Recipe Recommender System Using EDA

Done by : Gondesi Rudra Deepak

Batch .No : ADSBC-C02

Problem Statement

- As a Machine Learning Engineer at food.com, the objective is to design a recommendation system that elevates user engagement.
- The core purpose of this recommendation engine is to deliver personalized recipe suggestions to users, based on their choice and the current recipe they are looking at.
- In this project, the primary focus lies in the exploration of data and the crafting of features that form the foundational components for building a reliable and effective recommendation engine.

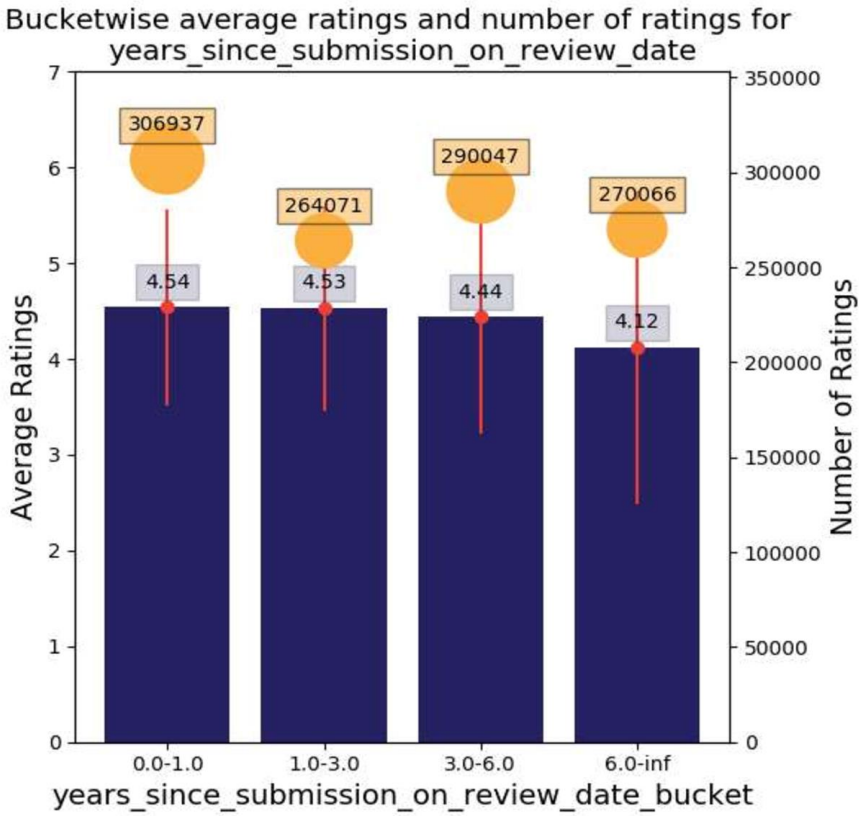
Business Objectives

- The recommendation engine is a way to increase the website's user engagement.
- If a user is shown relevant recipes, they are more likely to spend more time on the website reading about recipes. Higher user engagement will likely result in more business opportunities like collaborations, promotions, etc.
- The performance of a recommendation engine will significantly impact the revenue the recipe website can generate.
- Analysis and feature engineering is done using Spark on Elastic Map Reduce service (EMR).

Understanding Dataset

- The first file is the Raw_recipes.csv file. It contains all the recipe-related information. Each row in this file describes a recipe.
- The second file is the RAW_interactions.csv. Each row in this data file is one user reviewing one recipe. One user can review more than one recipe, and each recipe can be reviewed by more than one user.

Exploratory Data Analysis(EDA)



years_since_submission_on_review_date

(Review Time Since Submission)

