Heinz 95-845: How Food Can Help Us On Maintaining Healthy Cholesterol Level

Mengfei Lyu MLYU/mlyu@andrew.cmu.edu
Heinz College
Carnegie Mellon University
Pittsburgh, PA, United States

Steve Sroba SSROBA/ssroba@andrew.cmu.edu
Heinz College
Carnegie Mellon University
Pittsburgh, PA, United States

Abstract

Human body needs healthy level of cholesterol to function, and around 25 percent of our body's cholesterol comes from the daily foods. It's possible for people to adjust their dietary habit to maintain a normal level of cholesterol. Previous studies give advices on consuming certain foods to prevent high level of LDL cholesterol but not much details on ingredients. In this paper, I used dataset collected by National Health and Nutrition Examination Survey[NHA]to investigate the relationship between dietary habit and cholesterol level and identify ingredients taken from food which are correlated with high cholesterol level. Data from different surveys are combined to represent the integrated information of dietary habit and physical situation, and multiple machine learning classification models are constructed to identify individuals with high probability to develop high level of LDL cholesterol given the survey information. Model evaluation is conducted to assess performance of candidate machine learning classification algorithms including Support Vector Machines, Logistic Regression and Decision Tree. In addition, I analyzed trained weight of different features to investigate correlation of dietary habit and cholesterol level and provide dietary suggestion on maintaining normal cholesterol level with details on food ingredients.

1. Introduction

Cholesterol is a fatty substance made by the liver, and Human bodies use cholesterol to make vitamin D and hormones as well as bile acids.Cholesterol is distributed throughout the body by lipoproteins in bloodstream, and there are two kinds of lipoproteins we can use to transport cholesterol which are Low-density lipoproteins(LDL) and High-density lipoproteins(HDL). LDL carries cholesterol from liver to our body in the form of particles, and if there is too much LDL cholesterol in the blood, these particles will form deposits and narrow our arteries. This situation will have harmful effect on our body including chest pain, heart attack, numbness in the legs and blockage in the brain [Pietrangelo], and will increase the risk of coronary heart disease.

According to Pietrangelo, around 25 percent of our body's cholesterol comes from the foods we eat every day. Some kinds of foods, including sugar, cream, shellfish and chocolate, will boost the cholesterol level of our body which will lead to high cholesterol. Hence, changing dietary habit can lower cholesterol level, and it consists of two strategies: add foods that lower LDL cholesterol and reduce foods that boost LDL cholesterol level [Harvard]. Previous studies give advices on food selection to maintain healthy LDL cholesterol level.

Harvard Medical School published an article in 2015 [Harvard] to point out 11 foods that can lower cholesterol including oats, nuts, soy and fatty fish. However not much details are provided to uncover which ingredients within food actually boost LDL cholesterol and help to lower LDL cholesterol except soluble fiber. This paper aims to investigate correlation of different food ingredients and LDL cholesterol level within human body given details of demographic information, and provide integrated dietary suggestion on combination of different food ingredients to help maintain healthy LDL cholesterol level.

This paper has the following six parts. In section 2, background knowledge of machine learning approaches employed in the project is provided. In section 3, I present project pipeline and different tools I used to conduct different tasks of this project. In section 4, details of the project pipeline will be provided including NHANES survey information, data extraction and transformation, feature choices, classification models and evaluation. In addition, approaches to interpret the trained weight on different features of constructed classification model will be described. In section 5, analytical result of the different experiences will be discussed. And discussion, related work and conclusion will be presented in section 6 and section 7.

2. Background

Machine learning is a modern approach of data analysis to discover hidden insights represented by patterns within large dataset. It uses different algorithms to automate the process of analytical model building and different approaches can be applied to improve the performance of algorithms. A major task of machine learning is classification which is the main objective of this paper to classify individuals with high cholesterol level.

Common algorithms for binary classification includes Support Vector Machine, Logistic Regression and decision tree, and these algorithms compute iteratively to find a decision rule to separate data points of different classes. Support Vector Machine showned as Figure 1 aims to maximize the minimum distance from data points to decision boundary to reduce probability of misclassification. Logistic Regression computes the probability of data points belong to positive class by assign different weights to all the features. Data point with over 50% probability will be classified as positive class. Decision tree showned as Figure 2defines a set of consecutive decision rules to classify data point given values of different features.

