Stats 101C Final Project

Predictive Analysis of Obesity Diagnostic

Leo Cardozo, Allison Chen, Clajerson Gimena, Faith Satrya, Naina Sharma

Table of Contents

Abstract

The project aims to predict one's obesity status based on provided clinical characteristics with classification models. The report summarizes our process from exploring the dataset and imputing missing values to selecting the optimal model and refining model parameters. The final classification model is an XGBoost model that uses 19 of 29 total predictors given in the training dataset. The predictors include age, CH20, height, NCP, FAF, FCVC, race, TUE, gender, CAEC, CALC, MTRANS, family_history_with_overweight, FAVC, SCC, SMOKE, ever_married, cholesterol, and MaxHR. The model has an accuracy of 99.59% and is ranked 3rd on the Kaggle leaderboard.

1. Introduction

Obesity refers to a chronic condition characterized by excessive fat deposits which can not only impair one's health but also negatively affect one's quality of living. Statistics show that as recently as 2022, 890 million adults worldwide were living with obesity. This accounts for a portion of the 2.5 billion overweight adults worldwide in the same year. Obesity continues to be one of the most prevalent health concerns worldwide, with the statistics on worldwide adult obesity having more than doubled since 1990. Obesity increases the risk of type 2 diabetes, heart disease, and certain cancers. It also negatively affects bone health and reproduction while impacting sleep and movement (World Health Organization, 2024).

The objective of this project was to predict an individual's obesity status ("obese" or "not obese") using a comprehensive dataset collected from multiple sources. The dataset contains 42,686 observations, with each observation representing an individual, and includes 29 variables encompassing demographic information, lifestyle factors, and key health indicators. These

variables comprise both numerical and categorical data. Numerical variables include metrics such as cholesterol levels, while categorical variables capture factors like methods of transportation, smoking status, and socio-demographic characteristics. To achieve the project's objectives, key predictors were identified and used to develop a robust classification model designed to effectively assess obesity risk. This approach aims to support the global effort to combat obesity by facilitating early interventions and enabling personalized health recommendations.

2. Data Analysis

a. Data Background

The dataset, comprising 42,686 observations, was split into two subsets: 32,014 observations were allocated for training, while the remaining 10,672 observations were reserved for testing. These observations included 18 categorical variables and 11 numerical variables.

An analysis of the dataset revealed that 61% of observations were classified as "not obese," while the remaining 39% were classified as "obese," indicating a moderately imbalanced dataset. This imbalance implied a maximum baseline error rate of 39% for any predictive model. Consequently, robust classification techniques capable of handling imbalanced data were imperative to ensure reliable and accurate predictions.

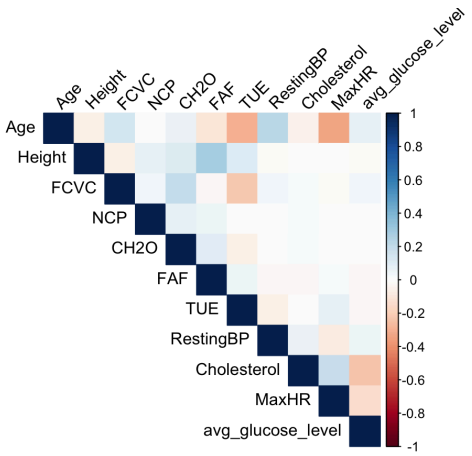
Another noteworthy aspect of the dataset was the significant proportion of missing data. Specifically, 8% of the data in the training subset and an additional 8% in the testing subset were missing. This consistent level of missingness across both subsets emphasized the need for appropriate imputation techniques to ensure the integrity of the analysis.

