert the nutritional values to per 100 calories.

Task List

- Task 3:
- Standardize the nutrition values: Convert the nutritional values to per 100 calories.
- Task 4:
- Convert the tags column from a string to an array of strings: Convert the tags column from a string to an array of strings.
- Task 5:
- Read the second data file: Read the RAW_interaction.csv and join this interaction level file with the recipe level data frame. The resulting data frame should have all the interactions.
- Task 6:
- Create time-based features: Create features that capture the time passed between one review and the date on which the recipe was submitted. Use the review_date and the submitted columns after you join the two data files.
- Task 7: Processing Numerical Columns (Optional): Convert all numerical columns to categorical columns using the percentile approach to decide the category boundaries. After creating buckets, study the variation of the average rating for each bucket and decide whether or not a particular bucketed column should be kept in the analysis.

- Task List

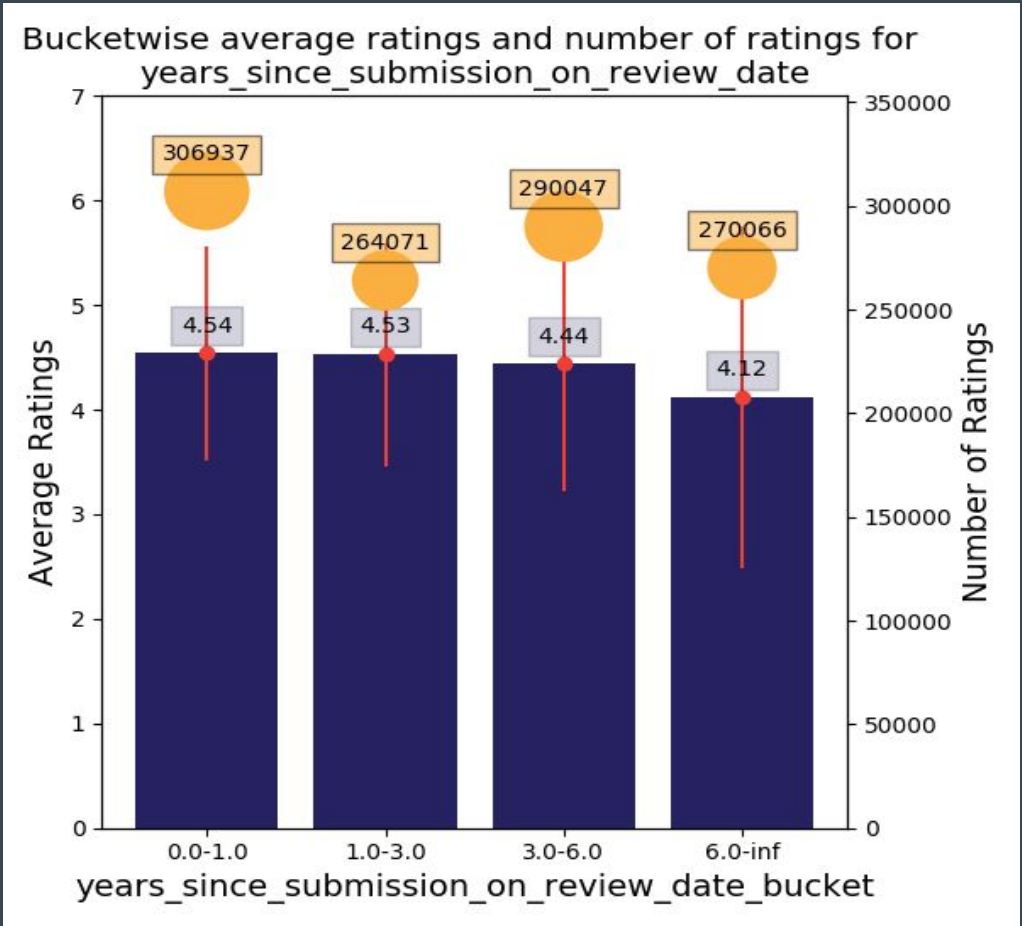
Task 8:

Create user-level features (Optional): 1.Create user-level features to capture intrinsic feedback. 2.Create columns such as user_avg_rating, user_avg_n_ratings, user_avg_years_betwn_review_and_submission, user_avg_prep_time_recipes_reviewed, user_avg_n_steps_recipes_reviewed, user_avg_n_ingredients_recipes_reviewed, user_avg_years_betwn_review_and_submission_high_ratings, user_avg_calories_recipes_reviewed, user_avg_total_fat_per_100_cal_recipes_reviewed, user_avg_sugar_per_100_cal_recipes_reviewed, user_avg_sodium_per_100_cal_recipes_reviewed, user_avg_protein_per_100_cal_recipes_reviewed, user_avg_saturated_fat_per_100_cal_recipes_reviewed, user_avg_carbohydrates_per_100_cal_recipes_reviewed, user_avg_prep_time_recipes_reviewed_high_ratings, and user_avg_n_steps_recipes_reviewed_high_ratings. 3.After these columns are created, do a thorough data check. You might have introduced null values to the data during your transformations. You can also do the bucketing exercise on user-level features.