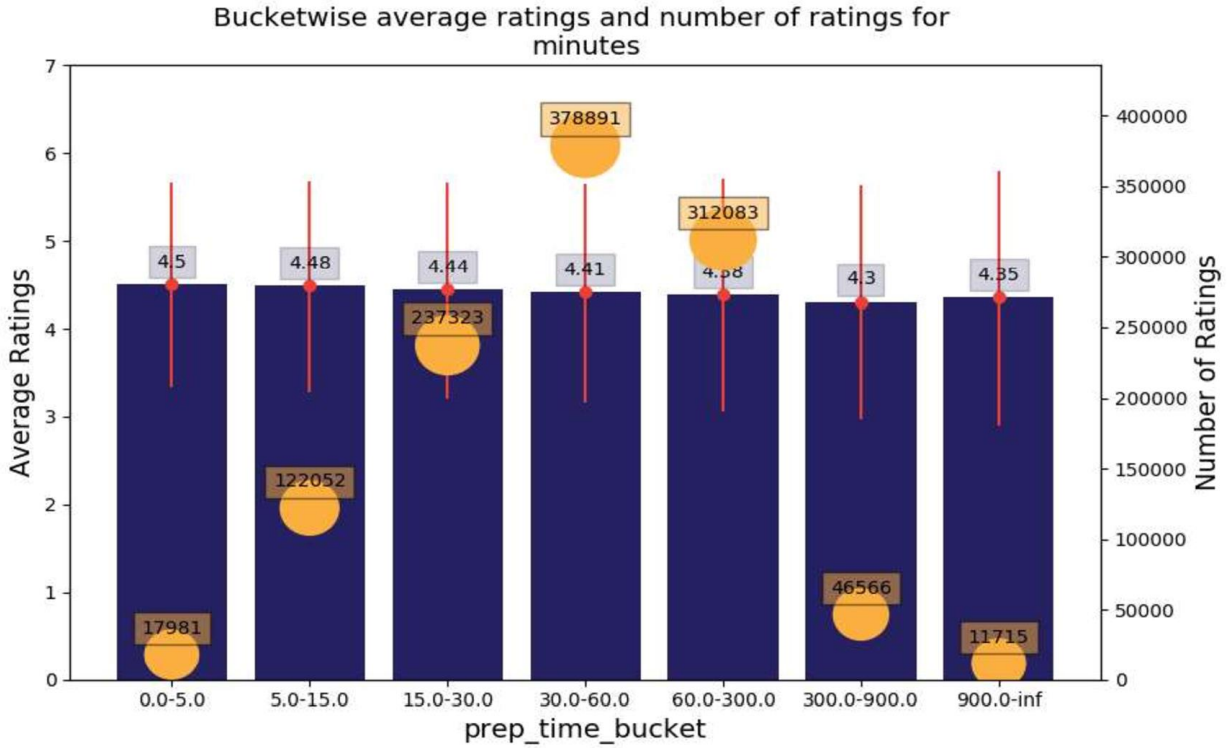
Upon analyzing the 'Review Time Since Submission' data, it becomes evident from the graphical representation that recipes older than six years tend to receive lower ratings.

Exploratory Data Analysis(EDA)

