Predicting Obesity with Statistical Learning Methods

ST4248 Term Paper

Done By:

A0225737Y

Executive Summary

In this term paper, we aim to predict the likelihood of an individual developing obesity given a set of clinical and non-clinical variables. Particularly, we are interested in finding out what non-clinical variables contribute to the risk of obesity. We shall build both classification and tree-based models to investigate this problem. From our results, the top 3 variables which contribute to the risk of obesity are age, a family history of being overweight and frequent consumption of food in between meals. Further analysis is also done to explain variability in model performance.

1. Introduction

1.1 Background Information

Obesity is defined as an excessive accumulation of fat, and it presents a risk to numerous chronic diseases such as heart disease or stroke. It is also linked to a number of cancers, and can lead to musculoskeletal disorders such as osteoarthritis (World Health Organization: WHO, 2020). Fortunately, many of the causes of obesity are reversible with a healthier diet and regular exercise. Our goal in this paper is to identify risk factors associated with obesity, and to predict the incidence of obesity.

1.2 Problem Statement

What are the non-clinical variables that contribute to the risk of obesity?

In this paper, we shall define clinical variables as age, sex, height and weight. Non-clinical variables are typically lifestyle choices, such as frequent consumption of alcohol. We would also like to build statistical models that can possibly classify whether an individual is obese. We intend to use both classification and tree-based methods.

2. Data Set

The dataset is obtained from the UCI Machine Learning Repository. It contains 17 attributes and 2111 observations. 23% of the data was collected directly from users from Mexico, Peru and Columbia through an online survey, while the remaining 77% of the data was synthetically generated through SMOTE. The 17 attributes contain clinical variables such as *gender, age, height and weight*. Other note-worthy variables include *favc* (frequency of consumption of high-caloric food) and *faf* (frequency of physical activity). A full list of the variables and their description can be found in Appendix A.

3. Data Cleaning

3.1 Feature Engineering

First, we noted that there are no missing values in our dataset. The column *weightclass* classifies an individual into 7 different levels, ranging from “Insufficient_Weight” to “Obesity_Type_III”. Since we are interested in simply predicting whether an individual is obese, we shall create a new variable, *obesity,* for if an individual is obese (regardless of obesity type). This helps us set up a two-class problem.

3.2 Variable Transformation

Since a portion of the data was synthetically generated, certain categorical variables such as *fcvc* and *ncp,* which are supposed to be integer encoded, contain trailing decimal values. We rectified this problem by rounding the variables.

3.3 Removal of Erroneous Variables

BMI is the ratio of an individual's weight (in kg) to the square of their height (in m). An individual with a BMI of 30 or more is considered obese (Palechor & De La Hoz Manotas, 2019). We calculated the BMI of each individual using their height and weight records, and verified that each individual is correctly denoted in the *obesity* variable we created earlier. Erroneous records are subsequently removed. We are left with 2105 observations.

3.4 Removal of Collinear and Unnecessary Variables

We removed the variables *height* and *weight*, as they are components to calculating BMI and thus, obesity. We also removed the variable *smoke,* as there are very few positive incidences of smoking denoted in our dataset (approximately 2% of all records), hence it will not contribute to the predictive power of our models. Lastly, we removed the variable *weightclass,* with the intention of using *obesity* as our response.

4. Exploratory Data Analysis (EDA)[1]

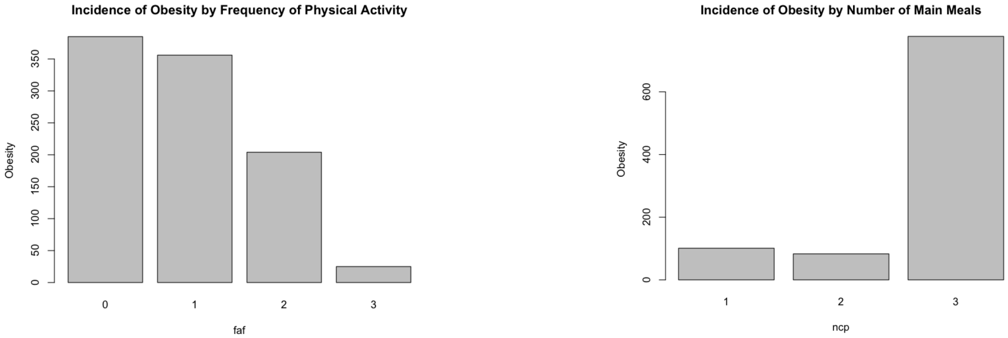

Figure 1: Bar Chart of Obesity against Frequency of Physical Activity (left) and Number of Main Meals (right)

Figure 1 shows some of our more noteworthy EDA graphs. It appears that as the frequency of physical activity decreases, and the number of main meals increases, incidence of obesity increases.

