LDA + Classification using cluster probability

Description of the dataset:

The dataset contains the following features namely

'goal_id', 'goal_short_name', 'expert_assessment', 'user_id', 'preferred_locale', 'meal_id', 'meal_type', 'meal_title', 'meal_ingredients', 'expert_explanation', 'carbs_grams', 'protein_grams', 'fat_grams', 'fiber_grams', 'calories', 'carbs_RD_explanation', 'protein_RD_explanation', 'fat_RD_explanation', 'fiber_RD_explanation', 'calories_RD_explanation'

Our main focus is on the following features:
'goal_short_name', 'expert_assessment', 'meal_title', 'meal_ingredients'

goal_short_name: This feature is basically a goal that has been set by the user that they need to accomplish

The goals can be any of the following choices

Make 1/4 of my meal grains and starchy vegetables
Make 1/2 of my meal non-starchy vegetables
Make ¼ of my meal protein
Choose low fat foods
Choose foods without added sugar
Drink water with my meal
Reduce the portion size of my meal
Choose whole grain carbs
Choose whole fruits
Choose a variety of fruits and vegetables
Choose lean proteins
Choose plant proteins
Choose vegetable fats instead of animal fats
Choose non-starchy vegetables
Drink water instead of sugary beverages

> **Note:** For this problem, we will be focusing only on the goals that do not have any quantities (Eg: Choose low fat foods and not Make ¼ of my meal protein)

expert_assessment: This feature is basically a binary feature having the values {'not_really', 'yes'}. This is an assessment given by a physician whether a user goal has been met or not based on the description of the features {'meal_title', 'meal_ingredients'}. For this feature, ‘not_really’ indicates that the user has not met the set goal and ‘yes’ means that the user was able to meet the set goal.

meal_title: This is the tile of the meal that the user has eaten

meal_ingredients: This describes the ingredients that a particular meal eaten by the user contains

Problem Description:

Here we are tackling two of the following problems:

1. Perform topic modeling on meals so that meals are segregated into 6-7 topics
2. Use the probabilities of a meal belonging to cluster to map whether a user goal is met or not

Our approach:

Since our dependent variables are free form responses (text), we need to preprocess them in order to get a cleaner dataset. We perform the following pre-processing steps

Preprocessing

- Replace all punctuations
- Lower case all the words
- Perform spell spell correction
- Remove all stop words
- Perform lemmatization (Eg: Convert history, histories to histor)
- Replace the digits with blank spaces as digits do not hold any importance for this particular use case
- Convert the dependent variable (expert_assessment) to 1s or 0s

LDA

Here we create a new feature combining meal title and meal ingredients as this will tend to give more robust and trustworthy information about the meals. We then create a dictionary out of this new feature to train our LDA model. After playing around with several topics, 8 topics tend to give the best average coherence score. Hence, we create our LDA model on 8 topics

Classification using cluster

Once the LDA model is built, using ‘get_document_topics’ we can get the probability of each meal being in a particular cluster. Running this on every meal we are able to find the probability of each meal being in a particular cluster. These probabilities then form our new dataset with the probabilities being the independent variables and the expert assessment being the dependent variable. We then perform a classification task to see if these cluster probability assignments for every meal help in determining whether a particular user goal is met or not.