The dataset I am using is a dataset about diabetes. There are people with or without diabetes in the dataset, and I want to research what's affecting the shift from getting diabetes to not. This would be important for anyone who wants to check if they are at risk of getting diabetes, helping people keep track of their risk level and stay in good shape. This includes researching the positive or negative impact on the results and if they are significant. After finding the significant values, I will check if they correlate with other variables and capture them. My question is, what is causing people to get diabetes? How much is it, positive or negative? Is there any data that has a close correlation between them? For the dataset of the people with a diabetes outcome (outcome = 1), is age important? Is there anything significant relating to age?

The dataset I will use includes pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, Diabetes Pedigree Function, age, and outcome. Most of it is biodata, and the outcome consists of only 1 and 0 indicating if one has diabetes or not. There are two different models in my modeling: the original model and the model where I excluded all the data that either skin thickness, insulin, or blood pressure is 0. I also want to minimize the effect of outliers using the cook's distance. Other than that, there's no normalization used in this data, as most of the function I use is not number-sensitive.

I plan to use a scatter plot matrix to first get an idea of how things are related, and if there's any correlation problem in the data, I will include their multiplies as a new column. Then, I plan to use the logistic regression to get a brief idea of which column is significant to the outcome; for all the columns, I plan to use the Variance Inflation Factor to check if the columns are correlated; if the correlation is not problematic, I will use extract the coefficients and use odds ratio to determine how much the column is affecting the outcome of the model, and if it is positive or negative. This can be used to find how much the effect is and if it is positive or negative, and it is controlled for other factors.

This is done for both models, one deleting all 0s in skin thickness, insulin, and blood pressure. The two models are compared to check how 0s (should not be 0) affect the dataset and whether the significance is true or not.

Moreover, if age is confirmed as an important factor, find if any other factor strongly correlates with age and inspect the effect on the model.

The first model is the result of a simple logistic regression model.

文本, 表格

描述已自动生成

For the matrix, here's the image. We can see many dots over the 0 position, which is not a good sign, but the correlation between the variables is quite low, which is a good sign. The 0s also largely impacted the mid-value of the variables.

图片包含 日历

描述已自动生成

To prove the model is significant overall, use the likelihood ratio test.

Here's the result

文本

描述已自动生成

The model is significant overall.

To test whether the model is a good fit for the data, use the roc curve to make sure the model fits the data well; a result close to 1 means a good fit. Here's the result. 

图示

描述已自动生成

After finding the model is in good shape, to examine the impact on the outcome, use the coefficient to investigate whether the impact is positive or negative. From the result, we can see that DiabetesPedigreeFunction, BMI, Glucose, and Pregnancies are the variables that greatly impact the model.

文本

描述已自动生成

Among those variables, DiabetesPedigreeFunction has an odds ratio of 2.57, which strongly indicates that an increase will lead to a shift towards 1, which means getting diabetes. BMI, glucose, and pregnancies all have an odds ratio bigger than 1, but they don't cause a quick shift as DiabetesPedigreeFunction does. We can also see that age has almost no impact on the data, which is somewhat counterintuitive.

Moreover, I did some pairwise tests over the dataset. Here's the results.

文本

中度可信度描述已自动生成

文本

低可信度描述已自动生成

文本

低可信度描述已自动生成

Pregnancies\_x\_Age

Glucose\_x\_DiabetesPedigreeFunction

Glucose\_x\_Age

SkinThickness\_x\_Insulin

These are the 4 significant variables. By adding them into the dataset as new columns and excluding all the 0s in the dataset, here's the new model result.

文本, 表格

描述已自动生成

We can see here that DiabetesPedigreeFunction, BMI, Glucose, and DiabetesPedigreeFunction\* Glucose are significant in the more complete dataset but less than in the original one. However, we see that using the roc curve is a better fit. It has a 3% increase in fit in the model.



When we look at the interactions between columns, we see that now there's only 1, which is significant.

文本

描述已自动生成

Considering the columns in the model, the cook's distance of 0.95 is used to ignore the outliers. We have a new model.

图表, 散点图

描述已自动生成

These are more significant to the model by changing the columns. The model fits the data very well. Compared to the last model, it has a 1% increase.



A graph with a line

Description automatically generated

Then, we can go back to check this refined model. What is more important, it has a larger effect on people who are going to have diabetes.

文本

描述已自动生成

文本

描述已自动生成

Ultimately, we can see that Age, Diabetes Pedigree Function, BMI, and Glucose mostly affect diabetes.

Age

Coefficient: Approximately 0.054 (or 5.4% per year)

Odds Ratio: 1.056

The statistical significance of age suggests it's a relevant factor in predicting diabetes. As people get older, their risk of developing diabetes increases.

Diabetes Pedigree Function

Coefficient: Approximately 2.04

Odds Ratio: 7.759

This variable had a very large odds ratio, indicating a strong association with the outcome. suggests that each one-unit increase in the Diabetes Pedigree Function is associated with an approximately 675.9% increase in the odds of having diabetes.

BMI

Coefficient: Approximately 1.75

Odds Ratio: 1.919

BMI is another significant predictor. The positive coefficient indicates that higher BMI values are associated with higher odds of diabetes, which is consistent with medical understanding of the disease. An odds ratio of approximately 1.919 implies that each one-unit increase in BMI is associated with an approximately 91.9% increase in the odds of having diabetes.

Glucose

Coefficient: Approximately 5.725

Odds Ratio: 1.059

Glucose levels are highly significant in predicting the presence of diabetes. The coefficient suggests that as glucose levels rise, so does the likelihood of diabetes. An odds ratio of approximately 1.059 means that each one-unit increase in glucose level is associated with a 5.9% increase in the odds of having diabetes.

By checking the VIF, we can confirm that there is almost no multicollinearity in the data reviewed besides the one we did for interactions.

文本

描述已自动生成

Summarize:

Age

The analysis indicates that age is a significant predictor of diabetes, with the odds of being diagnosed increasing by approximately 5.6% with each additional year. This is consistent with epidemiological studies that demonstrate an increased risk of type 2 diabetes as individuals age, possibly due to decreased insulin sensitivity and other age-related physiological changes.

Diabetes Pedigree Function

The Diabetes Pedigree Function, which provides a measure of the influence of heredity on diabetes risk, shows a remarkably high odds ratio of 7.759. This suggests a strong genetic component in the risk of developing diabetes, as a higher value, indicates a greater historical prevalence of diabetes in the individual's family. While the magnitude of the odds ratio seems quite high, it underscores the importance of genetic factors in diabetes risk, which has been well documented.

BMI

BMI is a crucial factor, with an odds ratio of 1.919 indicating that the risk of diabetes increases significantly with higher BMI levels. Numerous studies have established obesity and overweight status—often measured by BMI—as major risk factors for the development of type 2 diabetes. The relationship is thought to be mediated by factors such as insulin resistance, which is more common in individuals with higher body fat percentages.

Glucose

Finally, fasting glucose levels are a well-known marker for diabetes risk and a criterion for diagnosis. The model's coefficient for glucose confirms its strong predictive value, with a 5.9% increase in the odds of diabetes for each unit increase in glucose levels. This finding is in line with the pathophysiology of diabetes, where impaired insulin function leads to elevated blood glucose levels.

Limitations

In my dataset there are several important factors not considered such as diet and physical activity. These are usually considered important affective factors on diabetes, without them it can be bias on diagnose on the results. Moreover, the Diabetes Pedigree Function has a 600% increase, though this is specifically for diabetes, such big results are not usual.