This data analysis project contains 2 parts, data analysis on the correlation between daily stats and diabetes and real-life business application on how healthcare for employees is important - and the negative effect it has on the organizations.

**Part 1: Analysis**

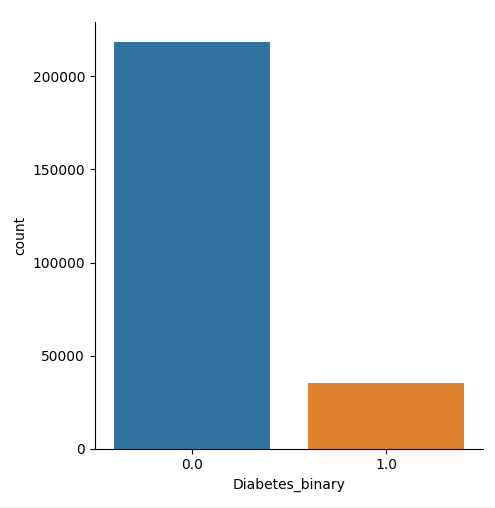
Diabetes Health Indicators Dataset: This dataset is a clean dataset of 253,680 survey responses to the CDC's BRFSS2015.

Here is the list of variables that I am going to use in this dataset:

| Variable | Description |
| --- | --- |
| Diabetes\_binary | 0 = no diabetes, 1 = prediabetes/diabetes |
| HighBP | 0 = no high BP 1 = high BP |
| HighChol | 0 = no high cholesterol 1 = high cholesterol |
| CholCheck | 0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years |
| BMI | Body Mass Index |
| Smoker | Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes |
| Stroke | You had a stroke. 0 = no 1 = yes |
| HeartDiseaseorAttack | coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes |
| PhysActivity | physical activity in past 30 days - not including job 0 = no 1 = yes |
| Fruits | Consume Fruit 1 or more times per day 0 = no 1 = yes |
| Veggies | Consume Vegetables 1 or more times per day 0 = no 1 = yes |
| HvyAlcoholConsump | (adult men >=14 drinks per week and adult women>=7 drinks per week) 0 = no 1 = yes |
| AnyHealthcare | Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes |
| NoDocbcCost | Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes |
| GenHlth | Would you say that in general, your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor |
| MentHlth | days of poor mental health scale 1-30 days |
| PhysHlth | physical illness or injury days in past 30 days scale 1-30 |
| DiffWalk | Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes |
| Sex | 0 = female 1 = male |
| Age | 13-level age category (\_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older |
| Education | Education level (EDUCA see codebook) scale 1-6 1 = Never attended school or only kindergarten 2 = elementary etc. |
| Income | Income scale (INCOME2 see codebook) scale 1-8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more |

This model developed tries to classify whether a person has “Prediabetics/Diabetes” or not. The response variable is binary (has only two possible values – in this case, 1 = “yes” or 0 = “no”). The predictor variables are 21 variables as mentioned in the table above.

After using seaborn to plot the decision variable “Diabetes\_binary”, it is clear that this dataset is not balanced ( as illustrated by the bar chart below)



In this specific scenario, the use of undersample and oversample is only done as a reference due to many reasons:

* In a real-life scenario, if the data is used to demonstrate demographic, undersample or oversample can lead to misleading results.
* This dataset is neither big enough nor small to need resampling.
* Resampling might increase model accuracy and will be referenced later on, but only the original dataset is used.

The model was split in an 80-20 ratio with the former for training and the latter for validation. After running the logistic regression, the result is produced below:

* *P < 0.05: statistically significant, often denoted with a \**
* *P < 0.01: more statistically significant, denoted with \*\**
* *P < 0.001: the most statistically significant a relationship gets, denoted with \*\*\**

| **Variables** | **P > |z|** | **Odd Ratios** | **Significant?** |
| --- | --- | --- | --- |
| HighBP | 0 | 2.134 | \*\*\* |
| HighChol | 0 | 1.778 | \*\*\* |
| CholCheck | 0 | 3.533 | \*\*\* |
| BMI | 0 | 1.062 | \*\*\* |
| Smoker | 0.352 | 0.986 |  |
| Stroke | 0 | 1.147 | \*\*\* |
| HeartDiseaseorAttack | 0 | 1.244 |  |
| PhysActivity | 0.002 | 0.951 |  |
| Fruits | 0.001 | 0.950 |  |
| Veggies | 0.017 | 0.958 | \*\* |
| HvyAlcoholConsump | 0 | 0.470 |  |
| AnyHealthcare | 0.004 | 1.113 | \* |
| NoDocbcCost | 0.622 | 1.013 |  |
| GenHlth | 0 | 1.713 | \*\*\* |
| MentHlth | 0 | 0.996 | \*\*\* |
| PhysHlth | 0 | 0.992 | \*\*\* |
| DiffWalk | 0 | 1.120 | \*\*\* |
| Sex | 0 | 1.304 | \*\*\* |
| Age | 0 | 1.129 | \*\*\* |
| Education | 0 | 0.969 | \*\*\* |
| Income | 0 | 0.946 | \*\*\* |

