Lifestyle and Depression: Predictive Modeling and Inference with NHANES Data

1. Abstract

Based on the National Health and Nutrition Examination Survey (NHANES), this report models the relationship between people's lifestyle and the occurrence of depression. Exploratory data analysis is conducted before modeling. Three separate predictive methods are applied, namely Random Forest, Gradient Boosting, and Adaboost. Cross-validation(CV) is utilized to tune hyperparameters, including maximum tree depth, shrinkage, and the number of iterations. The optimal hyperparameters for each method are selected based on the lowest CV deviance, and then compared based on test accuracy. All methods give high and relatively close accuracies, with Gradient Boosting achieving the highest test accuracy (91.20%) and lowest model complexity (maximum depth 2 and 94 iterations), and it is chosen as the optimal model. A logistic regression model is constructed separately for inference analysis, suggesting that females and people with alcohol or smoking addiction and bad dietary habits are more prone to get depressed than others; other lifestyle predictors demonstrate no significant association with depression.

2. Introduction

- Motivation

Depression has become increasingly prevalent in modern society, affecting individuals across various age groups and backgrounds. Specifically, understanding how daily habits, such as diet, smoking, and alcohol consumption, relate to mental health can provide valuable insights for prevention and intervention strategies.

This project explores the relationship between lifestyle factors and depression using data from the National Health and Nutrition Examination Survey (NHANES). By analyzing and understanding their associations with depression, we aim to identify lifestyle characteristics that may serve as potential risk factors or protective factors.

- Goal

Specifically, this project focuses on solving two main questions:

- **Prediction Question**: Can we predict whether an individual is depressed based on their lifestyle? (Example: Diet, Smoke, Alcohol)
- **Inference Question**: How do lifestyle factors (Example: Diet, Smoking, Alcohol) affect the odds of developing depression, and which of these factors are statistically significant?

3. Data Description and EDA

The raw data has 5985 rows and 27 columns. Since the project focuses on lifestyle factors, the target dataset has 12 columns in total: 2 columns of basic information(gender and age), 9 columns of lifestyle, and 1 column of depression result. The table below explains each feature:

Variable	Description	Type
RIAGENDR	1: male; 2: female	Categorical
RIDAGEYR	Age in years (20 to 80)	Continuous
ALQ111	Ever had a drink of any kind of alcohol: 1: Yes; 2: No (Will not be used since everyone drinks)	Categorical

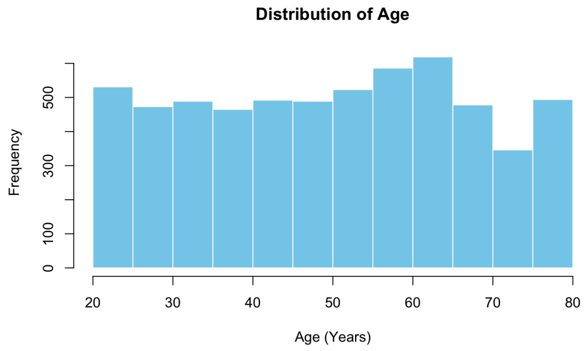
ALQ121	Past 12 months, how often drink alcoholic beverages: 0 to 10	Categorical
ALQ151	Ever have 4/5 or more drinks every day?: 1: Yes; 2: No	Categorical
DBQ700	How healthy is the diet: categorical cov grouped: 1: Excellent, 2: Very good; 3: Good; 4: Fair; 5: Poor	Categorical
DBQ197	Past 30-day milk product consumption: 0: Never; 1: Rarely; 2: Sometimes; 3: Often; 4: Varied	Categorical
DBD895	# of meals not home prepared during the last 7 days	Continuous
DBD905	# of ready-to-eat foods in the past 30 days	Continuous
DBD910	# of frozen meals/pizza in past 30 days	Continuous
SMQ020	Smoked at least 100 cigarettes in life: 1: Yes; 2: No	Categorical
depressed	1 if the individual showed signs of depression and 0 if not	Categorical