5. Methodology

For our investigation, we employed two main statistical learning methods: classification and tree-based methods. We shall use *obesity* as our response variable. We noted that the classes are relatively balanced, with 46% of our data denoted as positive incidences of obesity. We shall use accuracy as our metric to evaluate model performance.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Number\ of\ Observations}$$

To tune our models, we employed 10-fold cross-validation (CV), selecting the best set of hyperparameters that give us the highest CV accuracy.

[1] More EDA Graphs can be found in Appendix B.

5.1 Classification

First, we implemented two classification methods: logistic regression (LR) and support vector machines (SVM). We noted that our categorical variables are integer encoded with different scales. Since SVM is scale-sensitive, we scaled our variables. This ensures that the SVM margin will not be dominated by variables with larger scales.

LR			**SVM**		
Hyperparameter	Values	Optimal	Hyperparameter	Values	Optimal
threshold	[0.4, 0.5, 0.6, 0.7, 0.8]	0.4	*cost*	[0.01, 0.1, 1, 5, 10, 50]	10
			gamma	[0.1, 0.5, 1, 3, 5]	0.5
			kernel	[linear, radial, polynomial]	radial
LR CV Accuracy		84.6%	SVM CV Accuracy		90.2%

Table 1: Range of Hyperparamaters for LR and SVM

From Table 1, the optimal choice of *threshold* for LR is 0.4. The optimal choice of cost and gamma for the SVM is 10 and 0.5 respectively, which suggests that while the margin is relatively large, a less flexible classifier was chosen to minimise overfitting (Yıldırım, 2021). When tuning the ‘polynomial’ kernel, the range of values of degree we used is 1:5. Ultimately, a radial kernel was chosen, which suggests that our decision boundary is highly non-linear.

5. 2 Tree-Based Methods

Next, we implemented two tree-based methods: random forest and extreme gradient boosting (XGBoost). Note that for tree-based methods, scaling a feature does not affect the position in which a feature is split.

Random Forest			**XGBoost**		
Hyperparameter	Values	Optimal	Hyperparameter	Values	Optimal
mtry	[1:14]	4	*nrounds*	[50, 100, 150]	100
			max_depth	[3, 5, 7]	7
			eta	[0.1, 0.3, 0.5]	0.5
			gamma	[0, 1, 2]	1
			colsample_bytree	[0.4, 0.6, 0.8]	0.6
			subsample	[0.8, 1]	1
			min_child_weight	[0.5, 1, 1.5]	1
Random Forest CV Accuracy		92.9%	XGBoost CV Accuracy		91.9%

Table 2: Range of Hyperparameters for Random Forest and XGBoost

From Table 2, the optimal choice of *mtry* for random forest is 4, which suggests that CV attempted to find a model within the sweet spot in the bias-variance tradeoff. Unlike earlier models, we tuned the hyperparameters[2] in the XGBoost model using GridSearchCV.

6. Results

6.1 Model Evaluation

We evaluated our models on our test set by computing test accuracy. A higher test accuracy corresponds to a greater predictive power. The results of our models are as follows.

Methods	**Models**	**Test Accuracy**
Tree-Based	XGBoost	92.4%
	Random Forest	92.2%
Classification	SVM	81.3%
	LR	75.8%

Table 3: Test Accuracy for each model

From Table 3, we can see that XGBoost performs the best on our test set, slightly outperforming random forest with a high test accuracy of 92.4%. It appears that in general, tree-based methods performed better compared to classification methods.