The results indicate that as the variables are marked with one to three asterisks *(the more asterisk mean the more statistically significant a variable is)* are statistically significant in determining whether the individual is diabetic or not *(p < 0.05 and p < 0.001, respectively).* The variables with no asterisks are not statistically significant. Using the testing data, the results indicated the following percentages of:

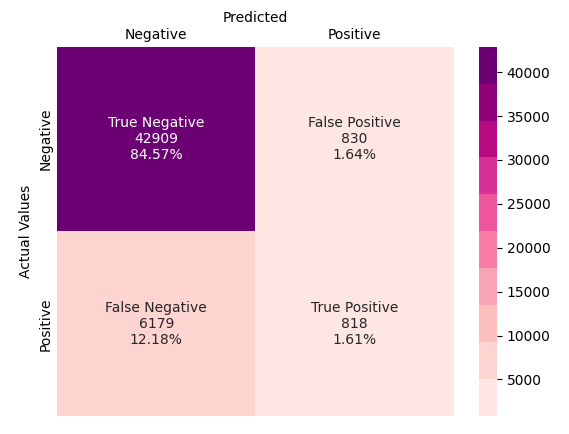
1. True Positives – accurately predicted “Yes” values

2. True Negatives – accurately predicted “No” values

3. False Positives – mislabeled “Yes” values when they are actually “No”

4. False Negatives – mislabeled “No” values when they are actually “Yes”.

Here is the confusion matrix plot to illustrate the accuracy of the model:



In total, the results of the logistic model indicate the following:

• True Negative – 84.57%

• True Positive – 1.61%

• False Negative – 12.18%

• False Positive – 1.64%

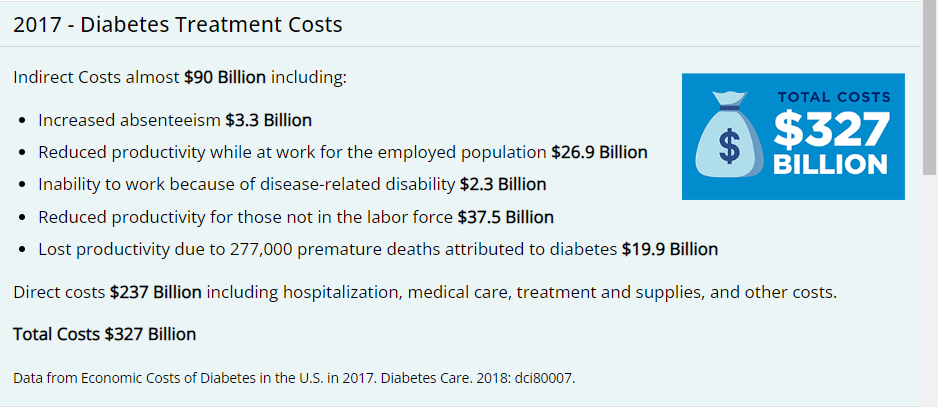
Accuracy = True Negative (TN) + True Positive (TP) = 84.57% + 1.61% = 86.18%

The model accuracy is 86.18%. This means that 86.18%, of the model, accurately predicted whether the individual is diabetic or not. In addition to the accuracy, McFadden’s Pseudo-R2 was calculated, which resulted in 20.88%. Given that McFadden’s Pseudo-R2 is considered good at around 40%, the model here did not surpass that indicating a low level of correlation in the model.

**Part 2: Business Application**

As mentioned above, whether a person is at risk of diabetes or not depends mostly on general health and mental health. Along with that, genetics and education are also playing an important role. With the increasing number of people who are diabetic, it is important for the company to acknowledge the potential negative effect it has on the operation of the organization if employees who are either diabetic or prediabetic are unaware.

According to CDC, in 2017 diabetes cost around $327 billion with at least $90 billion due to absenteeism, reduced productivity, or unable to work due to diabetic-related diseases. Other $237 billion are including hospitalization, medical care, treatment and supplies, and other costs.



*Source: cdc.gov*

It is important for companies to provide employees with healthcare benefits, so they can be aware of their health - and to help them live healthier. It is good for both employers and employees. As the logistic model suggests above - general health is the best indicator that a person will be diabetic or not, along with age and educational background. It is beneficial for employers to provide employees with a good healthcare plan that minimizes the risk of diabetes. By cutting down the budget for healthcare, employers tend to save money short term but ended up losing more money in the long term, which can either be absenteeism or human error due to health problems.