RECIPE RECOMMENDATION ASSIGNMENT

EDA USING PYSPARK

SOUMYA ROY

TABLE OF CONTENTS

PROBLEM STATEMENT

- Let us step into the shoes of an ML engineer working at food.com. Our job is to design a recommender system to recommend recipes to users based on their choice and the current recipe they are looking at.
- Utilizing a recommendation engine can effectively enhance user engagement on your website. When users are presented with pertinent recipes, their inclination to stay longer on your site and explore various culinary offerings increases.
- This heightened user engagement has the potential to lead to expanded business prospects, such as partnerships, promotional opportunities, and more.
- However, designing a recommender from scratch is a time-consuming task. So, in this assignment, let us explore the data and create features that will be used to build the recommender.

STEPS TO APPROACH THE PROBLEM

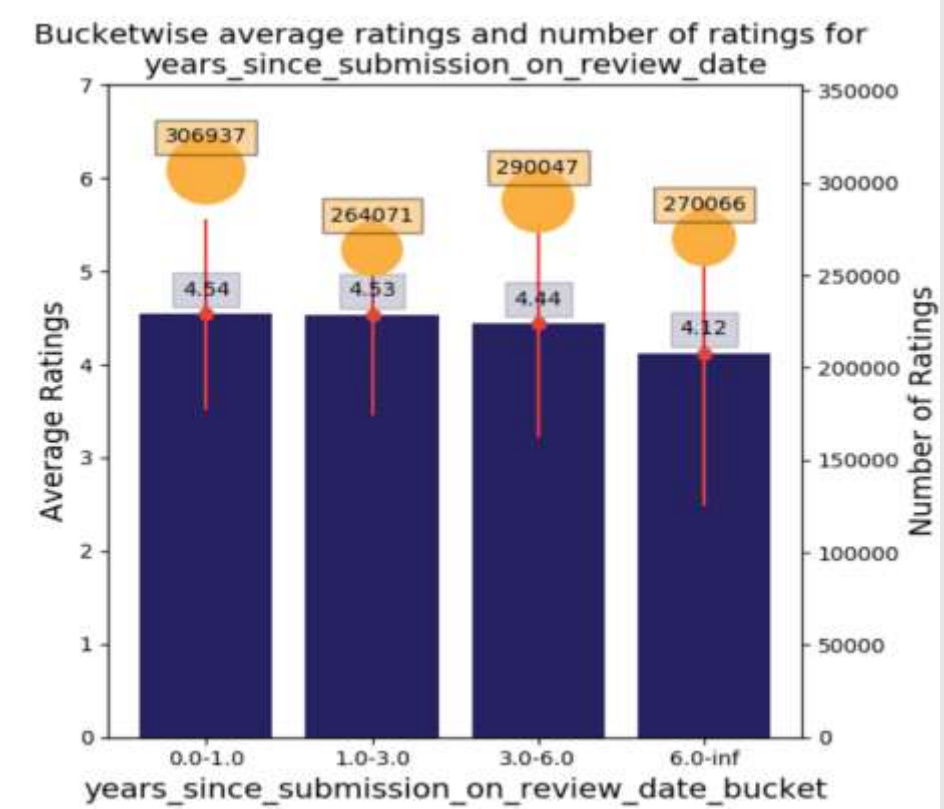
- To get started, launch AWS (Amazon Web Services).
- Create an EMR cluster on AWS
- Create a bucket list to put it to use in the later stages to store data set.

PROBLEM METHODOLOGY

In this section, we have elaborated each proactive task that was taken to create features for the recipe recommendation system.