6.2 Feature Importance

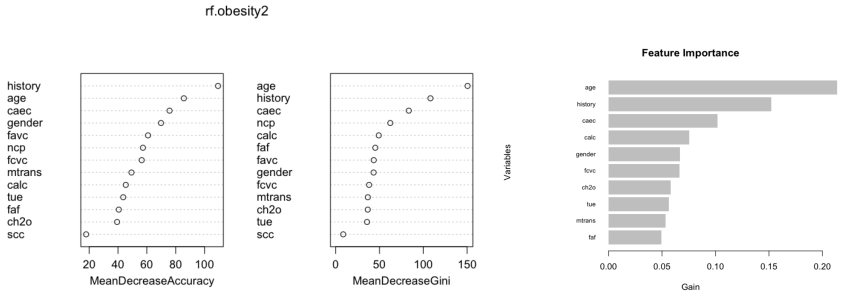

Figure 2: Random Forest (left) and XGBoost (right) Feature Importance plots

From Figure 2, both random forest and XGBoost agree that the top 3 most important variables in predicting obesity are *age, history* and *caec.* This implies that an individual of a higher age, with a family history of being overweight and frequent consumption of food between meals has a higher likelihood of being obese. Among these variables, *history* and *caec* are non-clinical variables.

[2] Full Description of XGBoost hyperparameters can be found in Appendix C.

7. Discussion

7.1 Analysis of XGBoost

XGBoost is an ensemble tree method which uses the gradient descent architecture to combine weak learners to form a stronger model (Morde, 2021). The implementation of XGBoost comes with a vast number of hyperparameters we can vary for greater accuracy, contributing to its high performance. Additionally, boosting is generally a robust method which is able to curb overfitting easily.

7.2 Analysis of Logistic Regression

We have seen earlier in tuning the choice of kernel for SVM, a highly non-linear decision boundary (radial) was chosen. This suggests that the relationship between the predictors and our response variable is complex, which is perhaps the reason why logistic regression does not perform well on our dataset. However, the ease of interpretability of the logistic regression model still makes it a popular statistical learning tool (Lekhtman, 2021).

7.3 Limitations

This paper aims to simply predict whether an individual is obese. In reality, there are different categories of obesity/overweight an individual can belong to, which is reflected in our original dataset. Each category presents a different level of risk of chronic ailments, with type 3 obesity being the most severe. Hence, moving forward, it will be useful to look into predicting these specific categories, i.e, a multi-class problem. We can also consider a regression approach, by predicting an individual's BMI. Lastly, we can look into using recall as our evaluation metric, as recall emphasises on reducing the number of false negatives produced. False negatives are dire as individuals may fail to realise they are at risk of becoming obese and developing its associated ailments.

8. Conclusion

We have identified the most important non-clinical variables which contribute to the risk of obesity: a family history of being overweight and frequent consumption of food between meals. Surprisingly, the frequency of physical activity and number of main meals are not as important variables, despite our EDA plots. While the accuracy produced by the models is satisfactory, the dataset is only reflective of individuals in Latin American countries. Different cultures and policies may affect the risk of obesity, a phenomenon we cannot observe with our dataset. In spite of this, and the limitations outlined above, the statistical learning methods we employed helped us to uncover insights into what increases the risk of obesity. While we cannot possibly change our family medical history, we can make changes to our lifestyle, that is, to limit our consumption of food between meals.

References

1. World Health Organization: WHO. (2020). Obesity. *www.who.int*. https://www.who.int/health-topics/obesity#tab=tab_1.
2. Palechor, F. M., & De La Hoz Manotas, A. K. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data in Brief*, *25*, 104344. https://doi.org/10.1016/j.dib.2019.104344
3. Yıldırım, S. (2021, December 16). SVM Hyperparameters Explained with Visualizations - Towards Data Science. *Medium*. https://towardsdatascience.com/svm-hyperparameters-explained-with-visualizations-143e48cb701b#:~:text=Gamma%20is%20a%20hyperparameter%20used,of%20a%20single%20training%20point.
4. Morde, V. (2021, December 9). XGBoost Algorithm: Long May She Reign! - Towards Data Science. *Medium*. https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d.
5. Lekhtman, A. (2021, December 31). When logistic regression simply doesn't work - Towards Data Science. *Medium*. https://towardsdatascience.com/when-logistic-regression-simply-doesnt-work-8cd8f2f9d997#:~:text=The%20reason%20is%20that%20the,even%20on%20the%20training%20data).