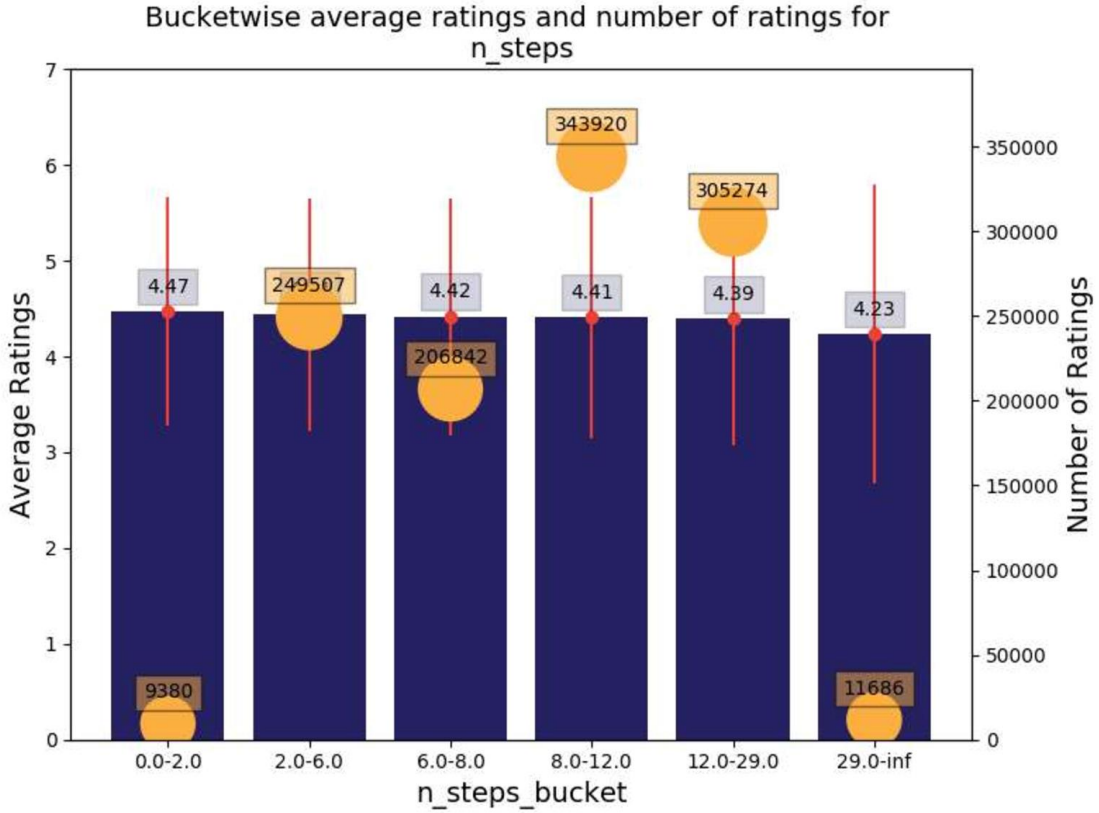
minutes

(Preparation Time)

Recipes with shorter preparation times tend to have higher average ratings compared to those with longer preparation times.



Exploratory Data Analysis(EDA)



n_steps

(Number of Steps)

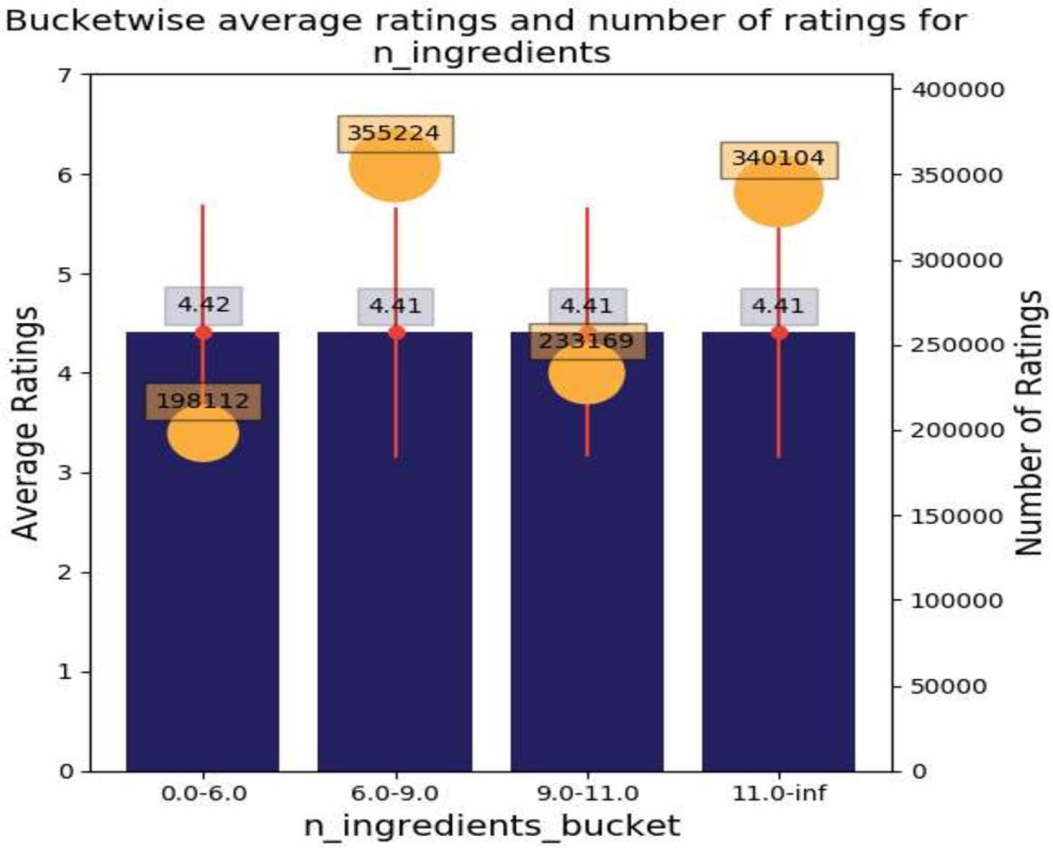
The 'n_steps' feature emerges as a clear determinant of recipe ratings. Recipes featuring fewer than 2 steps receive high ratings, while those with more than 29 steps are rated very low, indicating the strong relevance of this feature in determining average ratings.