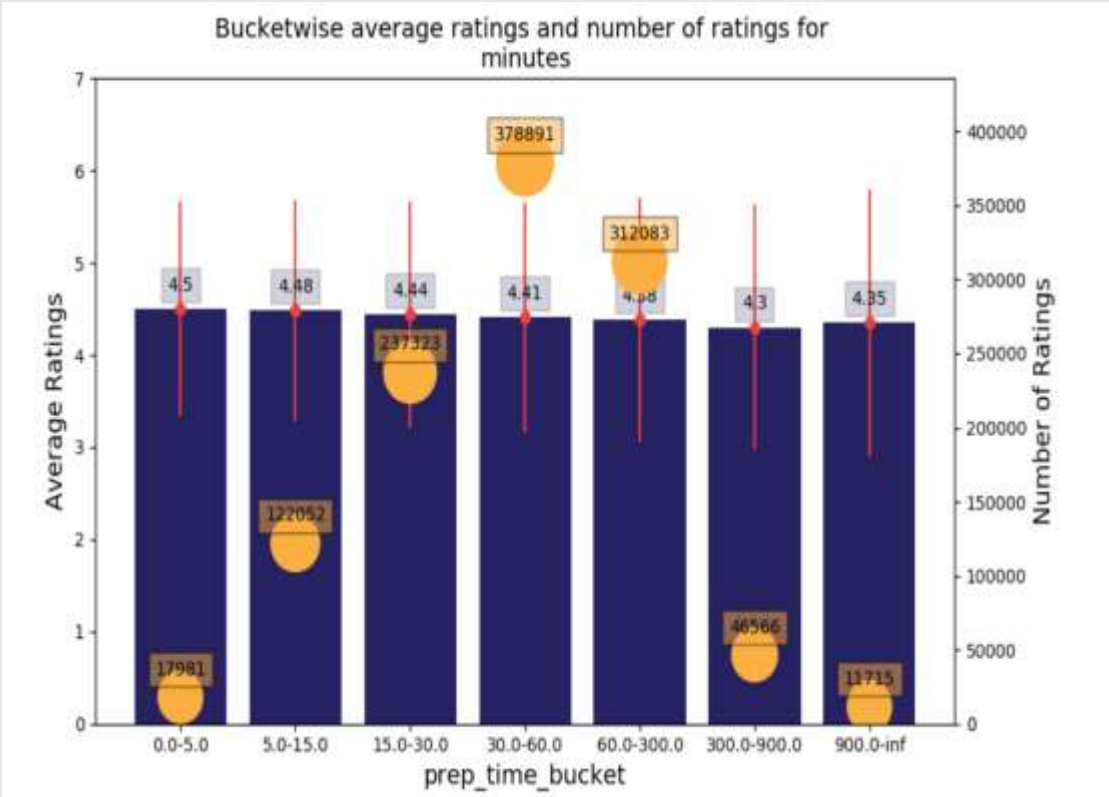
- **Task 1:** Read the data: Read RAW_recipes.csv from S3 bucket. Ensure each field has the correct data type.
- **Task 2:** Extract individual features from the nutrition column, and split them into seven new columns: calories, total_fat_PDV, sugar_PDV, sodium_PDV, protein_PDV, saturated_fat_PDV, and carbohydrates_PDV.
- **Task 3:** Standardize the nutritional values by converting them to values per 100 calories.
- **Task 4:** Change the format of the tags column to an array consisting of multiple strings.
- **Task 5:** Read the second data file which is RAW_interaction.csv, and then merge/join this interaction-level file with the recipe-level data frame. The resulting data frame should encompass all interactions.
- **Task 6:** Generate time-related features that reflect the duration between a review and the recipe submission date. Utilize the review_date and submitted columns from the merged data files to compute these features.
- **Task 7:** Optionally, categorize numerical columns based on percentiles to define category boundaries. After creating these buckets, analyze the average rating variation for each bucket to determine whether to retain the categorized column in the analysis.
- **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 (EDA)