Variable	Variable Type	Proportion Missing Train	Proportion Missing Test
work_type	Discrete categorical	0.0833	0.0860
RestingECG	Discrete categorical	0.0817	0.0837
Hypertension	Binary	0.0817	0.0775
CALC	Ordinal categorical	0.0817	0.0793
Smoke	Binary	0.0815	0.0819
Heart Disease	Binary	0.0811	0.0793
TUE	Numeric	0.0810	0.0787
Race	Discrete categorical	0.0808	0.0766
MaxHR	Numeric	0.0808	0.0804
NCP	Numeric	0.0807	0.0776
FAVC	Binary	0.0806	0.0800
FCVC	Numeric	0.0804	0.0796
Gender	Binary	0.0803	0.0829
FAF	Numeric	0.0800	0.0750
Cholesterol	Numeric	0.0798	0.0774
FastingBS	Binary	0.0798	0.0809
SCC	Binary	0.0797	0.0781
Residence Type	Binary	0.0797	0.0790
CH2O	Numeric	0.0796	0.0801

Family History with Overweight	Binary	0.0796	0.0796
Height	Numeric	0.0793	0.0805
Ever Married	Binary	0.0793	0.0790
CAEC	Ordinal categorical	0.0792	0.0817
Exercise Angina	Binary	0.0789	0.0792
Age	Numeric	0.0787	0.0806
Resting BP	Numeric	0.0784	0.0802
Avg Glucose Level	Numeric	0.0781	0.0807
MTRANS	Discrete categorical	0.0778	0.0832
Stroke	Binary	0.0799	0.0813