Appendices

Appendix A

Variable[3]	**Description**
gender	Gender.
age	Age.
height	Height in metres.
weight	Weight in kilograms.
history	Presence of a family member who suffered from/is overweight.
favc	Frequency of consumption of high-caloric foods.
fcvc	Frequency of consumption of vegetables.
ncp	Number of main meals.
caec	Frequency of consumption of food between meals.
smoke	Whether an individual smokes.
ch2o	Consumption of water daily.
scc	Whether an individual monitors their caloric consumption.
faf	Frequency of physical activity.
tue	Time spent using technology devices.
calc	Frequency of consumption of alcohol.
mtrans	Mode of transportation.
weightclass	Weight Class: Insufficient, Normal, Overweight 1-2, Obesity 1-3.

Table 4: Description of each variable in original dataset

[3] All variables were renamed to lower case for ease of programming. The variables *weightclass* and *history* were also renamed; they were originally *NObeyesdad* and *family_history_with_overweight* respectively.

Appendix B

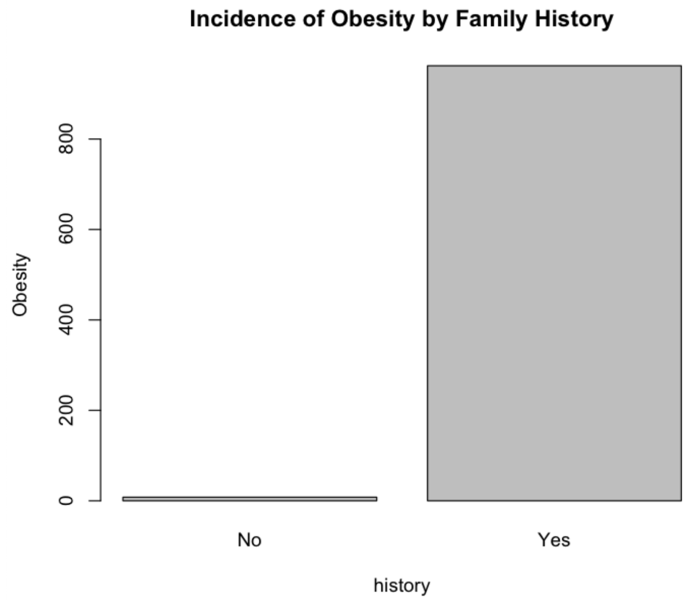

Figure 3: Bar Chart of Incidence of Obesity against Family History with Overweight

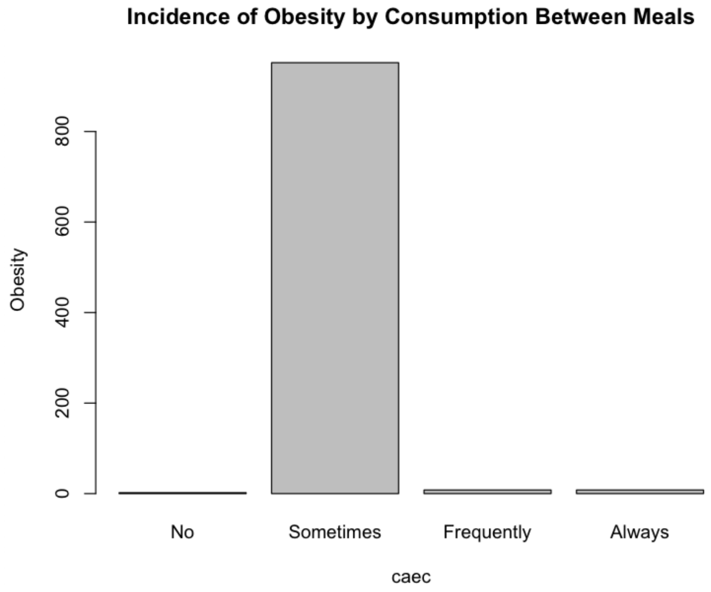

Figure 4: Bar Chart of Incidence of Obesity against Frequency of consumption between meals

Appendix C

XGBoost Hyperparameter	**Description**
nrounds	Maximum number of boosting iterations.
max_depth	Maximum depth of tree.
eta	Learning rate: scale the contribution of each tree by a factor of 0 < eta < 1 when it is added to the current approximation.
gamma	Minimum loss reduction required to make a further partition on a leaf node of the tree.
colsample_bytree	Subsample ratio of columns when constructing each tree.
subsample	Subsample ratio of the training instance.
min_child_weight	Minimum sum of instance weight (hessian) needed in a child. If the tree partition step results in a leaf node with the sum of instance weight less than min child weight, then the building process will give up further partitioning.

Table 5: Description of Hyperparameters tuned for XGBoost