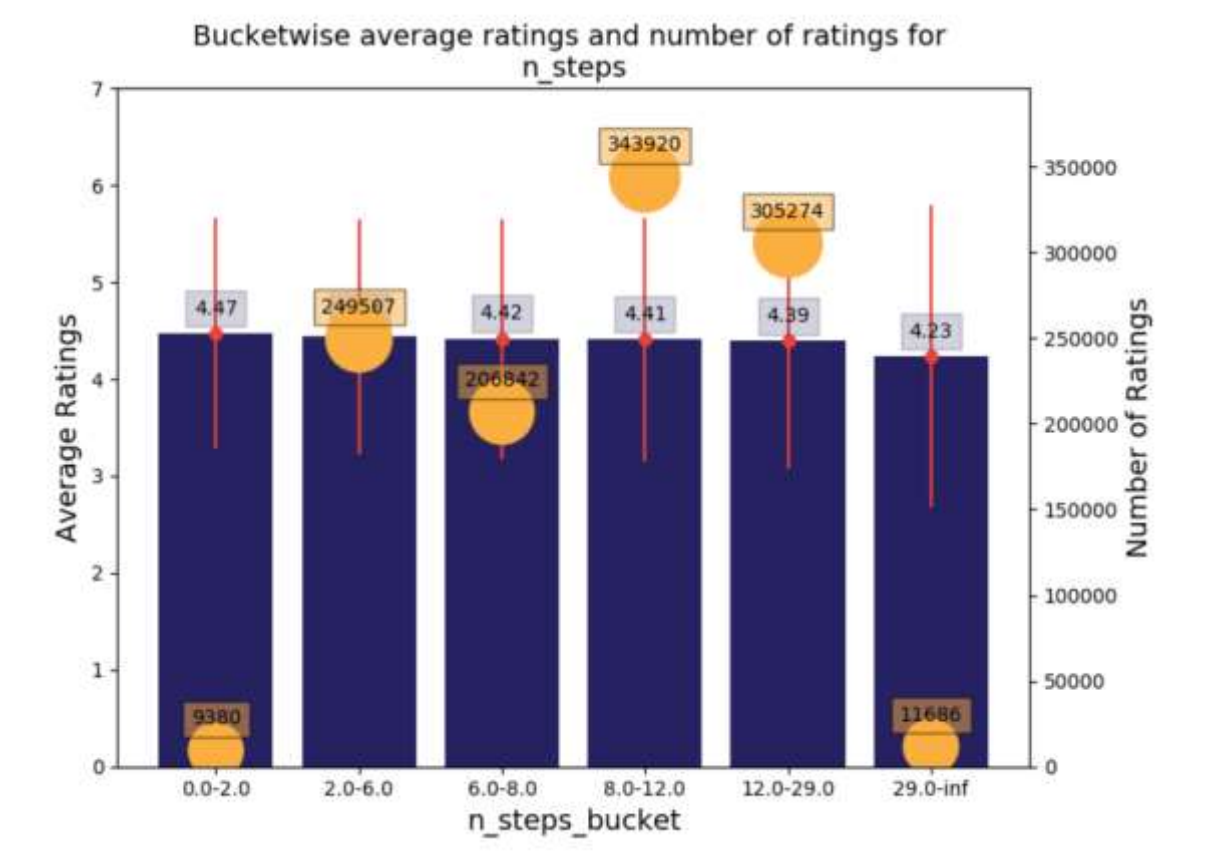
Observation

- Recipes more than 6 years old are rated low



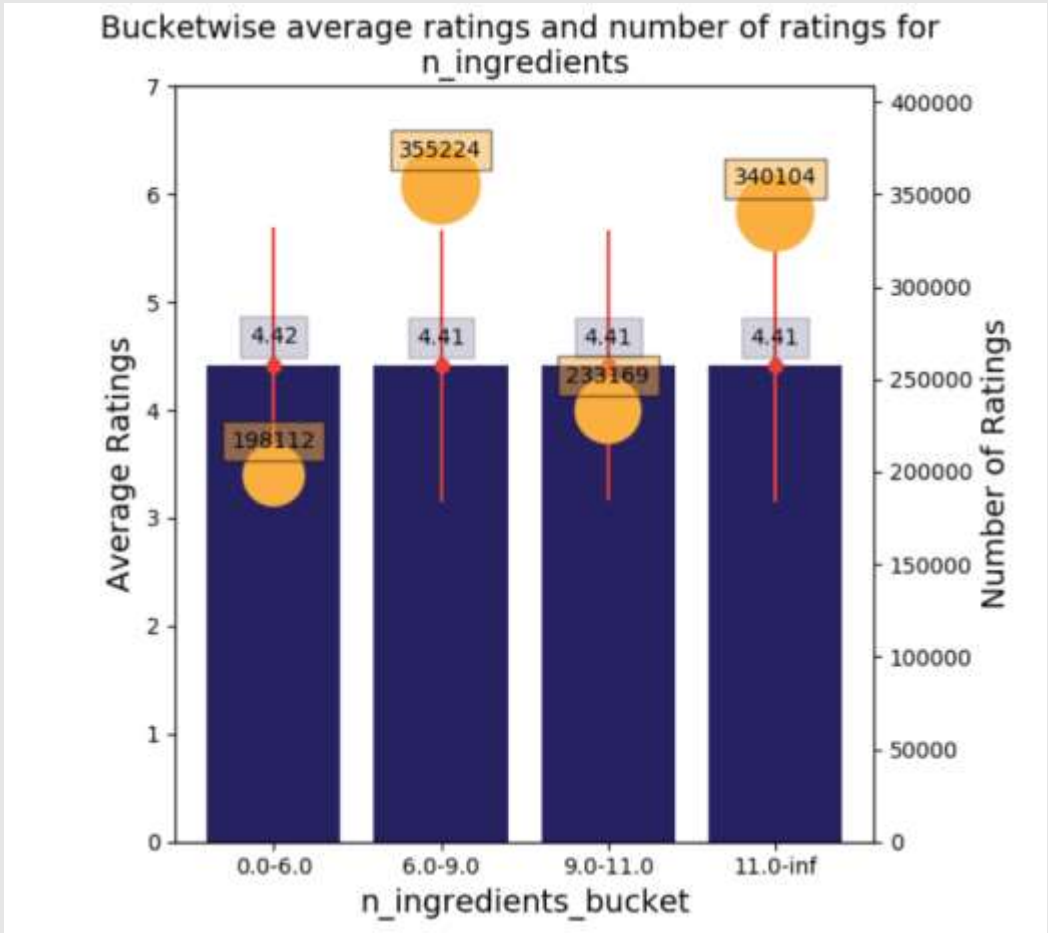
Observation

- The prep time is clearly relevant.
- Recipes with low preparation time received Highest Ratings.



OBSERVATION

- Recipes with less than 2 steps are rated high.
- Recipes with more than 29 steps are rated very low.



OBSERVATION

- No. of Ingredients are not relevant.

TOP 20 TAGS WITH HIGHEST COUNT OF USER RATING

individual_tag	avg_user_rating	n_user_ratings
preparation	4.411354948648275	1129076
time-to-make	4.4138409064517266	1110881
course	4.411947223592061	1076769
dietary	4.411331005472605	906698
main-ingredient	4.423460446350231	868466
easy	4.41807043683746	632226
occasion	4.41390652993341	622916
equipment	4.414997188352374	499707
cuisine	4.416292206025025	481359
low-in-something	4.4141343197390395	448497
main-dish	4.395603853828529	385383
60-minutes-or-less	4.405608852652683	343671
number-of-servings	4.406268866680343	341210
meat	4.4078258997788184	320551
taste-mood	4.412528439448443	311627
north-american	4.412832602505994	284837
30-minutes-or-less	4.426849710766375	267258
vegetables	4.454102520853779	260864
oven	4.417068897708177	251721
4-hours-or-less	4.383291661130505	248367

NUTRITION COLUMNS

- calories - Calories per serving seems irrelevant
- fat (per 100 cal) - Calories per serving seems irrelevant
- sugar (per 100 cal) - Calories per serving seems irrelevant
- sodium (per 100 cal) - Calories per serving seems

