Olivia Martin
December 2023

Identifying Diabetes Risk Factors

EDA

Combining Datasets

I evaluated each of the three datasets separately before merging to get an idea of comparing the variables in a smaller setting. For each dataset, I looked at the distribution of the variables. Each dataset has the DIABETE3 variable which has seven possible values. Overall, there are 5000 unique people in the dataset.

Evaluating Variables with Visualizations

I evaluated the DIABETE3 variable after merging the three datasets together to have an idea of the distribution of people with diabetes. There are seven possible responses to the question (Ever told) you have diabetes". The dataset is unbalanced when it comes to incidences of people with diabetes in this dataset (Figure 1)

For the numeric data, I wanted to check that the distributions made sense. I noticed that the WEIGHT2 variable had a few very high values and was able to corroborate the values as indicators for specific instances in the data dictionary (Figure 2).

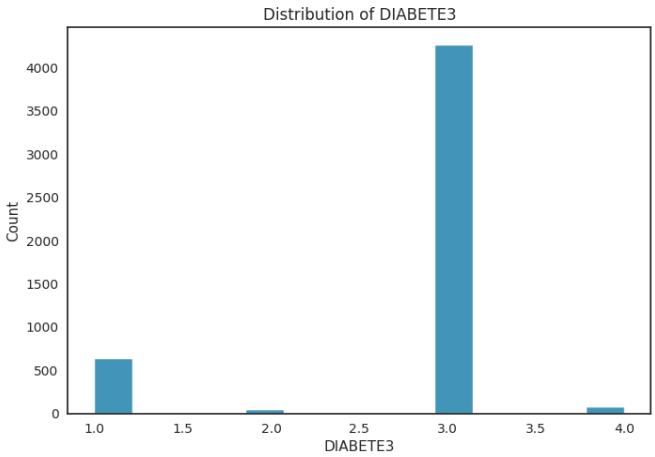

Figure 1 Distribution of DIABETE3 variable. 3.0 response indicates No, never been told I have diabetes. 1.0 indicates yes, I've been told I have diabetes.

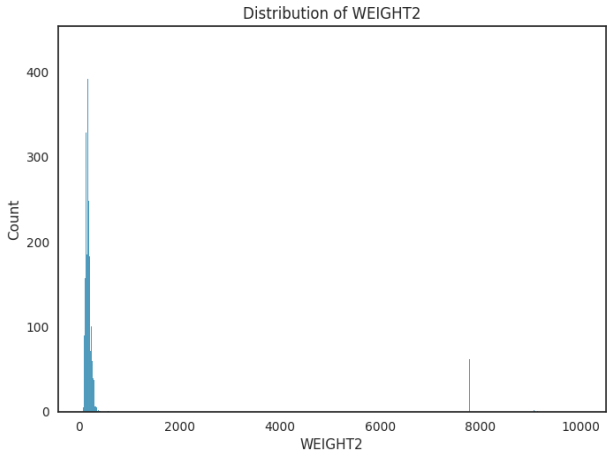

Figure 2 Evaluating distribution of WEIGHT2 variable from numeric dataset. The spike by 8000 represents the 62 responses from people where they indicated, "don't know" for their weight and the indicator value is 7777.

For the categorical data, I looked at a correlation matrix and noticed that there were potentially two repeated variables, DIEABETE3.1 and MARIATAL 1. The other categorical variables they were colinear with were DIABETE3 and MARIATAL, respectively. Similarly, _TOTINDA and EXERANY2, both ask participants about their physical activity (Figure 3).