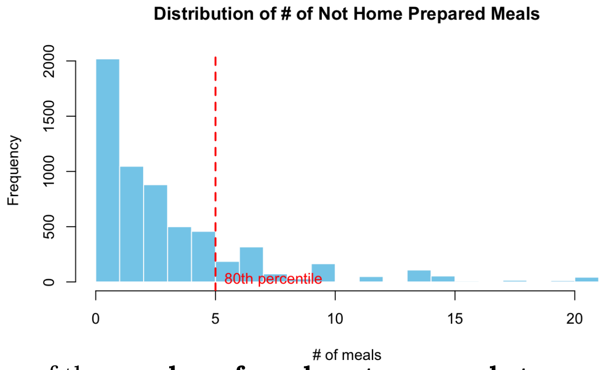
- NA checking

There were no missing (NA) values in the dataset, so no imputation or additional handling of missing data was required.

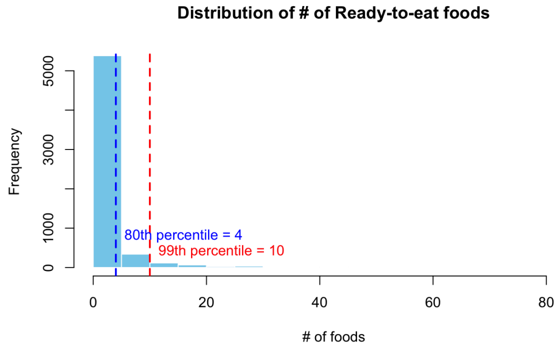
- Univariate Analysis

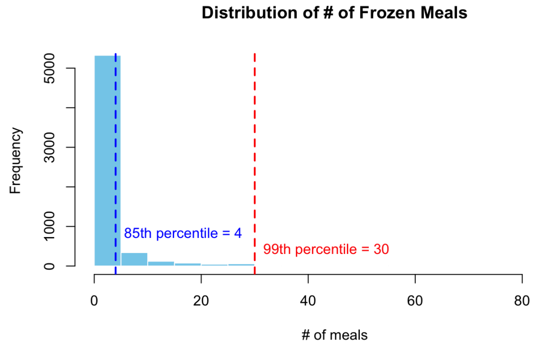
Basic demographic information, such as **gender** and **age**, is relatively balanced in the dataset, with **3,004 males** and **2,981 females**, indicating no major gender imbalance. The age distribution is visualized in the left plot below, with no extreme outliers present.





The right plot above illustrates the distribution of the **number of meals not prepared at home** over the past 7 days. Approximately **80%** of respondents consume **five or fewer** such meals. However, there is a small group of **43** individuals who reported consuming all **21** meals during the week outside the home.





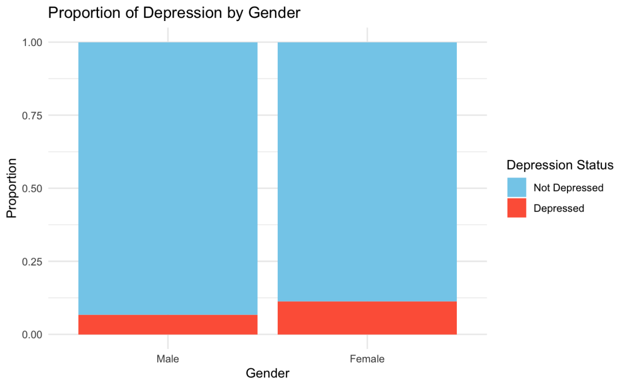
The left plot above illustrates the distribution of the **number of ready-to-eat foods** in the past 30 days. **80%** of respondents consume **4** or fewer such foods, and **99%** of respondents consume **10** or fewer such foods. However, there is a small group of **11** individuals who consume over **50** foods.

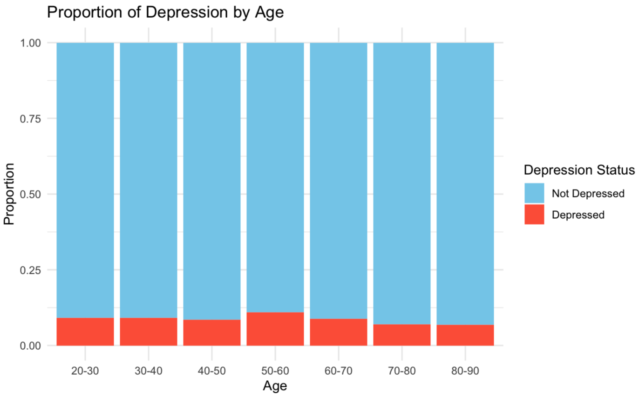
The right plot above illustrates the distribution of the **number of frozen meals** in the past 30 days. **85%** of respondents consume **4** or fewer such foods, and **99%** of respondents consume **30** or fewer such foods. However, there is a small group of **12** individuals who consume over **50** foods.

- Bivariate Analysis

The right plot below shows that the depression rate is higher among **females (11.20%)** compared to **males (6.72%)**.

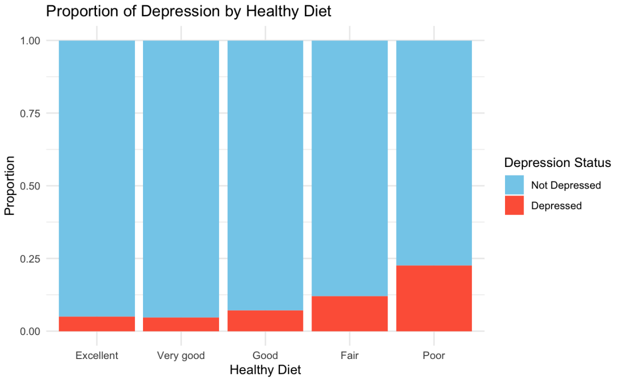
The left plot below shows that the age between **50** and **60** has the highest depression rate **(11.01%)**, while the age between **80** and **90** has the lowest depression rate **(6.83%)**.

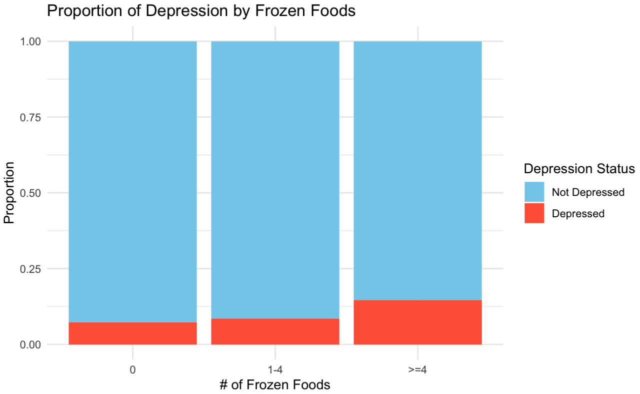




The left plot below shows that as **self-rated diet health decreases, the proportion of individuals with depression increases**. For example, only **4.69–4.96%** of individuals who rated their diet as **excellent** or **very good** were depressed, while the rate jumped to **22.49%** among those who rated their diet as **poor**.

The right plot below shows that as **the consumption of frozen food increases, individuals with depression increases**. Only **7.34%** of individuals who **never** consumed frozen food were depressed, while the rate jumped to **14.56%** among those who consumed more than **4** frozen foods in the past month.





4. Methods

● Predictive Question

Since it’s a binary classification task, predicting whether a person is depressed or not based on lifestyle features, this project first extracted lifestyle features from the raw data. Then, three different tree models are used for prediction, and their performance is compared:

- **Random Forest**: Random Forest builds an ensemble of trees where each tree is constructed using a random subset of features and bootstrap samples of training data. By aggregating predictions across many trees, this method reduces overfitting and variance compared to single-tree methods.
- **Gradient Boosting**: Gradient boosting builds predictive models in stages by merging the strengths of weak learners to enhance overall predictive accuracy. It is relatively robust to outliers, as the ensemble approach mitigates individual data points' influence. Additionally, this approach also reveals the relative influence of each feature.
- **AdaBoost**: AdaBoost utilizes a slightly different approach than Gradient Boosting. It initializes a model that equally weights each observation and explicitly assigns more weights to poorly predicted observations in each iteration. AdaBoost has the ability to reveal feature influence like Gradient Boosting, but is more sensitive to outliers than Gradient Boosting.

● Inference Analysis

This project uses logistic regression to perform inference analysis. Logistic regression is particularly suitable for modeling binary outcomes such as depression status and offers interpretable results. It enables the use of p-values to assess the statistical significance of each predictor, helping to identify which lifestyle factors are most strongly associated with depression. Additionally, the estimated coefficients from the model can be used to construct the final regression equation, providing a quantitative understanding of how each variable influences the probability of depression.

5. Prediction Results

● Random Forest

The Random Forest model was implemented using the randomForest package in R, with depression status as the binary response variable and the same lifestyle factors as predictors used in the Gradient Boosting model. Following the same methodology, the dataset was split into a training set (70%) and a testing set (30%) using the same random seed (42) to ensure direct comparability between models. We also factorized the "RIAGENDR", "ALQ121", "ALQ151”, "DBQ700", "DBQ197", "SMQ020", ‘depressed” variables.

The initial Random Forest model was constructed with 500 trees, with the default number of variables tried at each split (3 variables). This basic model showed a strong overall accuracy of 90.98% but demonstrated a notable class imbalance challenge, with low sensitivity (0%) for identifying depressed individuals.

Initial Model Performance:

- Accuracy: 90.98%
- Cohen's Kappa: -0.0011
- Sensitivity: 0.00%
- Specificity: 99.94%

These metrics indicated a significant class imbalance, affecting the model's ability to correctly identify cases of depression (Sensitivity).

The dataset is heavily imbalanced, with significantly more instances of the "not depressed" class. The model showed high accuracy but struggled with sensitivity due to the overwhelming majority of class predictions. This may lead us to think about different methods. Random Forests, while powerful, may not always capture complex interactions effectively without sufficient tuning and consideration of interactions. In this model, there was minimal transformation of continuous variables into categorical bins, which might be significant in reducing model sensitivity to outliers and capturing non-linearity.

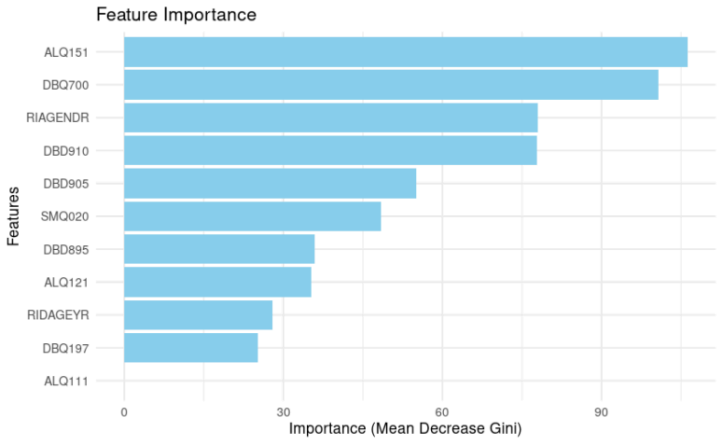
We explored and tested out some methods to improve the model's accuracy. We introduced cross-validation to improve model robustness and tuned hyperparameters (e.g., `mtry`) to find optimal values for the model using ROC as a metric. This method aims to improve the model's accuracy. The cross-validation model's results are

- Accuracy slightly decreased, but overall the performance accuracy stays the same (90.76%)
- Additional note: Despite using 10-fold cross-validation and optimizing for ROC, the model continued to struggle with the minority class. Best model had mtry = 2 with ROC = 0.7093. The model still showed 0% sensitivity for the depressed class.

Oversampled with ROSE Package to address class imbalance:

- Class distribution was balanced to 2,085 non-depressed vs. 2,104 depressed
- Sensitivity improved (32.92%), but accuracy dropped slightly (82.52%).
- Specificity decreased to 87.40%
- Balanced accuracy improved to 60.16%
- Kappa increased to 0.1594, indicating a slight improvement in predictive agreement beyond chance.

The feature importance metrics reveal potential limitations in predictor strength. Alcohol consumption (ALQ151) and diet-related variables (DBQ700) showed the highest importance scores (108.65 and 98.53, respectively), followed by additional dietary factors (DBD910, 78.18) and gender (RIAGENDR, 77.69). Smoking status (SMQ020) had moderate importance (47.67). Despite these relative rankings, the overall predictive power of even the top features was insufficient to accurately classify the minority class, suggesting that these variables may have complex, non-linear relationships with depression that the Random Forest algorithm struggled to capture effectively.



There are some limitations to this model. Random Forest alone struggled with imbalance unless oversampling or resampling techniques were explicitly applied. To apply this model in practice, ongoing rebalancing efforts are very much needed. Compared to Gradient Boosting or AdaBoost, Random Forest lacks internal boosting mechanisms to adapt and focus on harder-to-classify instances, which is inherent in ensemble boosting. Even with cross-validation optimization, the model failed to improve sensitivity. One other reason is that the variables selected may not provide a sufficient signal for the Random Forest algorithm to distinguish between depressed from non-depressed individuals effectively.

- Gradient Boosting

The model was trained using the gbm package in R, with depression status as the binary response variable and lifestyle features as predictors. The dataset was randomly split into a training set **(70%)** and a testing set **(30%)** to enable model training and evaluation.

The basic model was specified with the Bernoulli distribution to model the binary outcome, and the **maximum number of boosting iterations** is set to **500**. The **maximum depth for each decision tree** is **4**. The accuracy on test data for this model is **89.42%**.

To improve the predicted performance of the basic mode, we did some feature engineering to solve the problem of sample imbalance and used k-fold cross-validation to find the best hyperparameters.

Four continuous features were transformed into categorical variables, which helps to reduce sensitivity to outliers.

- **RIDAGEYR** Age in years: This variable was grouped into "20-30", "30-40", "40-50", "50-60", "60-70", "70-80", "80-90".
- **DBD895** Number of meals not home-prepared (past 7 days): This variable was split into six categories to reflect the increasing frequency of eating out: "0", "1", "2", "3", "4-5", ">=6".
- **DBD905** Number of ready-to-eat foods consumed (past 30 days): This variable was grouped into broader consumption levels: "0", "1-2", and ">=3".
- **DBD910** Number of frozen meals/pizza (past 30 days): This variable was grouped into "0", "1-4", and ">=4".

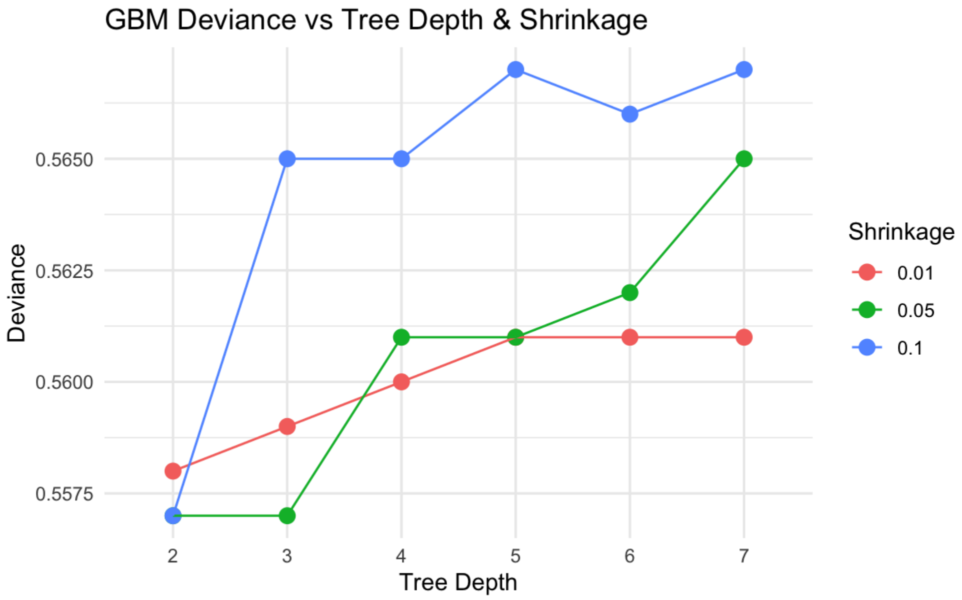
We used **5-fold cross-validation** to identify the optimal hyperparameters for the gradient boosting model, including:

- **Maximum tree depth**, ranging from 2 to 7
- **Shrinkage rate** (i.e., learning rate), tested at 0.01, 0.05, and 0.1
- The **optimal number of boosting iterations**, less than 500.

The plot below summarizes the model deviance during the cross-validation across different combinations of these parameters. Four configurations achieved the lowest deviance **(0.557)**:

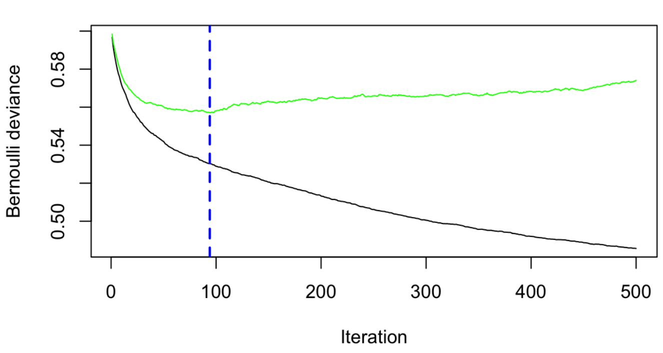
- Depth = 2, shrinkage = 0.05, iterations = 163
- Depth = 2, shrinkage = 0.10, iterations = 94
- Depth = 3, shrinkage = 0.05, iterations = 163

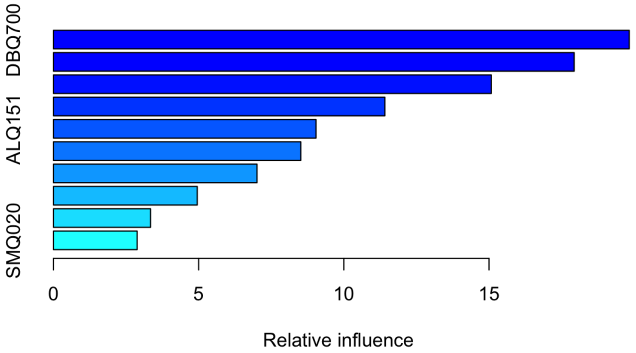
We selected the second configuration as the final model for Gradient Boosting, due to its low deviance, low tree depth, and lower number of iterations, which helps reduce the risk of overfitting.



The left plot below shows the deviance changes between the train set (black) and the test set (green), and reveals that the best iteration with the lowest deviance on the train set is **131**.

The plot on the right displays the relative influence of each feature in the gradient boosting model. The top four features— **healthy diet status (DBQ700), age in years (age_group), number of meals not home-prepared (meal_group), and frequency of drinking alcoholic beverages (ALQ121)** —contribute the most to the model's predictive performance.





Finally, after applying this model to the test data, the accuracy is **91.20%**.

- AdaBoost

The model is trained using the gbm package in R. The initial model is constructed with lifestyle variables as predictors. Like gradient boost, the dataset is randomly split into 70% training and 30% testing.

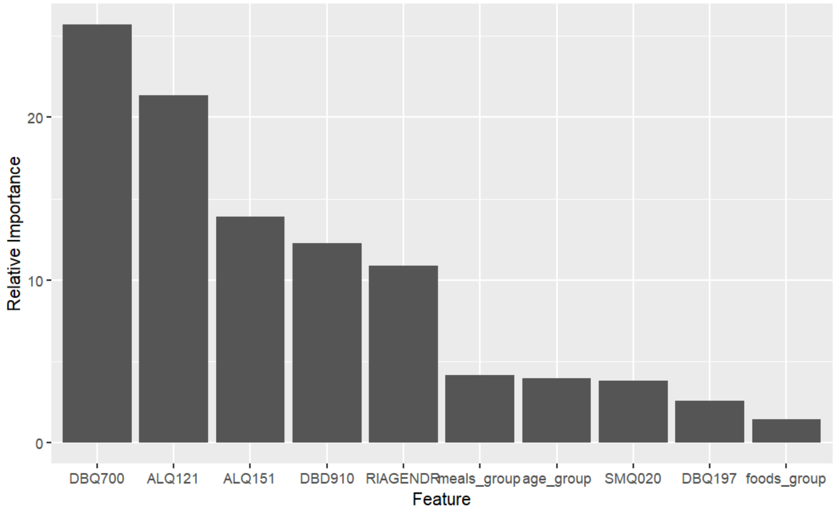
The initial model is constructed with 1000 trees, and the maximum depth for each tree is 1. Already, it yields a high accuracy of 90.98% on the test set, almost equal to the test accuracy of the Gradient Boost model after parameter tuning. Still, we will apply 5-fold cross-validation to seek improvement. We should expect the optimal number of boosting iterations to be relatively small to avoid overfitting.

The same feature engineering in Gradient Boost is done to reduce the model's sensitivity to outliers: Adaboost is known to be more sensitive than Gradient Boost, so this step is maintained. After feature engineering, the accuracy of the model is improved to 91.04%.

The hyperparameters we are tuning are still:

- **Maximum tree depth**, ranging from 2 to 7;
- **Shrinkage rate**, tested at 0.01, 0.05, 0.1. Smaller shrinkage rate usually requires more iterations, so we are not testing for even smaller values.
- **Optimal number of iterations**, no higher than 1000. Note that the upper limit is lower than that of Gradient Boosting, since Adaboost is prone to overfitting–it is thus not necessary to run as many iterations.

The optimal setup is given by maximum depth 2, shrinkage 0.05, and iterations 103, which has the lowest deviance from cross-validation. The final model yields a test accuracy of 91.04%.



The plot above displays the relative importance of each feature in the model. The feature importance ranking is slightly different from the Gradient Boosting: the top features for the Adaboost model are **healthy diet status (DBQ700), frequency of alcohol consumption (ALQ121), experience with large amounts of alcohol daily (ALQ151)**, and **number of frozen meals (DBQ910)**, in that order.

- Model Comparison and Evaluation

The table below shows the three models' prediction performance under the scale of accuracy and their complexity (number of trees, maximum tree depths).

	Models		
Scale	Random Forest	Gradient Boosting	AdaBoost
Accuracy	90.76%	91.20%	91.04%
Iteration(number of trees)	500	94	103
Maximum tree depth	2	2	2

The accuracy of each model is pretty close (**~91%**), as is the maximum tree depth. However, the number of trees varies a lot. Random Forest model uses the maximum number of trees (**500**), while Gradient Boosting uses the minimum number of trees (94).

Since the model with higher accuracy and less model complexity is preferred, the Gradient Boosting model is selected as our final model.

The Gradient Boosting model has very high predictive performance on the non-depressed sample. Within the **1635** non-depressed test samples, **all** samples were correctly predicted as non-depressed; the **True Negative Rate (Specificity)** is **100%**. However, the model's predictive performance on the depressed samples is super low. Within the **161** depressed test samples, only **3** samples were correctly predicted as depressed.

There are **2** possible reasons why the model performs really badly on predicting the depressed sample.

- **Class Imbalance**: The original data doesn't contain enough positive (depressed) data; only **10%** of the data is positive. This skewed distribution might cause the model to be biased toward the majority of the class (non-depressed), which leads to a low recall rate.
- **Limited Feature Scope**: We focus on how lifestyle features predict the depression result. Thus, the model highly depends on these features to predict depression. Focusing only on lifestyle features might not provide enough predictive power to identify the depressed sample. The depression might also rely on other types of features, like disease history, individual economy, etc.

6. Inference Result

Our logistic regression analysis examined how various lifestyle factors affect the odds of developing depression. Given the significant threshold p-value < 0.05, we found 7 statistically significant predictors:

Predictors	**Odd Ratio**	**95% Confidence Interval**	**P-value**	**Interpretation**

RIAGENDR=2	2.15	1.76-2.63	<0.001	Females have 2.15 times higher odds of depression than males.
ALQ151=2	0.43	0.34-0.54	<0.001	People who don’t ever have 4/5 or more drinks every day have 0.57 times lower odds of depression than people who do.
DBQ700=4	2.28	1.44-3.81	<0.001	People whose diet is fair have 2.28 times higher odds of depression than people whose diet is excellent.
DBQ700=5	4.20	2.56-7.21	<0.001	People whose diet is poor have 4.20 times higher odds of depression than people whose diet is excellent.
DBQ197=4	4.36	0.93-15.42	<0.05	People who consumed milk variably have 4.36 times higher odds of depression than people who never consumed milk in the last 30 days.
DBQ910	1.03	1.02-1.04	<0.001	For each additional frozen meal or pizza consumed during the last 30 days, the odds of depression increase by 3%.
SMQ020=2	0.67	0.55-0.82	<0.001	People who smoke fewer than 100 cigarettes have 0.33 lower odds of depression than people who smoke at least 100 cigarettes.

Briefly, females tend to be more prone to depression than males. People who have 4/5 or more drinks and smoke a lot tend to be prone to depression. People who have bad dietary habits tend to be prone to depression.

Other predictors, like age in years, alcohol beverage consumption, and number of ready-to-eat foods in the past 30 days, did not show significant associations with depression.

7. Conclusion

Through hyperparameter tuning, we are able to produce and select the optimal model for depression prediction using Gradient Boosting with shrinkage 0.1, depth 2, and 94 iterations. We are able to achieve a high test accuracy of 91.20%. Our inference result shows that females and people with unhealthy lifestyles (alcohol and smoking addiction, and unhealthy dietary habits) have stronger associations with depression. These insights suggest that these demographics should raise their awareness towards susceptible depression symptoms, and shouldn’t hesitate to seek assistance. They also suggest that society raise awareness on female mental health, and should promote and provide guidance on how to lead a healthier life

8. Responsibility

Work Section	Jiahao Cheng	Kevin Fang	Vy Dang
Project Proposal	Work together		
EDA	Code, Visualize, and Interpret	-	-
Predictive Question	Build, evaluate, and tune the Gradient Boosting Model	Build, evaluate, and tune the AdaBoost Model	Build, evaluate, and tune the Random Forest Model
	Compare the three models' performance		-
Inference Analysis	Code Interpret	-	-
Rmd file	EDA, Gradient Boosting, Inference Analysis, Merge three files together	AdaBoost	Random Forest
Final Report	Introduction, Data Description and EDA, Method, Prediction Result (Gradient Boosting, Model Comparison), Inference Analysis, Conclusion, Responsibility	Abstract, Method, Prediction Result (AdaBoost, Model Comparison), Conclusion	Method, Prediction Result (Random Forest)