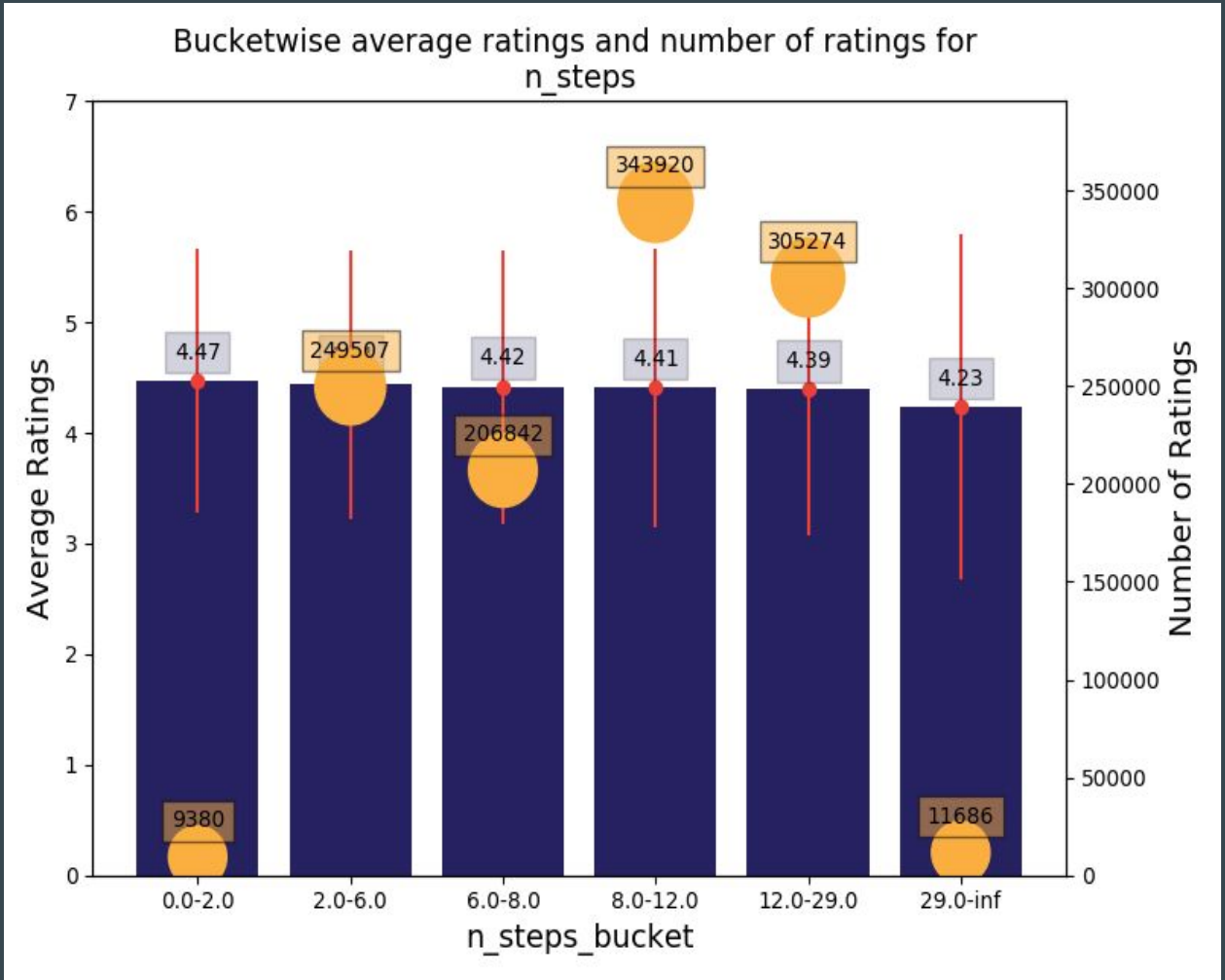
- Task 9:
- Create tag-level features (Optional): Extract tags-level features by exploring all the available tags. Create new columns to capture the unique tags and their frequency in the dataset.

