1 Executive Summary

This report presents findings from a logistic regression model developed to enhance understanding of factors contributing to the predictability of cardiac events. Utilizing data from a national cross-sectional study on health status in the United States, variables such as gender, age, ethnicity, education level, sleep hours, diabetes status, smoking habits, and BMI were investigated. The analysis revealed the significance of each factor in predicting the probability of cardiac events. Consequently, a logistic model was constructed, offering a tool to predict cardiac event probabilities based on an individual's age, BMI, diabetes status, ethnicity, education level, and smoking status.

Significant findings were uncovered through the logistic regression analysis. Age emerged as a noteworthy factor, indicating an increase in the odds of a cardiac event for each additional year. The log-transformed BMI demonstrated a substantial impact, revealing an increase in the odds of cardiac events for every BMI unit increase. Notable predictors also included diabetes, ethnicity (particularly being Black) education levels, and smoking status. Both higher education level and smoking status are linked to a decrease in the odds of a cardiac event, while the other predictors lead to an increase. However, sleep duration did not exert a considerable influence within the randomly generated subset under examination.

In conclusion, understanding key factors such as age, BMI, diabetes status, ethnicity, education level, and smoking status proves essential for healthcare professionals and researchers seeking insights into the multifaceted determinants of cardiac events. This comprehensive approach provides valuable knowledge for advancing cardiac event prediction and improving public health outcomes.

2 Introduction

This study focused on adults aged 20 and above, incorporating data collected from a random sample of 1910 individuals spanning from 2017 to March 2020. The dataset contains vital information on demographic variables such as age, gender, ethnicity, and education. Additionally, health-related measures, including sleep hours per night, diabetes status, smoking status, and BMI (Body Mass Index) were compiled. The primary focus of this investigation is the presence or absence of a cardiac event, with a participant considered to have experienced one if diagnosed with conditions such as congestive heart failure, angina, coronary heart disease, hypertension, myocardial infarction (heart attack), or stroke, as reported by a healthcare professional.

Our project seeks to analyze and identify predictive factors for cardiac events. In this paper, we will introduce the methods, procedures, and results using our model, along with a comparison to select the most effective one. Furthermore, we will incorporate evaluation and quantitative methods to elucidate how factors contribute to the increased or decreased risk of a cardiac event. As a result, our project serves as a valuable resource for both researchers and the general public, offering insights into the factors or combinations thereof that are most indicative of cardiac events.

3 Description of Subjects

3.1 Data checking and cleaning

The original dataset has 1910 rows and 11 columns. The first two columns are just IDs or indices that do not have a practical meaning. Therefore, the remaining 9 columns are meaningful to use. We have observed the

presence of some missing values (NA), and we will delve into the discussion regarding our approach to addressing them.

3.2 Filling in Missing Values

We observed the presence of missing values in columns including ***sleep.hrs***, ***diabetes***, **smoker**, and ***bmi***. There is a total of $1910 - 1659 = 251$ rows of NA missing values, which is 13% of the entire dataset. Besides the two rows that contain "don't know" for ***educ*** and ***diabetes*** that we deleted, we believe it is essential to fill the NA accurately. To address this, we opted to employ various models for predicting these missing values. Initially, we eliminated all rows containing NA values, resulting in a dataset without missing values. Utilizing this no-NA dataset, we employed stepwise linear regression to forecast the numerical variables ***sleep.hrs*** and ***bmi***. For the ***smoker*** variable, we applied a glm model (family=binomial). As for the variable ***diabete***s, which encompasses three categories, a multinomial model was utilized. In the case of ***sleep.hrs***, we utilized the model to predict missing values based on the no-NA dataset. For example, we introduced a new column called ***full_sleep.hrs*** which uses the logic as follows: if ***sleep.hrs*** is NA, we filled the cell of the new column with the prediction from the stepwise regression model; conversely, if ***sleep.hrs*** is not NA, we fill the cell of the new column with the original value from the original column. Consequently, we generated three additional columns, namely ***full_diabetes***, ***full_smoker***, and ***full_bmi***, which successfully addressed the missing values in their respective original columns.