b. Variables exploratory data analysis

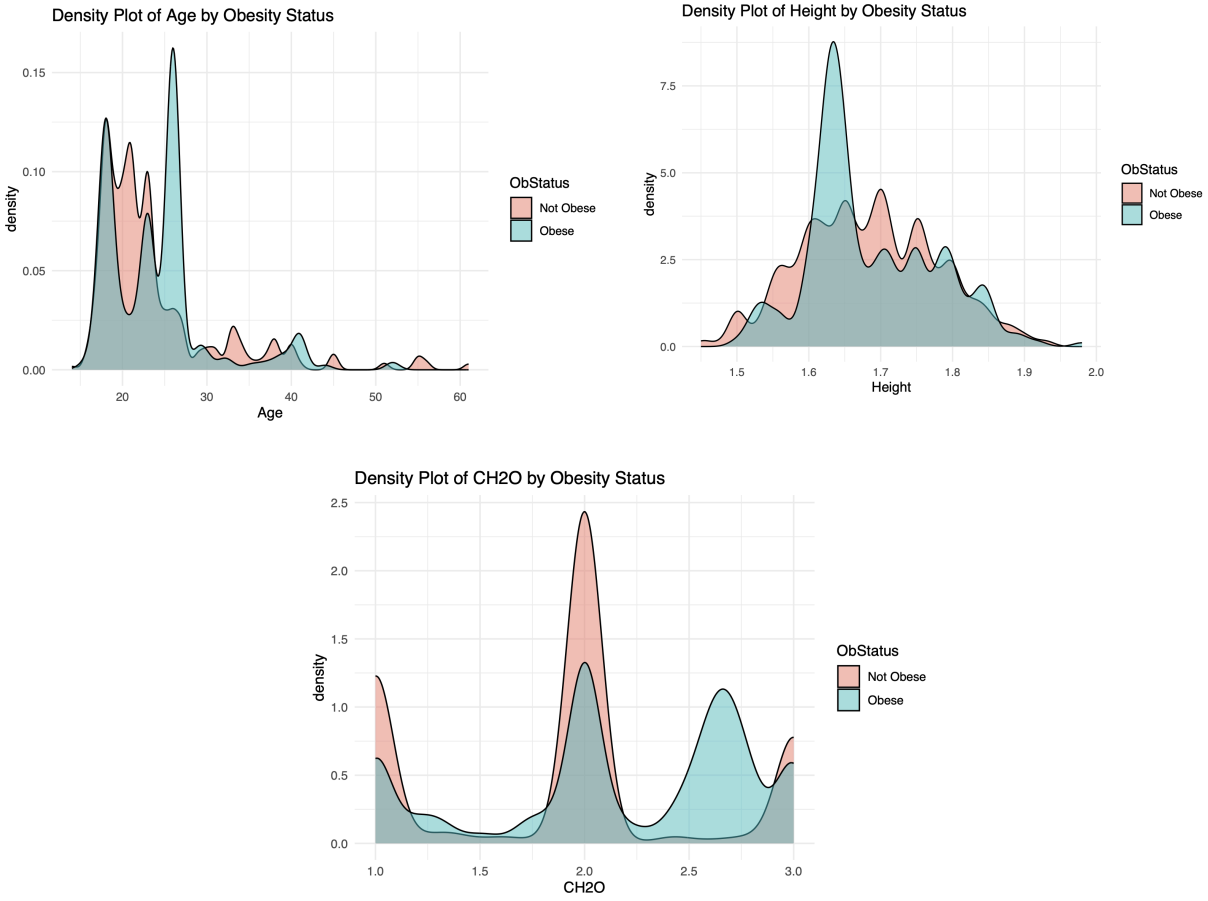
i. EDA - numerical



The above plot shows the correlation matrix for the numerical variables. Darker-colored boxes indicate stronger correlations. According to the plot, TUE is negatively correlated with

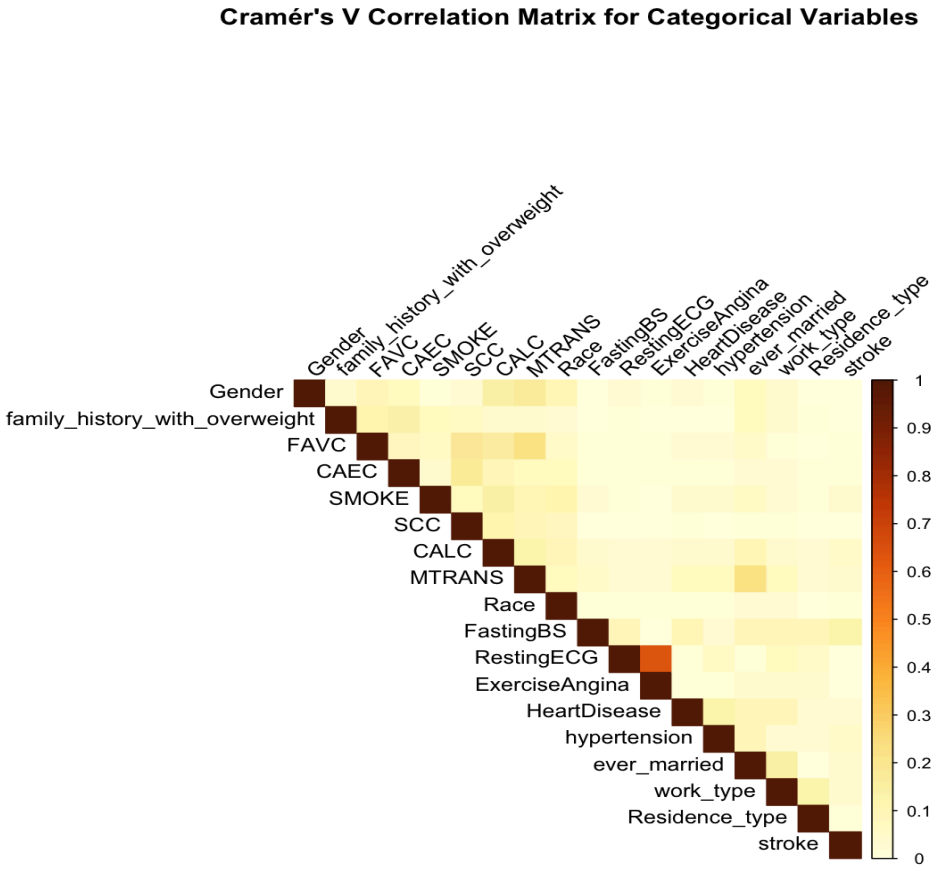
Age and FCVC, while MaxHR is negatively correlated with Age. Cholesterol and avg_glucose_level also exhibit a negative correlation. In contrast, RestingBP is positively correlated with Age, and FAF and Height show a stronger positive correlation. However, none of the variables have a correlation higher than |0.5|, indicating that the relationships between these variables are relatively weak.

