Project Proposal

1. Problem

In the ether of existential and absolute questions throughout mankind, one question maintains its consistency despite age, nationality, political inclination, or even IQ: What's for dinner?

Using features of recipes I aim to help clients decide on what kind recipe to cook--either for personal or more vocational reasons. Furthermore, my goal is to breakdown the factors of recipe popularity, i.e. perhaps quantifying the subjectivity of taste buds.

2. Who is the client and why do they care about the problem?

In terms of helping casual consumers of internet recipes, potential clients include anyone whose heart rate on their FitBit climbs up a few BPMs when thinking of the overwhelming amount of recipes on the Internet they need to sift through for a simple Mother's Day dessert. People who track their macronutrients can see what kinds of recipes are associated with a certain amount of protein or carbs.

Beyond the casual consumer, this model can benefit authors of culinary blogs, websites, or even recipe books who want to put out innovative recipes but are unsure of any trends in recipe popularity across certain categories.

Using the results of the model, these authors and content creators can predict the success of recipe type (quantified by user reviews) based on both nutritional and culinarily categorical factors. In this way, they can innovate or tweak existing recipes with a solid guideline as to what will be effective in receiving positive reviews.

3. Data & Data Acquisition

Dataset: "Epicurious - Recipes with Rating and Nutrition" from Kaggle

This dataset includes over 20,000 recipes with recipe rating, nutritional information, and assigned categories (680 features, including the city or state associated with the recipe, or whether it's for a particular holiday).

4. Solution Approach

a. Is this a supervised or unsupervised problem?

This is a supervised problem. There is historical data, through which I will approximate relationships between the observed inputs and output.

b. If supervised, is it a classification or regression problem?

This is a regression problem, as the relationship will aim to map the input to a continuous (rather than categorical) output.

c. Predicted Variable

This model aims to predict the reviewer score for a recipe.

d. Predictor Variables

The variables that will be used to predict the recipe score will include numerical factors such as Calories, Protein, Fat, and Sodium as well as categorical factors denoting what kind of meal it is (snack, brunch, vegan, alcoholic, etc.)

e. Training Data

The training data will be 80% of the total data set, leaving the remaining 20% for testing.

5. Deliverable

The deliverable for this capstone project will be a paper delineating the approach, results, and any relevant conclusions. Additionally, I will provide an Ipython notebook demonstrating the model construction and evaluation.