3.3 Description of Characteristics of Subjects

3.3.1 Descriptive Statistics for Categorical Variables

Predictor	Categories	Total	Cardiac Event Absent, N	Absent %	Cardiac Event Present, N	Present %	Chisq Test
Gender	Male	924	518	56.1%	406	43.9%	x^2 = 1.5048, df = 1, p-value = 0.220
	Female	984	580	58.9%	404	41.1%	
Ethnicity	White	627	335	53.4%	292	46.6%	x^2 = 44.22, df = 2, p-value < 0.0001
	Black	528	261	49.4%	267	50.6%	
	Other	753	502	66.7%	251	33.3%	
Education	Less than 9th grade	143	85	59.4%	58	40.6%	x^2 = 10.162, df = 4, p-value = 0.038
	9-11 grade (12th grade w/o diploma)	220	116	52.7%	104	47.3%	
	High school graduate/GED or equivalent	456	242	53.1%	214	46.9%	
	Some college or AA degree	616	362	58.8%	254	41.2%	
	College graduate or above	473	293	61.9%	180	38.1%	
Do They Have Diabetes?	Yes	305	75	24.6%	230	75.4%	x^2 = 167.65, df = 2, p-value < 0.0001
	No	1559	1003	64.3%	556	35.7%	
	Borderline	44	20	45.5%	24	54.5%	
Do They Smoke?	Yes	447	264	59.1%	183	40.9%	x^2 = 0.46926, df = 1, p-value = 0.493
	No	1461	834	57.1%	627	42.9%	
BMI Category	Underweight	21	18	85.7%	3	14.3%	x^2 = 80.932, df = 3, p-value < 0.0001
	Normal	445	327	73.5%	118	26.5%	
	Overweight	579	333	57.5%	246	42.5%	
	Obese	863	420	48.7%	443	51.3%	

Table 1: Descriptive Statistics for Cardiac Event Predictors (Categorical)

Table 1 shows that ethnicity, education, diabetes status, and BMI all demonstrate significant correlations with cardiac event occurrences, with the Chi-squared test yielding p-values less than 0.05.

3.3.2 Descriptive Statistics Numerical Variables

Predictor	Age	Sleep Hours
Min	20.00	2.00
Mean	50.95	7.55
Median	52.00	7.50
StDev	17.50	1.63
Max	80.00	14.00
p-value	<0.0001	0.56

Table 2: Descriptive Statistics for Cardiac Event Predictors (Numerical).

Table 2 indicates that age is significantly associated with the outcome variable since the p-value < 0.0001. Sleep hours does not show a significant association since the p-value is 0.56 > 0.05. The p-values are for the likelihood ratio test against a null intercept-only model.

4 Results

4.1.1 Contingency Analysis

We made two-way contingency tables for the categorical variables. The categorical variables at this stage are ***gender, ethnic1, educ_new, full_diabetes, full_smoker,*** and ***full_bmi_categorical***. After generating two-way contingency tables, we also conducted Chi-squared test on each categorical variable to determine whether the specific categorical variable has any association with cardiac event. The p-values that we obtained from the Chi-squared test are correspondingly 0.2199 for ***gender***, 2.499e-10 for ***ethnic1***, 0.03778 for ***educ_new***, 2.2e-16 for the ***full_diabetes*** variable, 0.4933 for ***full_smoker,*** and 2.2e-16 for ***full_bmi_categorical***.

Through comparing these p-values with the significance level of 0.05, we conclude that ***ethnic1, educ_new, full_diabetes***, and ***full_bmi_categorical*** are significant predictors since these p-values are smaller than 0.05. The ***gender*** and ***full_smoker*** variables are not significantly associated with the outcome variable by itself. This result justifies our selection of main effect variables for our model.

4.1.2 Mosaic Plots

Based on the contingency tables we obtained, we drew mosaic plots to show whether each level of the categorical variable affects the outcome variable and how much the extent of their influence is. We generated mosaic plots for gender vs. event, ethnicity vs. event, and education vs. event and observed whether the trends were linear.

Firstly, through observation on the plots, we conclude that ***gender*** and ***full_smoker*** are not significant predictors since the occurrence of the event is around the same for different levels of these two categorical variables. Variables ***ethnic1, educ_new, full_diabetes***, and ***full_bmi_categorical*** are significant predictors. This result corresponds to the p-values of Chi-squared test of association.

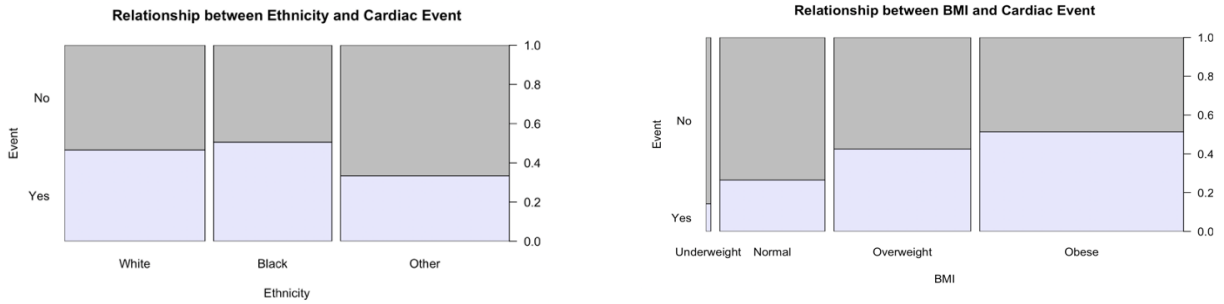

Figure 1: Mosaic plot of Ethnicity and BMI

Figure 1 consolidated the ethnic categories into three broader groups: “White”, “Black”, and “Other”. The categories “White”, “Black”, and “Other” show varying proportions of cardiac event occurrences. As for the BMI variable, there are four levels “Underweight”, “Normal”, “Overweight”, “Obese”. The categories show varying proportions of cardiac event occurrences and a quite linear trend.

