Using Machine Learning to Build a Scalable Tool to support Dietitians to Fight Chronic Diseases

Moumita Bhattacharya
Computer and Information Sciences
University of Delaware
Newark, Delaware
Email: moumitab@udel.edu

Debarati Roychowdhury
Computer and Information Sciences
University of Delaware
Newark, Delaware
Email: droyc@udel.edu

Michela Taufer
Computer and Information Sciences
University of Delaware
Newark, Delaware
Email: taufer@udel.edu

Keywords—Macro-nutrient, Spark, MapReduce, Feature Selection, Lasso Regression, Ridge Regression, ElasticNet.

I. Introduction

According to Centers of Disease Control and Prevention dietary and nutrition intake in the population of United States varies significantly based demographics of individuals. Nutritional status is an integral component of and has vital implications on the health of individuals. For instance, among children, nutritional status can affect growth, development, as well as the occurrence of nutrition-related health problems. Hence, identifying the associations between demographics, lifestyles, health characteristics, and dietary intakes is of growing interest among health care providers.

We aim to develop a framework using machine learning models that can assist dietitians to identify predictive patterns in nutrient intake, which can then help in fighting chronic diseases such as Obesity. Dietitians currently depend on behavioral changes and counseling methods to give nutrition advices. However, we plan to equip and assist dietitians by utilizing a data-driven approach and provide them information such as what factors are most indicative of fat or protein intake; what are the differentiating factors among people who intake more protein compared to those who intake more fats. Moreover, a prior knowledge of the demographic profile of an individual, will enable the dietitian to assess the kind of nutrients an individual would normally take.

We hypothesis that our approach can be extended to be used by dietitian in real-time to provide nutrient intake advice. We aim to identify the factors that are most informative of the five macro-nutrients. In an attempt to do so we first use simple linear regression where demographics features are used as predictors and one at a time among fat, protein, fibre and sugar is used as response variable. We then, use regularized linear regression, a combination of L1 and L2 regularization, namely *Elastic net*, for the task of feature selection. We present results obtained for predicting Fat, Carbohydrate and Protein for linear and regularized linear regression in this paper.

The rest of the paper is organized as follows: Section II describes the dataset used in this study, Section III describes the experimental setting, the data preprocessing steps, the feature selection, the regression methods and the evaluation methods used; Section IV presents and discusses the results; Section V summarizes our findings and proposes future work.

II. Dataset

Our dataset consists of demographics, vital signs and nutrient intakes gathered from several thousand individuals during nationwide surveys conducted by the National Center for Health Statistics and some other health agencies since 1971 [1]. The aim of the survey is to provide nationally representative information on nutritional status of the population and tracking changes over time. Specifically, we use subset of the dataset containing demographic information as well as macro-nutrient intake of individuals collected during 2013 and 2014. For each individual more than fifty demographic attributes are present in the dataset. We provide an individual's record-set as a example below:

Individual N <IDN,Fat,Age,Gender,Education Level, Income,. . . .,>

III. Method

In this section, we describe the framework we have developed to identify demographic factors that are highly informative of an individual's macro-nutrients intake. Figure 1 shows the developed framework.

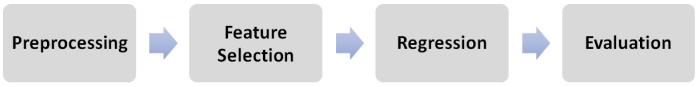


Figure 1: General Framework

Preprocessing: We preprocess and organize the data to obtain record-sets that can be used to train machine learning models. First, we remove four attributes that have a lot of missing values as well as the ones that have information only pertaining to children and old individuals. We then use MapReduce in Spark to obtain the sum of values of each of the five macro-nutrients for each individual and average the macro-nutrient intake values over two days.

Feature Selection and Regression: In order to identify the demographic attributes that are informative of the macro-nutrient intake from the dataset we conduct feature selection. Since the macro-nutrients are continuous numeric values we utilize regression methods to identify the relationship between the demographic attributes and the macro-nutrients intake. Specifically, we use a Regularized Linear Regression method which is a combination of LASSO and RIDGE regression, namely *ElasticNet*[2].

Linear Regression: Let Y denote the "dependent" variable whose values we wish to predict, and let $X_1, \ldots, X_k$ denote

the "independent" variables from which we want to predict Y.The equation for computing the predicted value of Y is:

$$y = c + \beta * x, \tag{1}$$

where y = estimated dependent, c = constant, β = regression coefficients, and x = independent variable.