Density plots for each numerical variable were also explored, and the density plots for the three most important variables in our model are shown below. None of the density plots clearly distinguish between the two levels of obesity. Ideally, the distributions for "not obese" and "obese" would be markedly different. However, the distributions do show some differences, suggesting that these variables still contribute meaningfully to the model.



ii. EDA - categorical

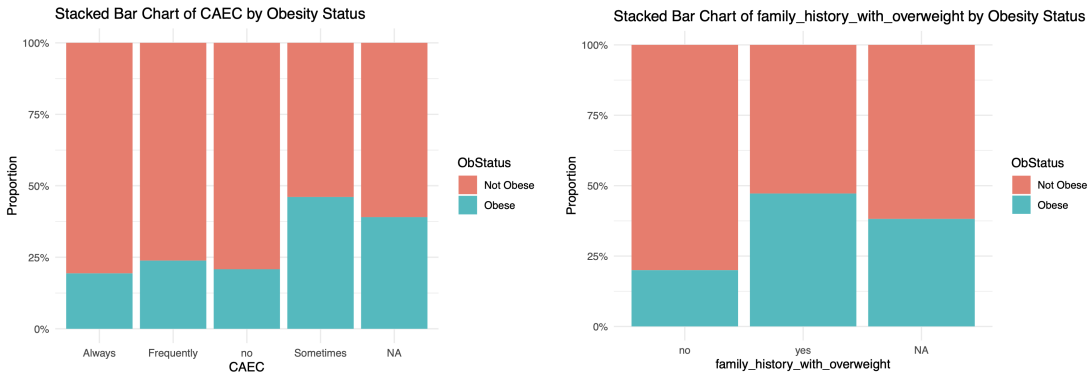
To determine correlations between categorical variables, Cramér's V was used. Cramér's V is a statistical measure used to determine the strength of association between two categorical variables. It is based on the chi-squared statistic and ranges from 0 to 1, where 0 indicates no association and 1 indicates a perfect association. Unlike the Pearson correlation for numerical variables, Cramér's V works for categorical data, making it a useful tool for understanding relationships in contingency tables.



The Cramér's V correlation matrix shows the strength of associations between all pairs of variables in the dataset. Darker colors in the plot indicate stronger associations, while lighter colors represent weaker or no association. Most of the pairwise relationships appear relatively weak, as seen from the abundance of light yellow shading. However, there are a few notable

associations—**RestingECG** and **ExerciseAngina**—show moderately stronger relationships, which stand out in deeper orange or brown tones. The diagonal values are all 1 because each variable is perfectly associated with itself. Overall, while no overwhelmingly strong associations exist, some relationships seem important, and the low overall redundancy suggests that the categorical features add meaningful diversity to the dataset.

Bar charts that plotted each categorical variable against Obesity Status were also explored to visually determine which variable may be the most important. Two of these bar charts are shown below - both of these variables show that their proportions differ between the different levels of obesity status, meaning they could be important in the model. These were important observations that were taken into consideration throughout the process of data analysis, particularly while conducting feature selection.



c. Imputation

The presence of missing values in the dataset necessitated imputation prior to proceeding with the project's objectives. Selecting an appropriate imputation method was critical to ensuring the highest possible predictive accuracy. To address this, various imputation methods were explored, ranging from simple techniques to more complex and computationally intensive approaches. The simplest approach involved imputing missing values by using the median for

numerical variables and the mode for categorical variables. This initial imputation enabled the fitting of logistic regression models, which informed the selection of the final model for this project. Once the model was selected, alternative imputation strategies were revisited to ensure optimal performance.