Secondly, we noticed that in Figure 2, the “Less than 9th Grade” level does not follow the general linear decreasing trend with the order of levels being from “Less than 9th Grade” to “College Grad and Above”. Therefore, we introduced a new binary variable that only includes “less than 9th grade” and “Others” (which is a combination of “9-11th Grade”, “High School Graduate”, “Some college or AA Degree”, and “College Grad and Above”) to mitigate this situation. By re-categorizing the ***educ_new*** into two levels - “Less than 9th Grade” and “Others”, the plot now shows a general linear decreasing trend in Figure 3. However, based on the p-value being 0.6978 for the education variable from the new Chi-squared test, the education variable is no longer a significant variable.

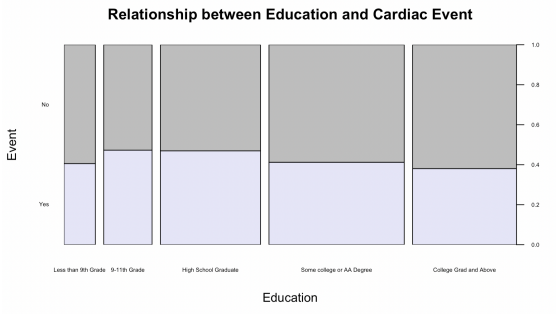

Figure 2: Mosaic plot of Education

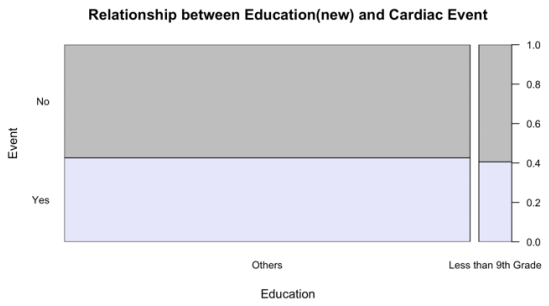

Figure 3: Mosaic plot of Education (grouped)

Thirdly, we observed a linear increasing trend in the ***diabetes*** vs event mosaic plot with order of levels being “No”, “Borderline” and “Yes”. Therefore, we recoded ***diabetes*** with “No” being 0, “Borderline” being 1, and “Yes” being 2. Thus, reordering the levels from 0 to 2 allows us to observe a clearer linear trend. Through the new contingency table and Chi-squared test for the recoded ***diabetes***， we justify that diabetes should be treated as a numerical variable. Viewing diabetes as a numerical variable, we drew a “slicing-dicing” plot of the Empirical Log-Odds of Event by ***Diabetes (new)*** and observed a linear increasing trend.

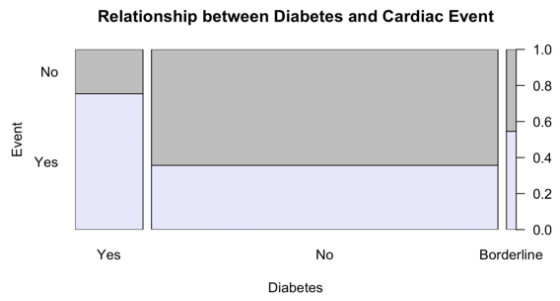

Figure 4: Mosaic plot of Diabetes

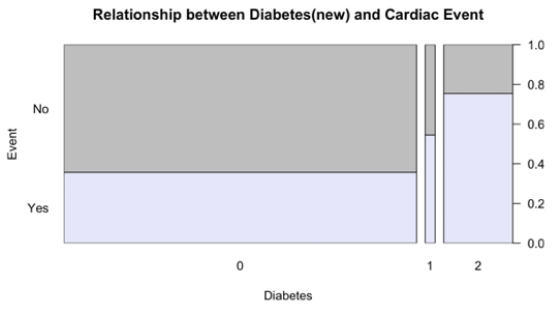

Figure 5: Mosaic plot of Diabetes (recoded)

4.2 Exploration and Choice of the Best Transformation For Numerical Variable

For every numerical explanatory variable, we drew a “slicing-dicing” plot of empirical log-odds to determine whether the variable’s relationship with the binary outcome was approximately linear, which would justify their inclusion as predictors in a logistic regression model.

Firstly, Empirical Log-Odds of Event by Age gives a relatively linear increasing trend justifying the significance of the ***age*** variable.