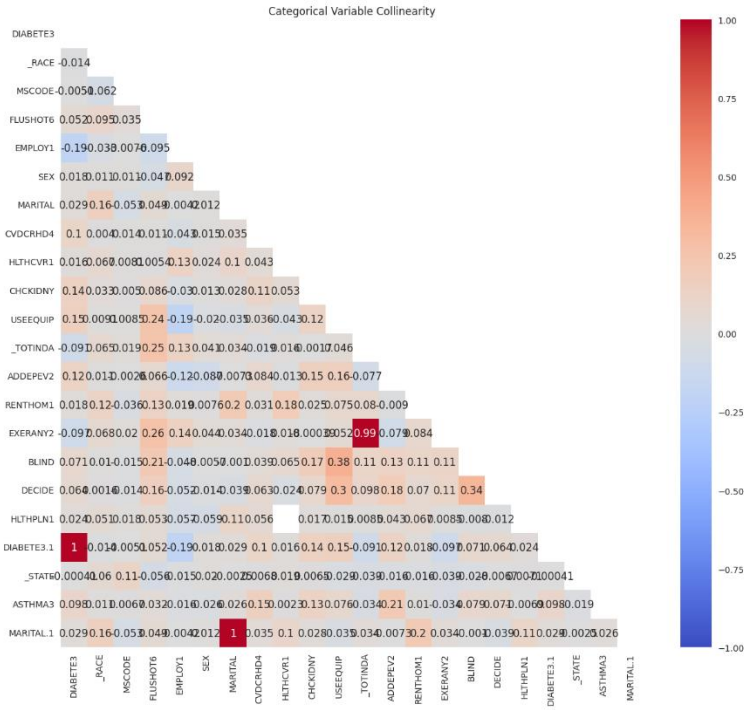

Figure 3 Correlation matrix of categorical variables. Identifying a few collinear variables related to diabetes, marital status, and physical activity.

Manage Missing Data

I decided to only keep weight data that were responses that had a number response from the participant and a response in pounds and not kilograms. Not including responses marked 7777 or I don't know, responses in kilograms, 9999 or refused responses was a strategy to remove data that could be manipulated, but in the interest of time, was removed. 272 peoples' responses were removed based on the WEIGHT2 variable. Based on the categorical collinearity matrix, DIAETE3.1, MARITAL.1, and _TOTINDA variable were dropped as they were repeats of another variable. The other important cleaning or reorganizing step was to change the predicted variable, DIABETE3, from a multi-class to binary. I recategorized having diabetes responses to be 1 and include the original responses 1.0 or Yes and 2.0 or Yes, due to pregnancy. The new no response or 0 became an aggregate of the original responses of 3.0 or no, 4.0 or pre-diabetes/borderline, 7 or not sure. This reorganization of the DIABETE3 variable does not create a balanced dataset but may simplify the classification and retain data.

Data Scale

After removing the indicator responses for the WEIGHT2 variable, the scale of the data was not so off for the remaining variables, so I did not rescale the data. From what I saw, the remaining variable responses were all within 1 to 9 as a response.

Model Development

I chose to go with a logistic regression model because of the ability to easily retrieve the coefficient of variables used in the model. I also explored using XGBoost classifier model but decided to continue looking into logistic regression model after the precision, recall, F1 score and mean AUC

score results. The logistic regression model performed only slightly better in the mean AUC score. I used precision, recall, F1 score and AUC score because the dataset is not balanced, and these metrics will take into account imbalanced datasets. The mean AUC score was determined with Stratified K Fold with 10 splits and for each model, was consistently above 0.5 indicating the models can differentiate between the two classes, with diabetes, 1, and without diabetes, 0.

	XGBoost		Logit	
	+	-	+	-
Precision	**0.88**	0.39	0.87	0.38
Recall	0.96	**0.18**	**0.98**	0.07
F1 Score	0.92	**0.24**	**0.92**	0.12
Mean AUC Score	0.77		**0.78**	

Table 1 Scoring results for XGBoost and Logit or logistic regression models.

Identifiable Risk Factors for Diabetes

Some of the top variables that predicted having diabetes were GENHLTH or respondents general health status, BMI5CAT or body mass index value, and AGEG5YR or age category of a respondent. I can base these predictors from Figure 4 where the coefficients from the logistic model are sorted and displayed in a bar chart. To what degree each of the mentioned variables affect the model in deciding if a respondent will have diabetes can be determined with some other manipulation, but manually, we could see what the GENHLTH, BMI5CAT and AGEG5YR data does look like for people with diabetes.

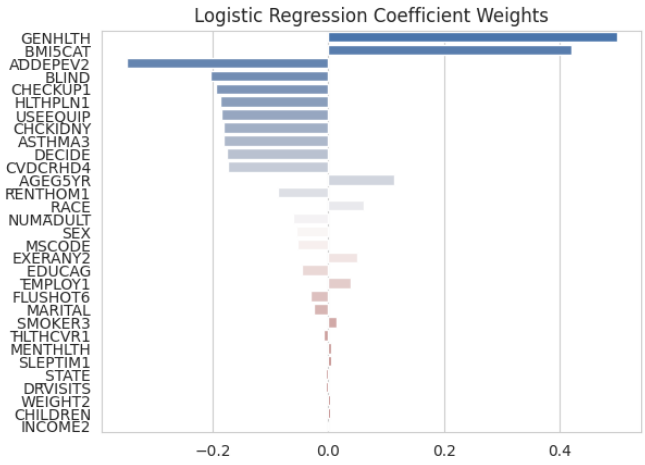

Figure 4 Final model coefficients for logistic regression in order of greatest impact to least impact.