The project's initial imputation attempt was done using the mice (Multivariate Imputation by Chained Equations) library in R, one of numerous libraries available in R to impute missing values. For the imputation of the missing values, predictors were first split into four categories: numerical variables, ordinal categorical variables, discrete categorical variables, and binary categorical variables. For each different category, various imputation methods were used. Predictive mean matching, linear regression with noise, linear regression without noise, linear regression without bootstrap, and linear regression with bootstrap were attempted for numerical variables. While imputations for the categorical variables were held constant, cross-validation was performed on each of the different numerical imputations. Cross-validation results showed that linear regression without noise performed better than the other methods, leading to the decision to choose this method as the final imputation method for numerical variables. The mice library only allows for logistic regression to be used for binary variables, hence this method was used to impute all the binary variables. Proportional odds logistic regression was used for ordinal categorical variables, allowing for a clear interpretation of how much the odds of y in a higher category will increase (by a certain ratio) for each increase in input variable for x. Discrete categorical variables were imputed using the Bayesian polytomous regression model. Using only proportional odds logistic regression on all categorical variables was also attempted. However, using the selected methods for the different variables proved to perform better during

cross-validation. As a result, the aforementioned methods were used for imputations in the final model with parameters m = 5 and maxIter = 50.

Imputations using other libraries in R were also attempted. The missForest library in R, which uses random forest models to predict numerical variables, was utilized as it captures non-linear trends in the data. These imputations lead to an increased overall prediction accuracy of 99.86%. Unfortunately, due to the computationally expensive nature of this imputation method, this work was completed after the project deadline and was not included in the final submission.

3. Methods and Models

a. Logistic Regression

Logistic regression was initially used to fit the training data. The first logistic regression was modeled without feature selection and yielded 74.7% accuracy. To explore if feature selection would improve the model, forward selection based on AIC score was utilized. This method selected 20 features, but decreased testing accuracy to 72.13%. Since the general accuracy of the logistic regression models was low, logistic regression was not selected as the final model. The high misclassification error rates also suggested a non-linear boundary between the two classifications of obese and not obese and the need to use more complex models.

b. Random Forest

Random Forest is an ensemble learning method which uses independent decision tree models to fit the data. Each decision tree uses a different subset of data and predictors, allowing for better generalization. The initial random forest model was a significant improvement over

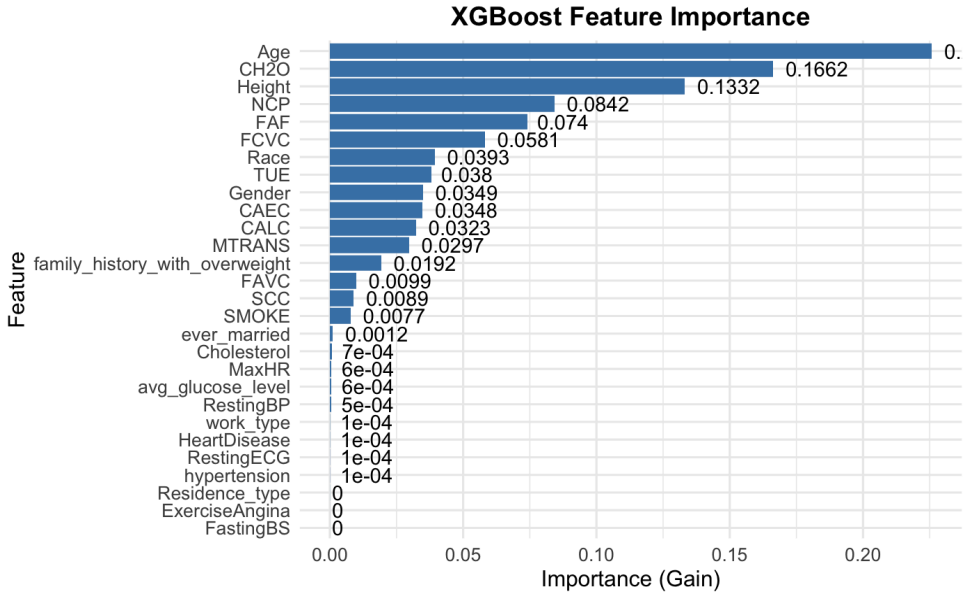
logistic regression, boasting 98.4% accuracy. Further exploration with feature selection and cross-validation allowed the model to achieve a 99.3% accuracy with 400 trees and 14 features selected using iterative feature addition.

c. XGBoost Model

XGBoost, formally known as extreme gradient boosting, is an ensemble learning model based on decision trees. XGBoost sequentially builds different decision trees, where each new tree focuses on correcting errors made in previous ones. Using default hyperparameter settings and all predictors, the model achieved an accuracy of 98.9%. Thus, the XGBoost model was selected to be the final model as it performed very well without tuning.