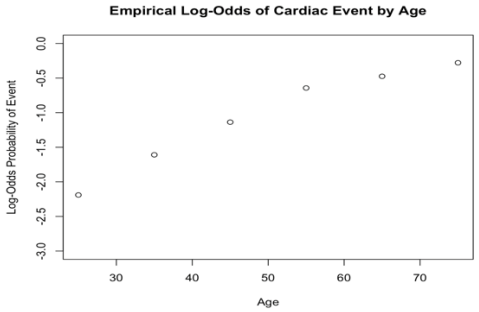

Figure 6: "slicing-dicing" plot of empirical log-odds for Age

Secondly, Empirical Log-Odds of Event by Sleep Hours doesn't explicitly give a linear trend and doesn't justify the significance of ***sleep.hrs***. Therefore, to determine whether ***sleep.hrs*** could potentially be a significant variable, we explored various transformations of ***sleep.hrs,*** including the square root of ***sleep.hrs,*** the log of ***sleep.hrs***, the exponential of ***sleep.hrs***, and the inverse of ***sleep.hrs***. Observing the new slicing-dicing plots of the transformations, Empirical Log-Odds of Event still don't show a clear linear trend. Therefore, we conclude that ***sleep.hrs*** is not a significant predictor for predicting cardiac event.

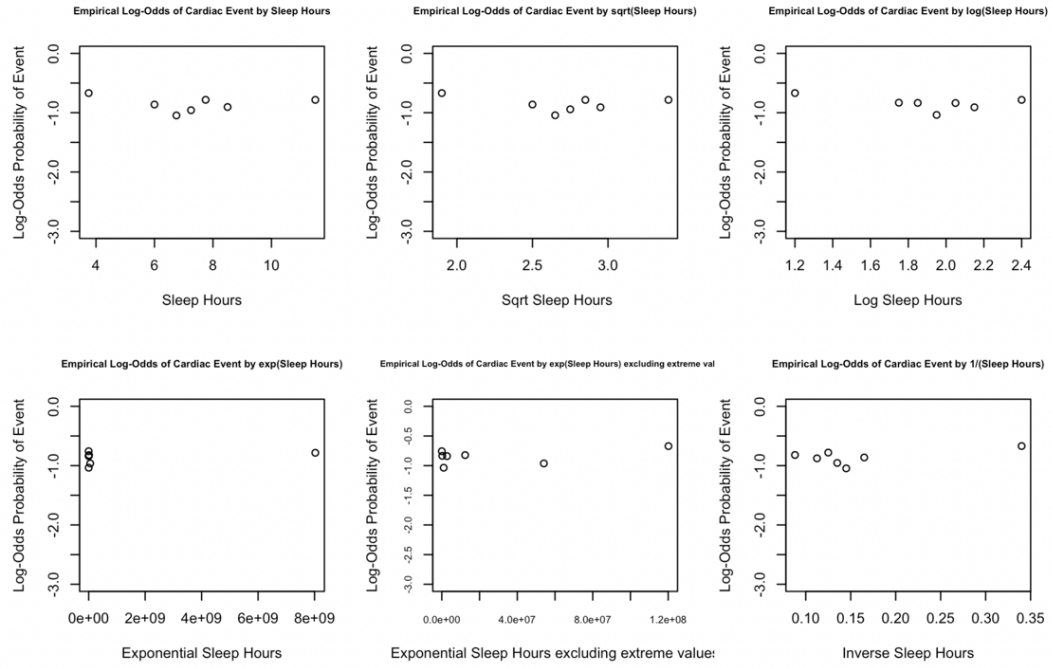

Figure 7: "slicing-dicing" plot of empirical log-odds for sleep hours and its transformation

Thirdly, we observed the slicing-dicing plot of Empirical Log-Odds of Event by BMI to check if it should be categorical or numerical. Our observation is that there is a linear increasing trend with few outliers. Considering the outlier and to determine whether ***bmi*** could potentially be a significant variable, we explored various transformations of ***bmi*** including square root, log, exponential and inverse of ***bmi.*** Since the new slicing-dicing plots with transformations show rather linear trend with outliers being seemingly less extreme, we could both consider ***bmi*** as categorical or numerical. Therefore, we went on to fit models considering ***bmi*** as either categorical or numerical.

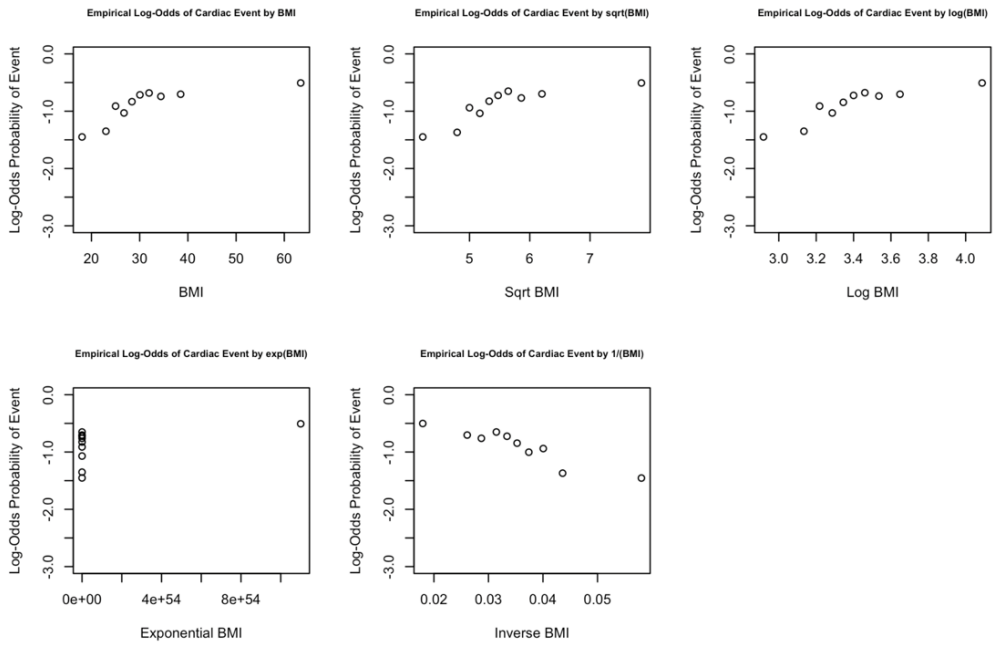

Figure 8: "slicing-dicing" plot of empirical log-odds for BMI and its transformations

First, it is assumed that ***bmi*** is categorical as it fits well in four categories and fitted a model using forward selection. We chose to use forward selection since it has computational advantage and also has the potential of giving us the best model. The model with the lowest AIC (2004.33) is selected: ***event*** ~ ***age*** + ***full_bmi_categorical*** + ***diabetes_new*** + ***ethnic1*** + ***educ_new*** + ***full_smoker.***

We then move on to finding whether interactions should be added to the model. We took an approach by grouping together the interactions with one of the interaction variables being the same.

Our process of testing significance of interactions is as follows:

i. Perform chunk test of variables

Chunk test:

$$H_o: \textit{all the coefficient of the interactions are } 0$$
$$H_a: \textit{at least one of the coefficients is not } 0$$

ii. If we observe a p-value less than 0.05, we will reject the null hypothesis. Therefore, the next step is to figure out which specific interaction is significant. Thus, we perform the following test on each interaction:

$$H_o: \textit{the coefficient of the interaction term is } 0$$
$$H_a: \textit{the coefficient of the interaction term is not } 0$$

For example, if the main effect variables are a, b, c, and d, we perform chunk test: the first test is on interaction terms ab, ac, ad, the second is on ba, bc, bd, the third on ca, cb, cd, the fourth on da, db, dc. If the p-value of the test on the model with ab, ac, and ad is less than 0.05, then we perform a single hypothesis test separately on ab, ac, and ad to determine whether the specific interaction term is significant.

Specifically, in our model interaction exploration, we observed a p-value of 0.03671(<0.05) in the chunk test related to ***ethnic1***.This means that we reject the null hypothesis that all the coefficients for the interaction terms related to ***ethnic1*** is 0. Therefore, there is convincing evidence that at least one coefficient of the interaction terms related to ***ethnic1*** is not 0.

Since all the other tests give a p-value greater than 0.05, we fail to reject their null hypothesis that the coefficients of those interaction terms are all 0 and there is no convincing evidence that at least one of those interaction terms are 0. After checking all the interactions related to ***ethnic1***, we found that the interaction between ***ethnic1*** and ***age*** is the most significant with a p-value being 0.023705. Although it is not less than the threshold of 0.01, we compared the AIC of the main effect model generated from forward selection and the model with the interaction ***ethnic1*** and ***age.*** The model with the interaction has an AIC of 2000.846, which is more than 2 units decrease of AIC compared to the main effect model, so we concluded that the interaction is significant. The model with ***ethnic1*** and ***age*** interaction is ***event ~ age + full_bmi_categorical + diabetes_new + ethnic1 + educ_new + full_smoker + ethnic1*age.***

To further lower AIC, we performed transformations on the interaction model above. After trying all the possible combinations of transformations, we found two models with the same AIC being 2000.676. One is exponentiating ***diabetes_new***, and the other one is exponentiating ***diabetes_new*** and squaring ***age***. Since they have the same AIC, we would prefer the simpler model, which is only exponentiating ***diabetes_new***. This will give the model: ***event ~ age + full_bmi_categorical + exp(diabetes_new) + ethnic1 + educ_new + full_smoker + ethnic1*age.***

Furthermore, we made ***bmi*** numerical since there is a general increasing linear trend for BMI vs event, we decided to find more models considering ***bmi*** as a numerical variable. Using forward selection again, we get the model ***event ~ age + full_bmi + diabetes_new + ethnic1 + educ_new + full_smoker***, with the AIC being 1993.8. Like before, we tried finding interaction terms through grouping them with one of the interactions being the same. However, this time, there is no p-value less than 0.05, and we fail to reject all the null hypothesis that the beta for the interaction terms is 0. Therefore, there is no convincing evidence that at least one of the betas of the interaction terms is 0.