This formula has the property that the prediction for Y is a straight-line function of each of the X variables, holding the others fixed, and the contributions of different X variables to the predictions are additive. The slopes of their individual straight-line relationships with Y are the constants b1, b2, ..., bk, the so-called coefficients of the variables. The additional constant b0 is the intercept.

Elastic Net: Suppose that the dataset has n observations with p predictors. Let $y = (y_i, ..., y_n)^T$ be the response and $X = (x_l ... X_p)$ be the model matrix, where $x_j = (x_j, ..., x_{nj})^T$, j = 1, ... , p, are the predictors. After a location and scale transformation, we can assume that the response is centred and the predictors are standardize. For any fixed non-negative λ_1 and λ_2

$$L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda_2|\beta|^2 + \lambda_1|\beta|, \tag{2}$$

$$|\beta|^2 = \sum_{j=1}^{p} \beta_j^2 \tag{3}$$

$$|\beta|_1 = \sum_{j=1}^{p} \beta_j \tag{4}$$

Evaluation: We empirically evaluate the performance of our models by plotting the total intake of macro-nutrients as a function of different factors identified as highly informative by our method. For instance, we plot macro-nutrients intake over discrete age groups and gender as shown in Fig. 2.

IV. Experiments and Results

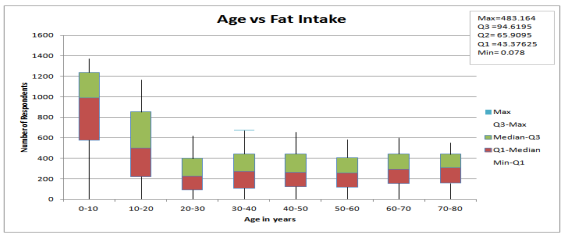
Experiments: First, we preprocess the dataset such that the variables which have more than 30 percent of values missing were removed. Then we create five different record-sets where each had one of the five macro-nutrient as the response variable (the value we want to predict). Next, we split each record-set into training and testing sets, where 75 percent of all the records are used as training data and 25 percent is used as testing data.Following which, we fit both linear regression and elastic-net regularized regression model to the training dataset.Then we predict the response variable's value on the test dataset and report the mean squared error value. We also report the coefficients of all the predictor variables in order to determine which variables are informative for the prediction task. Last, we report the variables that are most informative of each of the macro-nutrient and also present the observed difference among them. This analysis show how different macro-nutrient intake varies across individuals.

Results: From the experiments, we noted that certain demographic features affect some of the macro-nutrients while other demographic features affect other macro-nutrients. From Table 1, we see that demographic features such as Age, Spoken Language and Education Level affect the intake of four out of the five macro-nutrients, while Income affects the intake of all the macro-nutrients. Our results show that Gender is a highly informative feature for Protein and Sugar intake but not a clear indicator of Fat, Carbohydrate and Fiber intake.

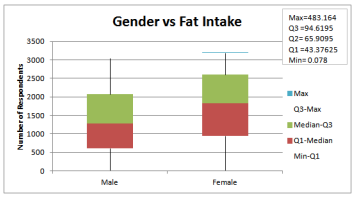
We empirically validate the above observation by plotting Fat intake for different Gender and Age ranges (See Figures 2) and observe that Fat intake indeed does not vary with respect to Gender. However, Fat intake significantly varies among different Age ranges.

	Fat	Carb	Protein	Fiber	Sugar
Age	X	X		X	X
Gender			X		X
Race			X	X	
Income	X	X	X	X	X
Education Level	X	X	X	X	
Country of Birth	X	X			X
Spoken Language		X	X	X	X
No. of People in HH		X	X	X	X
Pregnancy Status	X	X			X
Interpreter Used	X	X			X

Table I: Shows which demographic features impact the intake of macro-nutrient



(a) Variation of Fat intake with Age



(b) Variation of Fat intake with Gender

Figure 2: Shows that Fat intake varies significantly with Age but not with Gender

V. Conclusion and Future Work

We reported preliminary results obtained from our data-driven approach, using regularized linear regression to identify demographic factors highly informative of macro-nutrients intake. Our results indicate that there are many demographic factors that are indicative of all the macro-nutrient intakes. However, a few factors are only informative about a certain nutrient intake. We plan to implement the feature selection and regression models using the datasets from all the years available in the NHANES website.

References

[1] Centers for Disease Control and Prevention (2016). Nation Health and Nutrition Examination Survey. https://www.cdc.gov/nchs/nhanes/, last accessed 12/05/16.

[2] Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology).