SUMMARY OF RECIPE RECOMMENDER

Problem Statement:

Step into the shoes of an ML engineer working at food.com. Your job is to design a recommender system to recommend recipes to users based on their choice and the current recipe they are looking at.

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 your site 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 your recipe site can generate.

Designing a recommender from scratch is a time-consuming task. In this assignment, you are expected to explore the data and create features that will be used to build the recommender.

Steps for Solution:

1. Task 1 – Reading the data.
   Checking rows and number of columns i.e, there are 231637 rows and 12 columns.
2. Task 2 - Extract individual features from the nutrition column.
   a. In this step we firstly used removed square brackets and later using string operation, regular-expression replace function
   b. Split the nutrition column using the comma delimiter into seven individual columns and cast the new columns to float values i.e; calories, total_fat_PDV, sugar_PDV, sodium_PDV, protein_PDV, saturated_fat_PDV, carbohydrates_PDV Then you can use a for loop to iterate over the column names declared in the variable to extract the value at a specific index of the nutrition array.

3. Task 03: Standardize the nutrition values
   In this task we need to standardize the nutrition columns using calories as the base of standardization and Convert the nutrition from absolute values to per 100 calorie values.
4. Task 04: Convert the tags column from a string to an array of strings
   In this task we consider tags column which is a string column but holds an array of strings, Remove [ ] ' punctuation marks from the tags column. Split the tags column based on the comma delimiter and convert the tags columns from a string to an array of strings.
5. TASK – 5: Creating new df = interaction_level_df by joining raw_recipes_df data set and raw_ratings_df data set using left join. There are total 1132367 – rows and columns – 30.
6. TASK – 6: Type casting a column to DateType(), finding the number of days between dates, days_since_submission_on_review_date.

PART – 2 EDA

Exploratory data analysis was conducted on data set, where it showed the file consisted recipes that were older than 6 yrs. There were total 1132367 reviews. It was found that short preparation time was preferred by many users.

PART – 3 FEATURE EXTRACTION

Below were the list of more features extracted

More 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