Since no interactions are significant, we then moved on to finding transformations on current variables. We first tried single transformations on the three numerical predictors, which are ***age***, ***diabetes_new***, and ***full_bmi*** as we believe that if the single transformation is not significant, the combination of the insignificant transformations would also be insignificant. If we take log of BMI, then the model would be ***event ~ age + log(full_bmi) + diabetes_new + ethnic1 + educ_new + full_smoker,*** which has an AIC value of 1988.643, which decreases AIC of the main effect model (1993.8) by more than 2 units and is therefore considered a better model.

We further tried adding a transformation using the two other numerical variables after using a log transformation of ***full_bmi*** one at a time. For models with two transformations, taking square root of ***age*** and log ***full_bmi*** will give the model with the lowest AIC, with the model being: ***event ~ sqrt(age) + log(full_bmi) + diabetes_new + ethnic1 + educ_new + full_smoker*** and the AIC value being 1987.733. If we added one more transformation on top of the model with two transformations with the lowest AIC, we will get the model: ***event ~ sqrt(age) + log(full_bmi) + exp(diabetes_new) + ethnic1 + educ_new + full_smoker***, and an AIC value of 1987.265. Since both AIC values of more complex models (1987.733 and 1987.265) are not more than 2 units less than the AIC of the model with only one transformation, we prefer the model with less complexity.

Thus, the selected model is ***event ~ age + log(full_bmi) + diabetes_new + ethnic1 + educ_new + full_smoker***. This has a relatively low AIC with relatively low complexity.

Looking specifically at the selected model, ethnic1 has two reference categories being ethnic12 and ethnic13. Since the p-value for $\beta_{ethnic13}$ is greater than 0.05, we fail to reject the null hypothesis that $\beta_{ethnic13} = 0$. There is no convincing evidence that $\beta_{ethnic13}$ is not 0. Therefore, we should remove ethnic13 from the model. All other variables have a p-value less than 0.05, so we reject the null hypotheses that those betas are 0, and there is convincing evidence that those betas are not 0 and are significant.

#	Predictors	Added pred p-values	AIC	DF	Dev	ROC AUC
	Group 1 (BMI as Categorical Variable)					
1.1	Age,BMI,D,Ed,Eth,S	-	2004.33	9	617.0817	0.8135
2.1	Age,BMI,D,Ed,Eth,S, EthAge	0.0237	2000.846	11	624.5658	0.8151
3.1	Age,BMI,exp(D),Ed,Eth,S, EthAge	<0.0001	2000.676	11	624.7335	0.8152
	Group 2 (BMI as Numerical Variable)					
1.2	Age,BMI,D,Ed,Eth,S	-	1993.796	7	623.6152	0.8156
2.2	Age,log(BMI),D,Ed,Eth,S	<0.0001	1988.643	7	628.7686	0.8160
3.2	sqrt(Age),log(BMI),D,Ed,Eth,S	<0.0001	1987.733	7	629.6784	0.8157
4.2	sqrt(Age),log(BMI),exp(D),Ed,Eth,S	<0.0001	1987.265	7	630.1462	0.8158

Table 3: Model comparison

Table 3 provides model considered during final model selection step. Predictor abbreviations: Age, BMI, D(Diabetes), Ed(Education), Eth (Ethnicity), S(Smoker), and EthAge(Ethnicity*Age). There are two groups of models: Group 1 treats BMI as Categorical Variable and Group 2 treats BMI as Numerical Variable. In each group, each model adds one new variable at a time, so "Added pred p-values" column indicates if its categories are significant(green). The AIC values are listed for all predicators created by dummy coding for the added variable. AIC values also give the goodness of fit of each model.

4.3 Final Model and Findings

As a result, our final model includes Age, BMI, Diabetes, Ethnic12, Education, and Smoker.

$$logit(\pi_E) = \beta_0 + \beta_{age} \times Age + \beta_{log\_bmi} \times log(BMI) + \beta_{diabetes} \times Diabetes + \beta_{ethnic12} \times I_{ethnic12} + \beta_{educ} \times I_{educ} + \beta_{smoker} \times I_{smoker}$$

The numerical variables are defined as followed:

$\boldsymbol{Age(variable: age)}$ = Age of participants in years

$\boldsymbol{BMI(variable: full\_bmi)}$ = Body Mass Index measured in kg/m^2

$\boldsymbol{Diabetes(variable: diabetes\_new)}$ = 0 for not having diabetes; 1 for being on the borderline between having diabetes and not having diabetes; 2 for having diabetes

The categorical variables are defined as followed:

$\boldsymbol{I_{ethnic12}(variable: ethnic12)}$ = 0 when the participant's ethnicity is White; 1 when the participant's ethnicity is Black

$\boldsymbol{I_{educ}(variable: educ\_new)}$ = 0 for others (participant being in 9-11th grade, high school graduate, some college or AA degree, and college grad and above; 1 for when the participant is less than 9th grade

