# **DSCI- 5340**

# **PREDICTIVE ANALYTICS AND BUSINESS FORECASTING**

# **FINAL PROJECT**

# **PREDICTION OF CALORIES FROM NUTRITION DATA**

**INTRODUCTION:**

In our day to day lives, people always worry about the calorie intake they take, and it is very important for them to know the calories from the type of food they eat. Calories are units of energy that are available within food and absorbed by the body to maintain a healthy lifestyle. These are associated with nutrition values like protein, carbohydrates, fat, glucose etc. Each of the measures, energy will be available in the calories. Different type of foods will provide us with different amounts of calories. Additionally, other nutrients which are needed for the body also change according to the foods which we eat.

**MOTIVATION:** Daily diet is one of the prominent parts of our life. Intake of Nutrition and prediction of calories is something which everybody is very eager to know. This made us to consider the Nutrition dataset.

**OBJECTIVE:** The objective is to predict calories based on nutrition intake and various factors like carbohydrates, total fat, saturated acid, cholesterol, iron, protein, sodium etc. in this nutrition data set.

**LITERATURE REVIEW:**

Nutrients are expended through the nourishment that we eat, and through metabolic procedures in the stomach related framework these supplements are ingested at a cell level in the body (Gibney et al. 2009). Ideal sustenance adds to wellbeing, prosperity, ordinary turn of events, and high caliber of life (Gibney et al. 2009). In any case, malnutrition, overnutrition, and lack of healthy lifestyle are connected to imperfect wellbeing results (Gibney et al. 2009). Such less than stellar eating routines have been connected to the event of interminable maladies, including cardiovascular illness, Type-2 diabetes, malignancy, osteoporosis and frailty (Lytle et al. 2002). For instance, investigate reports that low admission of foods grown from the ground builds the hazard for creating malignancy (Steinmetz and Potter 1996), just as cardiovascular ailment (Hung et al. 2004), while low admission of dietary fiber has been connected to being overweight (Patrick et al. 2004).

The determinants of good dieting, the impact of sustenance on psychological wellness, sex contrasts in nourishment, and work environment programs are examined in this writing survey. Albeit no investigations which explicitly center around understudies' weight control plans have been distinguished, the writing will make inductions from related age gatherings and workplaces. Because of this lack of research, a few specialists have explicitly called for progressively centered sustenance mediations with youngsters (Walsh and Nelson 2010).

**DATASET:**

Data set considered is Nutrition Data with 8789 rows and 77 columns. Description of the columns:

Name – name of the food type

serving\_size- serving portion of food

calories- calories intake

total\_fat

saturated\_fat

cholesterol

sodium

vitamin\_a

vitamin\_b12

vitamin\_b6

vitamin\_c

vitamin\_d

vitamin\_e

vitamin\_k

calcium

copper

iron

magnesium

manganese

phosphorous

potassium

selenium

zinc

protein

carbohydrate

fiber

sugars

fructose

galactose

glucose

lactose

maltose

sucrose

fat

These values lead us to the calorie intake of a food item and to explore the important question such as “comparing calories intake with respect to the total fat, sodium, cholesterol, fiber, protein, etc.”

<https://www.kaggle.com/trolukovich/nutritional-values-for-common-foods-and-products>

**DATA CLEANING:**

The dataset at the beginning was largely left unaltered. It has many missing values, null values which needed to be replaced in order to make the dataset clean and prepared for the analysis.

All the attributes have units grams(g), milligrams(mg), mcg, IU after the values that needs to be removed. Missing values have been replaced by the mean value using “proc stdize”.

**REGRESSION ASSUMPTION CHECK:**

Normality assumption is checked by plotting Q-Q plot, it looks like the distant from the normality so using the log transformation again check then it shows normal but if we transform the data with square root transformation then the Q-Q plot looks normal.

![A close up of a map

Description automatically generated]()

**MULTI-COLLINEARITY CHECK:**

All variables need to be checked for the association between the independent variables.

H0: There is no association

H1: There is association between the variables

If the p- value is less than 0.05 then we can say that the correlation exists between the variables. These are called significant variables.

We must check for the VIF (Variance inflation factor) to detect the variables which are correlated.

Here we see that if Variance inflation factor greater than 10 then shows the cause of multi-collinearity. Now, we must remove those variables which are having VIF>10 in order to best fit the model.

If we use the only variable for which VIF less than 10 then we get the R-Square: 65%, so it seems that this model will not accurate as of able to more significant level prediction. Therefore, we must use the different variable selection criteria to find the most accurate model. We used Mallow’s CP method.

![A screenshot of a cell phone

Description automatically generated]()

Using this Mallow's Cp technique, we find a model which shows the 97 % accuracy with less variable. Also note that the variable comes from the VIF method almost cover in this model.