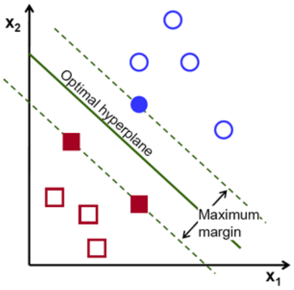

Figure 1: Support Vector Machine

Both Support Vector Machine and Logistic Regression conduct classification with underlying linear algebra computation, hence the data used as model input should be numeric

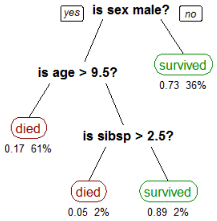

Figure 2: Decision Tree

data representing different features of data point. And regularization on feature weight can be applied to reduce the complexity of constructed model. Grid search is the approach to train models with different combination of candidate parameters and choose the parameters with best performance of machine learning model.

Accuracy of machine learning model means the ratio of correct classified data out of total number of data points, and it can be used to measure the performance of machine learning models. However, accuracy will not be a proper measurement for model evaluation when dataset has unbalanced class distribution. For example, if the size of data points with negative class is 10 times bigger than the size of data points with positive class, then even the model classifies all data points as negative, the accuracy will still be as high as 90%. When the case is to detect high cholesterol, the cost of inability to detect potential health risk will be huge. Alternately, sensitivity will be used to evaluate the classification model for this project, and it measures the ratio of correctly positive classification out of number of total data points in positive class. In addition, adjusted class weight will be used as a model parameter to enable the machine learning algorithm assign different weights to different classes.

3. Support Vector Machine

The pipeline of the project consists of three main part: data processing and exploration, classification model construction and analytical result inference. The structure of the pipeline is shown as Figure3.

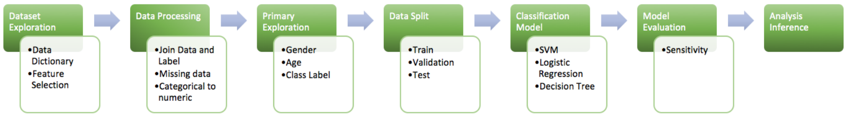

Figure 3: Project Pipeline

The data processing part is finished by R. I explored the whole data dictionary of NHANES and chose relevant data fields to the project. The I used NHANES's library for R RNHANES[RNH] to get index of different data files and download required data set. Then data processing is conducted including join data frames to integrate information for same respondent, select chosen features, label respondent class according to recommended cholesterol levels by age from National Heart, Lung, and Blood Institute[Goldman], deal with missing data and lastly convert categorical features to numeric ones. The processed data is exported to a text file for model construction.

The machine learning model is constructed by Python. Full dataset is split to three parts: training set, validation set and testing set. Then a SVM model is trained on the training using python's machine learning package sklearn. While the package contains build-in class and methods to train SVM model, grid search is also conducted to confirm the best kernel for SVM model. The class_weight parameter is set to 'balanced' for the model to automatically assign different weights to classes according to class distribution. And recall_score method of sklearn is used to calculate the sensitivity of training model. A ROC curve is also plotted to assess the performance of SVM model.

For result inference part, the model's trained coefficients are investigated to analyze correlation between different features and classification outcome, and specific dietary recommendation is yielded to maintain healthy LDL cholesterol level.

The R code and Python code are contained in two files, and they are available at `https://github.com/MavisLyu/95845FinalProject.git`

4. Experimental Setup

In this section, I will provide details of each step for the project and how I finalize the Support Vector Machine as the machine learning model to conduct classification.

4.1 NHANES

This project used dataset from National Health and Nutrition Examination Survey(NHANES). The NHANES is a program of variety of studies to assess the health and nutritional status of adults and children in the US. The survey starts from early 1960s and it combines questionnaire interviews with physical and laboratory examinations. The survey examines about 5,000 persons each year with different healthcare topics including demographics, dietary information, physical examination and laboratory tests. The population of the survey is selected to represent the U.S. population of all ages located in counties across the country. However, to produce reliable statistics, NHANES over-samples persons 80 and older, hence the analytical result of the project provides more profound insight for people older than 80.