Exploratory Data Analysis

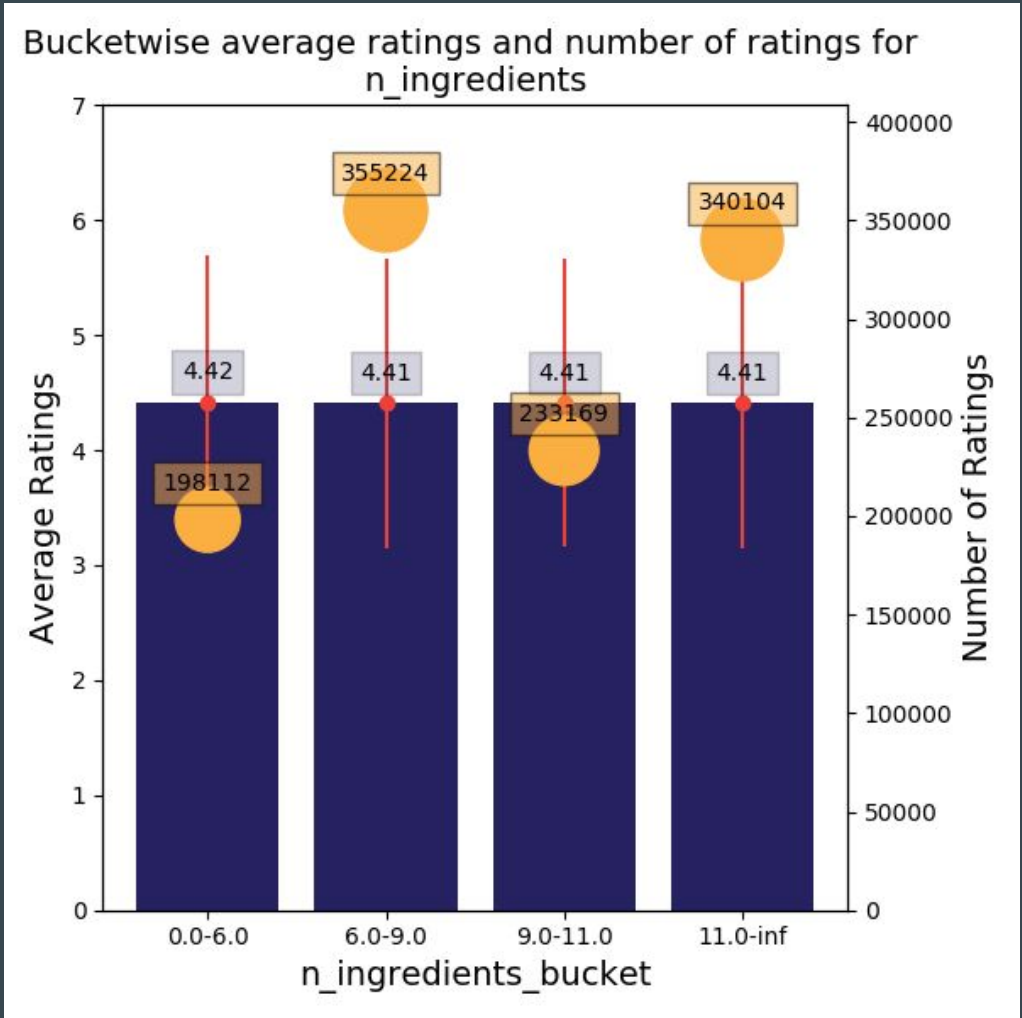
After import, inspect, converted and cleaning the dataset, we started exploring the data. We got major features which could help for better recommendation. Later we saved this in Parquet file which demonstrates in PySpark code.



Summary statistics for ratings in buckets.



Bucketwise average ratings and number of ratings for n_steps



Bucketwise average ratings and number of ratings for n_ingredients

Conclusion

Features	years_since_submission_on_review_date
Features	Minutes in preparation time
Features	Ingredients bucket
Features	• high_ratings = 5 rating • user_avg_years_betwn_review_and_submission_high_ratings • user_avg_prep_time_recipes_reviewed_high_ratings • user_avg_n_steps_recipes_reviewed_high_ratings • user_avg_n_ingredients_recipes_reviewed_high_ratings

The Team

Biplab Mondal

Vijaya

Indrajit Bose