![A screenshot of a cell phone

Description automatically generated]()

Now we will use the forward regression technique of the variable selection.

**FORWARD SELECTION:**

Forward selection method gives the most accurate result with 97 % accuracy

It also uses the smaller number of variables.

One more important inference from the forward regression model is , it gives the parameter estimates with the significant p-value.

It starts with step 1 and by the end of step 35, we will have the model with the all variables which have significant p-value i.e. <0.05.

![A screenshot of a cell phone

Description automatically generated]()

This is followed till 35th step. R-square value increases gradually in the stepwise from step 1 to step 35. Summary of the stepwise selection criteria is given below.

A screenshot of a cell phone

Description automatically generatedA picture containing window, large

Description automatically generatedA picture containing white, room

Description automatically generated

**BACKWARD ELIMINATION:**

Now, we apply Backward elimination and observe the model to remove the variables which are not significantly contributing to the model.

It starts from Step 0 with all the variables and continued till step 29 where all the variables are left with p-values <0.05 and which are significant.

![A screenshot of a cell phone

Description automatically generated]()

![A close up of a piece of paper

Description automatically generated]()![A close up of a piece of paper

Description automatically generated]()

Summary of the backward elimination is :

![A close up of a device

Description automatically generated]()

Hence, we can predict the calories using the multi regression equation of final model is:

sqrt(Calories)=5.99481 + 0.28154\*fat + 0.17154\*lysine + 0.00383\*ash

-0.03564\*fiber+ 0.14748\*carbohydrate -0.00272\*magnesium

+ 0.31759\*alcohol+ 0.19382\*protein - 0.43567\*arginine

-0.00013339\*lucopene + 0.16832\*maltose +0.00083062\*fatty\_acids\_total\_trans -0.00083616\*folic\_acid + -0.00067491\*vitamin\_c + 0.01786\*vitamin\_e

-0.54111\*cholesterol + 0.51249\*saturated\_fatty\_acids + 0.05334\*fructose

-0.03754\*polyunsaturated\_fatty\_acids + 0.1148\*vitamin\_a -1.07405\*serine +

0.02369\*glutamic\_acid - 0.01763\*monounsaturated\_fatty\_acids -0.01221\*iron

+ 0.92567\*methionine + 0.23787\*proline + 0.01493\*zink

-0.00061001\*theobromine + 0.61509\*cystine + 0.40608\*histidine

-0.00020576\*carotene\_beta- 0.00021432\*vitamin\_d +

0.00003574\*cryptoxanthin\_beta + 0.09053\*aspartic\_acid

If we put the independent values in the model and you will find the square root calories so for the actual calories square the obtain value from the model.

**CONCLUSION:**

Based on these all models performed, we can conclude that calories can be predicted and compared to the different nutrients in the food which are factors that will be affecting the calories in any food item. But this also requires ongoing exploration as to find out for a greater number of food items and for different serving sizes, proportion. There is lot of scope in the future and all these findings will be useful.

**REFERENCES:**

Multiple Stepwise Regression Analysis on Knowledge Evaluation(Lan Yuqing, Guo Shuhang,)

<https://libproxy.library.unt.edu:2301/stamp/stamp.jsp?tp=&arnumber=4656643>

VIF Regression: A Fast Regression Algorithm for Large Data (Dongyu Lin, Dean P. Foster). Published in

2009 Ninth IEEE International Conference on Data Mining.

<https://libproxy.library.unt.edu:2301/stamp/stamp.jsp?tp=&arnumber=5360322>

Predictive Data Mining in Nutrition Therapy( Diana Ferreira, Hugo Peixoto, José Machado and António Abelha) .Published in 2018 13th APCA International Conference on Automatic Control and Soft Computing (CONTROLO) June 4-6, 2018, Ponta Delgada, Azores, Portugal

<https://libproxy.library.unt.edu:2301/stamp/stamp.jsp?tp=&arnumber=8516413>

The importance of nutrition and physical activity in young people increased quality of life.(Sînziana-Călina Siliúteanu,Mihai ovaúă) Published in The 5th IEEE International Conference on E-Health and Bioengineering - EHB 2015.

<https://libproxy.library.unt.edu:2301/stamp/stamp.jsp?tp=&arnumber=7391476>

Approximate Estimation of the Nutritions of Consumed Food by Deep Learning( øbrahim Berkan AYDøLEK)

<https://libproxy.library.unt.edu:2301/stamp/stamp.jsp?tp=&arnumber=8093588>

http://www.nutritionaustralia.org/sites/default/files/Diet%20and%20nutrition%20chapter%20from%20Apprentices%20-%20young%20people%20