4.2 Data Extraction

This project uses 5 NHANES datasets from 2013-2014 cycle: Demographics Data, Dietary Data, Examination Data, Laboratory Data and Questionnaire Data.

SEQN from each data file represents unique ids of respondents, and 5 downloaded data file are merged by the value of SEQN. Each record contains laboratory examination result of LDL cholesterol level, and labels for each record is created under the reference of recommended cholesterol levels by age from National Heart, Lung, and Blood Institute.

All chosen features then are selected from the merged dataset and renamed. Missing data rate is checked for each feature, and two features with missing rate over 30% are filtered out because there is huge information loss within these features. The original LDL cholesterol measure is also dropped as well as respondent id on the purpose of classification.

After examining the features with missing data, around 10% records contain missing data about dietary information indicating they didn't take the corresponding interview. Hence these records are also filtered out. For the left missing data, imputation is conducted to simulated the data pattern and replace possible values for missing data.

In addition, full dataset is split into three parts for machine learning model construction: training data, validation data and testing data.

4.3 Feature Choices

Features are chosen after carefully exploring the full data dictionary of the survey to represent all relevant information for the target outcome. Chosen features represents total nutrient intakes from sample day(e.g. dietary sample weight, salt and water intakes, vitamin and calcium intakes as well as seafood and sugar intakes), alcohol use, demographics(e.g. gender, age, weight and BMI) and cholesterol examination history(e.g. high cholesterol level diagnosis and prescribed medicine treatment).

For categorical features including gender, use of sugar and cream, alcohol use, supplement use, frequency of salt intake and whether respondent is on diet, one-hot encoding features are created to represent these information with numeric features.

4.4 Primary Exploration

Primary data exploration is conducted on processed data through data visualization.

The distribution of binary classes is shown as Figure4. From the visualization, the ratio of positive data points is much lower than the negative data points. This indicates that true positive rate is a more proper model evaluation measurement than accuracy because even the model classifies all data points to negative, it can still achieve a high accuracy.

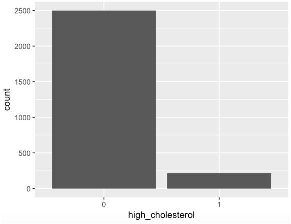

Figure 4: Label Distribution

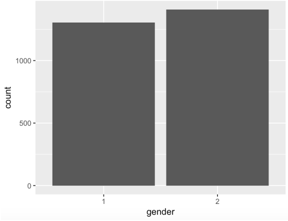

Figure 5: Gender Distribution

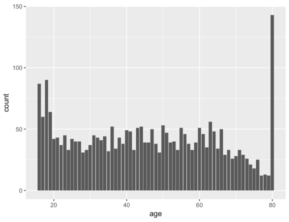

Figure 6: Age Distribution

The distribution of gender is shown as Figure5. From the visualization, the distribution of gender within sample population is approximately balanced. The distribution of age is shown as Figure6. The distribution of age within sample population has a higher ratio of people older than 80 corresponding to the oversampling approach of NHANES. Further discussions can be found in Discussion section.

4.5 Comparison Methods

The model is constructed by comparing three different classification algorithms: Support Vector Machine, Logistic Regression and Decision Tree. Parameters used to construct models are tuned first, then model is trained using training data set. Validation data is also used to assess the performance of training model.

For Support Vector Machine, grid search is applied to confirm the kernel with best performance and for this dataset is linear kernel. For Logistic Regression, L2 regularization is used with maximum 100 iterations. Both of the models use balanced class weight parameter to train the model. For Decision Tree, specifically Gradient Boosting Tree, grid search is applied to confirm parameters including learning rate and max depth of tree with best performance.

4.6 Evaluation Criteria

The outcome of the project aims to detect risk of high LDL cholesterol, and the cost of inability of detect potential risk is greater than misclassify negative classes. And with the unbalanced distribution of binary classes, sensitivity of the model is used to evaluate different machine learning models.

The sensitivity of different models are shown as Table 4.6. And for this project, Support Vector Machine with linear kernel performs the best out of three candidate models.