irrelevant

- protein (per 100 cal) - Calories per serving seems irrelevant
- sat. fat (per 100 cal) - Calories per serving seems irrelevant
- carbs (per 100 cal) - Calories per serving seems irrelevant

MORE FEATURES:

Highest Rating = 5

- 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

TOP 5 RATED TAGS

individual_tag	avg_user_rating
side-dishes-beans	5.0
cabbage	5.0
heirloom-historic...	5.0
middle-eastern-ma...	5.0
breakfast-potatoes	5.0

BOTTOM 5 LEAST RATED TAGS

individual_tag	avg_user_rating	n_user_ratings
cranberry-sauce	5.0	1
pot-roast	0.0	1
main-dish-seafood	0.0	1
ham-and-bean-soup	4.0	1
lamb-sheep-main-dish	0.0	1

CONCLUSION

From the notebooks, we can deduce the following:

- The time elapsed since a recipe's submission, along with factors like the number of steps, preparation time, and the quantity of ingredients, significantly influence the recipe's rating.

- Recipes that receive user reviews after a substantial time has passed since their submission, and possess fewer steps, shorter preparation times, and a limited number of ingredients, tend to garner high ratings, often achieving a perfect score of 5.

- In contrast, the number of ingredients within a recipe does not appear to correlate with the recipe's rating. Likewise, nutritional information such as calories, fat content, sugar, sodium levels, protein content, and saturated fat per serving do not seem to play a pivotal role in determining a recipe's rating.

THANK YOU