Exploratory Data Analysis(EDA)

n_ingredients

(Number of Ingredients)

The 'n_ingredients' feature, representing the number of ingredients, exhibits relatively consistent average ratings across various ranges, indicating that it may not be an influential feature in determining recipe ratings.



Exploratory Data Analysis(EDA)

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

Exploratory Data Analysis(EDA)

individual_tag	avg_user_rating	n_user_ratings	n_recipes	in_percent_recipies	in_percent_interactions
preparation	4.4119124813277715	1123326	229318	0.9952779007491125	0.9970859455232471
time-to-make	4.414416558383976	1105132	224098	0.9726222407402585	0.98093659823417
course	4.412402044928726	1071920	217130	0.9423799727437654	0.9514569828574067
dietary	4.412032038984685	901277	163918	0.7114311259255401	0.7999909462821618
main-ingredient	4.424040070642098	864074	169549	0.7358705936477349	0.7669688418963456
easy	4.4183637556952755	630786	125789	0.5459449840715953	0.5598978882646952
occasion	4.4144829634028655	619666	113433	0.4923179083878024	0.5500275605822428
equipment	4.415547752950291	496985	69892	0.3033427948924941	0.4411335254733452
cuisine	4.416942151349161	478853	90639	0.39338819301580685	0.42503921058681404
low-in-something	4.414730950603082	445959	85258	0.37003376648177566	0.39584185817794815
main-dish	4.395996656937766	384079	71531	0.310456324922094	0.34091596995940915
60-minutes-or-less	4.405568569863525	343212	69929	0.30350338098834234	0.30464162810700074
number-of-servings	4.407139294746751	338857	58410	0.2535090232025208	0.3007760456378389
meat	4.408259712746521	319091	55769	0.2420466480907615	0.28323136065840054
taste-mood	4.412428615527087	310992	52060	0.2259489770231678	0.27604253117097416
north-american	4.413212293557913	283433	48182	0.20911781811237554	0.25158062823925603
30-minutes-or-less	4.4268528818028265	267003	55059	0.23896513111637718	0.23699704156455345
vegetables	4.454577657305231	259718	53562	0.23246790448165414	0.23053073426539286
oven	4.417805174050443	249669	30777	0.1335772505924325	0.22161104695595366
4-hours-or-less	4.383299863701983	247986	49450	0.21462114701874083	0.22011718351264725

Top n most rated tags

These tags have received the highest number of user ratings, indicating their popularity and relevance among users.

Exploratory Data Analysis(EDA)

Top rated tags

individual_tag	avg_user_rating	n_user_ratings	n_recipes	in_percent_recipies	in_percent_interactions
side-dishes-beans	5.0	2	2	8.680329505308021E-6	1.775238791807983E-6
cabbage	5.0	1	1	4.340164752654011E-6	8.876193959039915E-7
heirloom-historic...	5.0	3	2	8.680329505308021E-6	2.662858187711975E-6
middle-eastern-ma...	5.0	2	1	4.340164752654011E-6	1.775238791807983E-6
breakfast-potatoes	5.0	1	1	4.340164752654011E-6	8.876193959039915E-7

Tags with the highest average user rating, which stands at a perfect score of 5, appear to have received a comparatively lower number of user ratings.

Conclusion

- Through in-depth Exploratory Data Analysis (EDA), we have gained valuable insights into the factors shaping average ratings.
- Specifically, 'Review Time Since Submission,' 'Preparation Time,' and the 'Number of Steps' have been identified as pivotal elements profoundly affecting recipe ratings.
- Conversely, our analysis indicates that the ‘Nutrition’ columns and the 'Number of Ingredients' lack significant relevance in determining the average rating.
- These insights will be instrumental in model development and will aid strategic decision-making for the business.

THANK YOU