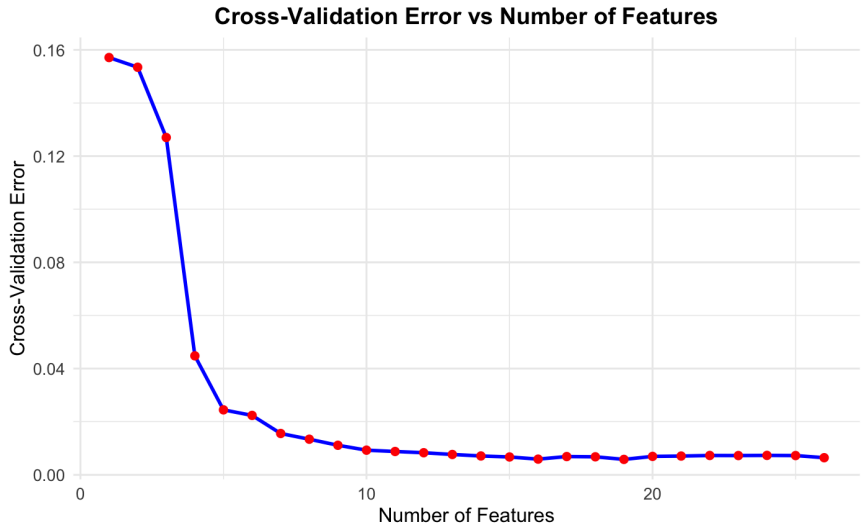
d. Feature Selection for XGBoost Model

The initial accuracy scores lead to the decision to use the XGBoost classifier as our final model. The XGBoost model provides a ranking of the most important features used, which can be seen in the plot below.



All features are listed on the left and their corresponding Importance measures are plotted as horizontal bar graphs. The higher the Importance score, the more important the feature was in building the model. The top 5 predictors of the model are age, “daily water intake” (CH20), height, “number of main meals” (NCP), and “physical activity frequency” (FAF) (Kaggle).

According to this ranking, cross-validation and forward selection were utilized to select the best subset of predictors to be used in the XGBoost model. The subsets are created by adding the next most important feature. For example, the first subset of features would consist of Age and the second subset of features would consist of Age and daily water intake (CH20). Each subset is then used to build 100 rounds of 5-fold cross-validation XGBoost model. The minimum cross-validation error is then obtained to evaluate the models built on various subsets of features and are plotted in the graph below.



The elbow method was initially used to select the best subset of features. The cross-validation error seems to stabilize starting from the first 11 most important XGBoost features. Thus, the first revised XGBoost model was trained on those 11 features, which resulted in a testing accuracy of 99.26%. However, the accuracy score may be further improved by selecting another feature subset.

For the second revised XGBoost model, the new feature subset was selected by minimum cross-validation error. This value was given by the XGBoost model trained on the first 19 most important features. The second revised XGBoost model was then built with those 19 features and resulted in an accuracy score of 99.59%, which is higher than the fine-tuned Random Forest model.

The two revised by 0.33% in model accuracy and 8 features. Since the goal was to improve the accuracy of the model as much as possible, it was decided that the trade-off of adding 8 more features to the model was worth the increase in accuracy. The first 19 most important features of the original model were ultimately selected.

4. Final Model

The final model selected for the analysis was an XGBoost model, including 19 key predictors: age, “daily water intake” (CH20), height, “number of main meals” (NCP), “physical activity frequency” (FAF), consumption of vegetables (FCVC), race, amount of “time using technology devices” (TUE), gender, “consumption of food between meals” (CAEC), caloric intake (CALC), mode of transportation (MTRANS), family history of being overweight, “frequent consumption of high-calorie foods” (FAVC), “consumption of sweet drinks” (SCC), smoking status (SMOKE), marital status (ever_married), “cholesterol level” (cholesterol), and “maximum heart rate” (MaxHR) (Kaggle).