Model	Sensitivity
SVM	0.76
Logistic Regression	0.48
GradientBoostingTree	0

Table 1: Model Evaluation

5. Results

After examination of weights assigned to all features by the trained classification model, two kinds of features yield different correlations with LDL cholesterol level.

5.1 Demographic Features

The trained weight of demographic features is shown as Table 5.1. The trained model indicates that males are more likely to have high LDL cholesterol than females, and higher BMI is correlated to higher probability of high LDL cholesterol level. The assigned weight of age is 30.33 which means older people are more likely to have high LDL cholesterol than younger one.

Features	Weight
male	14.76
female	-14.8
weight	1.6
height	-20.9
BMI	6.33
age	30.33

Table 2: Assigned Weight of Demographic Features

5.2 Dietary features

Part of the trained weight of dietary features is shown as Table5.2. The trained model indicates people on diet and taking vitamin supplements are less likely to have high LDL cholesterol. And if salt is added in cooking very often, it correlates with higher probability for a person to have high LDL cholesterol. For specific food ingredients, people having a dietary habit with high volume of protein, vitamin E, retinol, vitamin B1, vitamin B2 and copper are less likely to develop high level of LDL cholesterol, while a dietary habit with high volume of niacin, vitamin B6, vitamin D, iron, zinc and vitamin A is correlated with higher probability of high level of LDL cholesterol.

Features	Weight	Features	Weight
supplement	-6.75	salt often	19.69
on diet	-17.8	no diet	19.42
protein	-19	VE	-35
retinol	-63	VB1	-33.54
VB2	-35.53	niacin	18.56
VB6	72.82	VD	52.86
iron	47.92	zinc	6.88
copper	-9.1	VA	61.7

Table 3: Assigned Weight of Dietary Features

6. Discussion

This project constructed a machine learning classification model to investigate the correlation of dietary features and LDL cholesterol level. Suggestion on dietary habit is provided to help people maintain healthy level of LDL cholesterol based on the analysis of constructed model. And this model can also be used to predict the probability of high LDL cholesterol level given the demographic and dietary information of an individual.

Several limitations of the project also exist. Due to the data quality and limited number of respondents of NHANES, the processed data only contains around 2800 records. It's possible that certain pattern of the data is not captured and represented by available dataset due to relatively small data volume. In addition, the trained weight of different features from the classification model only reveals the correlation of dietary habit to LDL cholesterol level instead of the reason that why certain food ingredients are presented with corresponding LDL cholesterol level. Further investigation can be conducted to explore whether certain food ingredients reduce/increase the LDL cholesterol level or other ingredients contained in the same food have the effect to lower/boosting LDL cholesterol level.

There are further discussions regarding the model construction process. From the primary data exploration about distribution of gender and age on sample population indicates that the analysis can provide a reasonable insight to both gender of population and more profound insight to elder people within the whole population. And generally gradient boosting tree is expected to perform better than the result of this project. A possible reason behind this is that normally to construct a gradient boosting tree requires more information on complex patterns hidden in the data than the dataset I used for this project.

7. Conclusion

This paper describes the motivation, methodology and experimental result of the proposed project to investigate correlation between dietary habit and LDL cholesterol level.

The dataset from National Health and Nutrition Examination Survey is used to construct a machine learning classification model, and analytical result of the model is translated to dietary suggestion at the level of food ingredient to help people maintain normal LDL cholesterol level. While previous studies provide advices on which kinds of food helps to lower cholesterol level, this paper aims to give more details on dietary habit adjustment, and draw more attention of other scholars to investigate more approaches on adjusting dietary habit to maintain healthy physical condition.

References

Nhanes. URL https://www.cdc.gov/nchs/nhanes/about_nhanes.htm.

Introduction to rnhanes. URL https://cran.r-project.org/web/packages/RNHANES/vignettes/introduction.html.

Rena Goldman. What are the recommended cholesterol levels by age? URL http://www.healthline.com/health/high-cholesterol/levels-by-age?s_con_rec=false&r=00#Inadults4.

Harvard Heart Letter Harvard. 11 foods that lower cholesterol. URL http://www.health.harvard.edu/heart-health/11-foods-that-lower-cholesterol.

Ann Pietrangelo. The effects of high cholesterol on the body. URL http://www.healthline.com/health/cholesterol/effects-on-body.