$\boldsymbol{I_{smoker}(variable: full\_smoker)}$ = 1 if the participant is a smoker; 2 if the participant is not a smoker

$$logit(\pi_E) = -10.65 + 0.07 \times Age + 2.08 \times \log(BMI) + 0.46 \times Diabetes + 0.56 \times I_{ethnic12} - 0.62 \times I_{educ} - 0.29 \times I_{smoker}$$

4.4 Assessment of the Overall Goodness-of-fit of the Model

Performing a goodness of fit on the final model, the G^2 value is $2601.4 - 1972.6 = 628.8$. Since the this is distributed as a chi-square with 6 degrees of freedom, the p-value associated with this Chi-squared value and degrees of freedom is less than 0.0001. Because the p-value is so low, we would reject the null hypothesis and shows that the more complex model provides a significantly better fit to the data than the null model. Also, since the critical value is 12.592, which is smaller than the test statistics $G^2 = 628.8$, we should reject the null hypothesis that the model does not have good fit and there is convincing evidence that the model has a good fit.

Variable	Estimate	Odds Ratio	Odds Ratio 95% CI	p-value
β_0, Intercept	-10.65	0.000024	(0.0000039,0.00014)	<0.0001
β_{Age}, Age	0.07	1.07	(1.06,1.08)	<0.0001
β_{BMI}, log(BMI)	2.08	8.04	(5.00,13.05)	<0.0001
β_D, Diabetes	0.46	1.58	(1.36,1.85)	<0.0001
β_{Eth}, Ethnicity	0.56	1.76	(1.38,2.25)	<0.0001
β_{Ed}, Education	-0.62	0.536	(0.352,0.809)	0.003
β_S, Smoker	-0.29	0.745	(0.574,0.966)	0.026

Table 4: coefficients statistics

Table 4 provides Coefficients estimates, odds ratio estimates, odds ratio 95% confidence intervals, and p-values for our logistic model. Odds ratio confidence intervals are obtained from the 95% Wald confidence intervals for the coefficients. P-values are obtained from the Wald test.

Classification Table

	Actual	
Predicted	No	Yes
No	869	267
Yes	229	543

Table 5: Classification Table

We created a classification table to determine the specificity and sensitivity of our model, which is used to create the ROC curve. From the ROC curve, we can see that AUC = 0.82, which is relatively high, showing that our model does have a good fit.

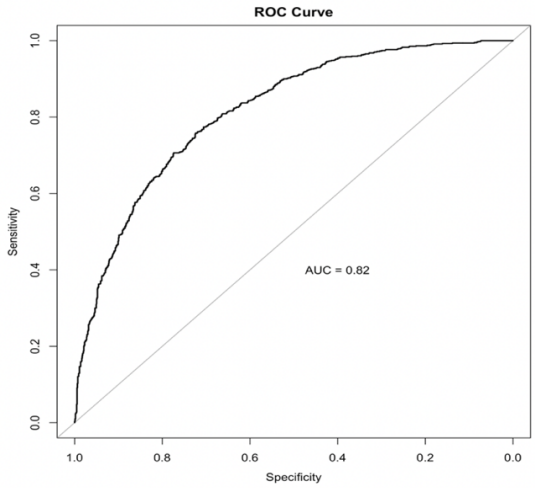

Figure 6: ROC where AUC = 0.82

From the classification table, we can see that the error rate is 26.0%, which is acceptable. Specificity is 79.1% and sensitivity is 67.0%. Although we would want both to be close to 100%, the model is still reliable in the sense that it will not produce too many false positives (false positive rate = 20.9%) and false negatives. (false negative rate = 33.0%)

We can observe the sensitivity and specificity from the ROC graph (Figure 6). The area under the curve is 0.82, which indicates that the accuracy of the predictions if the model was used would be good.

4.5 Success Probabilities of Events vs Age with Consideration of Smoking Habits

To find the probability of success, we must refer to the final logit model:

$$logit(\pi) = -10.65 + 0.07 \times Age + 2.08 \times log(BMI) + 0.46 \times Diabetes + 0.56 \times I_{ethnic12} - 0.62 \times I_{educ} - 0.29 \times I_{smoker}$$

π is the probability of success where logit(π) is log(π / (1- π)). Therefore, we can get π using the following formula:

$$\pi = \frac{logit(\pi)}{1 + logit(\pi)}$$

$$= \frac{-10.65 + 0.07 \times Age + 2.08 \times log(BMI) + 0.46 \times Diabetes + 0.56 \times I_{ethnic12} - 0.62 \times I_{educ} - 0.29 \times I_{smoker}}{1 - 10.65 + 0.07 \times Age + 2.08 \times log(BMI) + 0.46 \times Diabetes + 0.56 \times I_{ethnic12} - 0.62 \times I_{educ} - 0.29 \times I_{smoker}}$$