The model parameters were fine-tuned through an extensive hyperparameter tuning process to enhance its predictive performance. The final configuration is as follows:

- Objective: binary:logistic (optimized for binary classification)
- max_depth: 8 (allowing for a deeper tree to capture complex patterns)

- eta: 0.3 (balancing learning speed and step size)
- subsample: 1 (using all data for training each tree)
- colsample_bytree: 0.6 (using 60% of features for each split to prevent overfitting)
- min_child_weight: 1 (minimizing overfitting with fewer splits on small data)
- gamma: 0 (no additional regularization applied)
- nrounds: 100 (limiting the number of boosting rounds to avoid overfitting)

This model was selected for its ability to balance complexity and generalization, making it well-suited for the predictive classification task.

5. Conclusion and Recommendations

a. Results and Limitations

Throughout the process of conducting this project, the team was able to gain hands-on experience in conducting supervised machine learning while also exploring more complex models such as random forest and XGBoost models. The project also allowed the team to gain a deeper understanding of imputations and how to deal with missing values in data. The knowledge gained throughout the project allowed for the team to land on a final model yielding a 99.59% prediction accuracy.

While the final XGBoost model is highly effective in predicting obesity status, the model is not very interpretable as to what causes obesity status. The only method for understanding the decision-making process of the XGBoost model is the feature importance captured by the model, but it is not clear how those features contribute. Furthermore, though more complex models such as neural networks can be used, there may have been more improvement in accuracy simply by improving the imputation method. The imputation step was particularly difficult as with the use

of the mice package, there were many parameters to fine-tune according to the data's feature types. Additionally, in order to achieve more precise imputations, running the code was quite time-consuming. This resulted in limited imputation technique that affected how well the model was trained on the training data.

b. Recommendations

To address interpretability, techniques such as SHAP values can help with estimating the behaviors of the features, but there are still no direct coefficients to interpret like with logistic regression models. As discovered in the attempts to use the missForest library, using a more complex method for imputation than mice would have resulted in better imputed values, and ultimately a better model. This was a good reminder that no matter how much fine-tuning is done, quality data inputs are necessary to improve model performance. Finally, a general revision to the XGBoost model can be to tackle multicollinearity among the features.

Though there are still improvements to be made, the authors are content with the final model and model accuracy score.

6. Acknowledgments

The team would like to thank Dr. Akram Almohalwas and Stella Huang for their continuous support and guidance which was critical to the completion of this project.

References

Centers for Disease Control and Prevention. (n.d.-a). About adult BMI. Retrieved December 2, 2024, from https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

Centers for Disease Control and Prevention. (n.d.-b). Adult obesity facts. Retrieved December 2, 2024, from https://www.cdc.gov/obesity/data/adult.html

Centers for Disease Control and Prevention (CDC). (2022, September 2). Adult BMI calculator. U.S. Department of Health & Human Services. Retrieved December 2, 2024, from https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

Gray, D. S., & Fujioka, K. (1998). Use of relative weight and Body Mass Index for the determination of adiposity. Journal of Clinical Nutrition, 66(4), 1–9. https://doi.org/10.1002/j.1550-8528.1998.tb00690.x

Kaggle. (n.d.). Predicting obesity status: Data. Retrieved December 2, 2024, from https://www.kaggle.com/competitions/predicting-obesity-status/data

Westfall, P., & Henning, T. (2021). Ordinal regression. In People Analytics Regression Book. Retrieved December 2, 2024, from https://peopleanalytics-regression-book.org/ord-reg.html

World Health Organization. (2010). Global status report on noncommunicable diseases 2010. Geneva, Switzerland: World Health Organization. Retrieved December 2, 2024, from https://iris.who.int/bitstream/handle/10665/44579/9789240686458_eng.pdf

World Health Organization. (2024). Obesity and overweight. Retrieved December 2, 2024, from https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

Westfall, P., & Henning, T. (2021). Ordinal Regression. In People Analytics Regression Book.

Retrieved [Month Day, Year], from

https://peopleanalytics-regression-book.org/ord-reg.html