In our analysis of success probabilities, we incorporate two variables: ***age***, the numerical factor, and ***full_smoker***, the categorical variable. We compute the probability of cardiac event success across the entire spectrum of age and ***full_smoker,*** while maintaining all other variables at their mean values. The formula takes the following shape:

$$logit(\pi) = -10.65 + 0.07 * Age + 2.08 * mean(log(BMI)) + 0.46 * mean(Diabetes) + 0.56 * mean(I_{ethnic12}) + (-0.62) * mean(I_{educ}) + (-0.29) * I_{smoker}$$

$$logit(\pi) = -10.65 + 0.07 * Age + 4.08 + (-0.29) * I_{smoker}$$

Our graph depicts age on the x-axis across its entire range, while the y-axis represents the probability of cardiac event success. The two lines on the graph correspond to ***full_smoker*** = 1 (smoke) and ***full_smoker*** = 2 (do not smoke). Notably, we observe an overall higher probability of cardiac event success as individuals age. Additionally, it is evident that individuals who smoke (***full_smoker*** = 1) exhibit a greater probability of cardiac event success compared to individuals of the same age who do not smoke. The visual representation on the graph illustrates that the blue curve consistently surpasses the red curve, indicating a higher likelihood of cardiac event success.

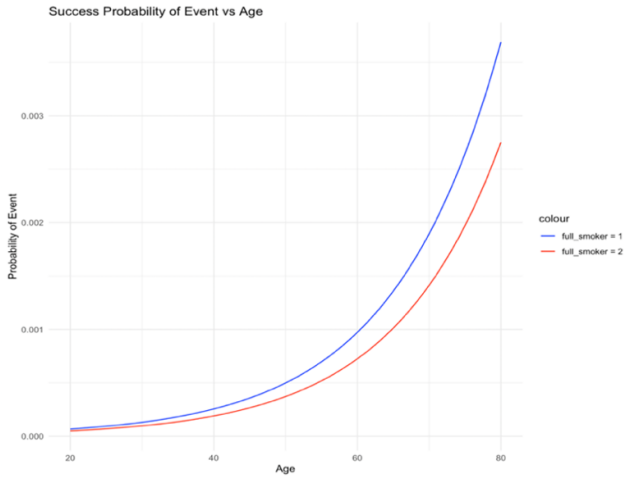

Figure 7: Probability plots for different levels of smoker.

5 Discussion

- ***age:*** For each one-year increase in age, the log-odds of a cardiac event increase by 0.067. The odds ratio of 1.07 indicates a 7.0% increase in the odds of a cardiac event for each additional year of age.
- ***log(full_bmi):*** A one-unit increase in the log-transformed BMI is associated with a 2.08 increase in the log-odds of a cardiac event. The odds ratio of 8.04 implies an increase in the odds for each one-unit increase of BMI by a multiplicative factor of 8.04.
- ***diabetes_new*****:** A one-unit increase in the diabetes category is associated with a 0.46 increase in the log-odds of a cardiac event. The odds ratio of 1.58 indicates a 58.5% increase in the odds for patients developing/already have diabetes.
- ***ethnic12*****:** Individuals with Ethnic12 = 1 (Black) have a 0.56 increase in the log-odds compared to Ethnic12 = 0 (White). The odds ratio of 1.76 suggests a 75.6% increase in the odds for the Black.
- ***educ_new*****:** Individuals with additional levels of education have a 0.62 decrease in the log-odds. The odds ratio of 0.54 implies a 46.4% decrease in the odds for additional education completed.
- ***full_smoker*****:** Individuals who smoke have a 0.29 decrease in the log-odds compared to smokers. The odds ratio of 0.75 indicates a 25.5% decrease in the odds for smokers.

Based on our prior examination of sleep duration as a predictor, we concluded that ***sleep.hrs*** is not a significant predictor of cardiac events. While we recognize that sleep duration is a significant variable in the original dataset, since the data we are working on is a randomly generated subset, we found that sleep duration has no significant impact on the incidence of cardiac events here.

Significant challenges encountered during the analysis and their resolutions:

- Addressing Missing Values: Various strategies for handling missing values were considered, including deletion, imputation using mean/median/mode, or employing model predictions.
- Model Selection Methodology: We explored multiple approaches for model selection, distinguishing between linear regression for numerical variables and general linear models and multinomial log-linear models for categorical variables.
- Metric for Model Comparison: Different metrics such as Mallow's Cp, AIC, and BIC were deliberated for comparing models. Due to AIC providing a relatively unbiased model selection criterion, it was chosen as the primary metric.
- Testing Interactions: Methods for assessing the significance of interactions between models were discussed, and the decision was made to examine interactions through grouping.
- Transformation of Variables: The consideration of whether numerical variables required transformations involved exploring techniques such as slicing-dicing plots, attempting various combinations, and fitting different transformations individually before determining the most effective approach to apply.

The model provides valuable insights into the factors influencing cardiac events. For further research, we could test the model on larger datasets to reduce as much bias as possible. We can also include more variables that the model does not currently include. For example, whether the individual has a family history of coronary disease, and the income level of the individual are both factors that might be helpful to test on. It would be intriguing to see if intervention treatment, such as weight loss procedures for obese patients, would have a reduction on the risk of cardiac event.