Springboard – DSC
Capstone Project 3

Predicting Which Patients Are Likely to Become Prediabetic or Diabetic in the Future

Lauren Sweeney

September 2022

Introduction

Problem Statement

Diabetes is among the most prevalent chronic diseases in the United States. It can lead to reduced quality of life and life expectancy, and can also lead to several complications, including heart disease, vision loss, amputation, and kidney disease. There is no cure for diabetes, but early diagnosis can lead to lifestyle changes and better treatments and outcomes for patients.

Our client, Very Fancy Hospital (hereinafter, "VFH"), is the largest hospital network in Northeastern Pennsylvania. VFH is also one of the leading hospitals for diabetes research, and patients from all over the country seek treatment for this chronic disease.

VFH is interested in determining whether it's possible to predict which of their current patients are likely to become diabetic in the future. This project will focus on finding ways to characterize the drivers of the two diabetic categories being studied ("nondiabetic" versus "prediabetic or diabetic") using machine learning models and by analyzing the results the models yield.

Criteria for Success

VFH wants to know which variables impact the likelihood of a patient developing prediabetes or diabetes. As a result, our class of interest is "prediabetic or diabetic." We will determine which variables are the most important for predicting prediabetes or diabetes of VFH's patients by calculating the highest Accuracy, Precision, and Recall scores by building machine learning models.

Data Sources

The dataset can be found via this link: Kaggle

The dataset includes 253,680 survey responses to the CDC's Behavioral Risk Factor Surveillance System (a health-related telephone survey that is collected annually by the CDC). It includes health data from those who either are diabetic, prediabetic, or non-diabetic (or, alternatively, only experienced gestational diabetes).

Raw Data

Target variable – Prediabetic or Diabetic

Variable Name:	Information re Variable:
Diabetes_binary	0 = no diabetes 1 = prediabetes or diabetes
HighBP	0 = no high blood pressure 1 = high blood pressure
HighChol	0 = no high cholesterol

	1 = high cholesterol
CholCheck	0 = they have not had their cholesterol checked in the last 5 years 1 = they have had their cholesterol checked in the last 5 years
BMI	Body mass index Numerical
Smoker	0 = they have not smoked at least 100 cigarettes in their entire life 1 = they have smoked at least 100 cigarettes in their entire life
Stroke	0 = they have not had a stroke in the past 1 = they have had a stroke in the past
HeartDiseaseorAttack	0 = they do not have coronary heart disease (CHD) or myocardial infarction (MI) 1 = they do have heart disease
PhysActivity	0 = they have not performed physical activity in the last 30 days (excluding their jobs) 1 = they have performed physical activity in the last 30 days
Fruits	0 = they do not consume fruit at least once per day 1 = they do consume fruit at least once per day
Veggies	0 = they do not consume vegetables at least once per day 1 = they do consume vegetables at least once per day
HvyAlcoholConsump	0 = they do not drink a high amount of alcohol per week 1 = they do drink a high amount of alcohol per week High amount = at least 14 drinks per week for men and at least 7 drinks per week for women
AnyHealthcare	0 = they do not have any kind of health care coverage 1 = they do have health care coverage
NoDocbcCost	0 = there was not a time within the past 12 months where they could not see a doctor due to cost 1 = there was a time within the past 12 months where they could not see a doctor due to cost
GenHlth	This survey question asked them to rate their general health on a 1-5 scale, where 1 =

	excellent, 2 = very good, 3 = good, 4 = fair, and 5 = poor
MentHlth	This survey question asked how many days participants experienced poor mental health in the last 30 days
PhysHlth	This survey question asked how many days participants experienced either illness or physical injury in the last 30 days
DiffWalk	0 = they do not have a serious difficulty walking or climbing stairs 1 = they do have a serious difficulty walking or climbing stairs
Sex	0 = female 1 = male
Age	This survey question asked for the ages of participants on a 1-13 scale, where 1 = 18 to 24, 2 = 25 to 29, 3 = 30 to 34, 4 = 35 to 39, 5 = 40 to 44, 6 = 45 to 49, 7 = 50 to 54, 8 = 55 to 59, 9 = 60 to 64, 10 = 65 to 69, 11 = 70 to 74, 12 = 75 to 79, and 13 = 80 or older
Education	This survey question asked for the highest level of education participants completed on a 1-6 scale, where 1 = never attended school or only attended kindergarten, 2 = grades 1 through 8, 3 = grades 9 through 11, 4 = grade 12 or GED, 5 = college 1 year to 3 years, and 6 = college 4 years or more
Income	This survey question asked for the income levels of participants on a 1-8 scale, where 1 = less than $10,000, 2 = less than $15,000, 3 = less than $20,000, 4 = less than $25,000, 5 = less than $35,000, 6 = less than $50,000, 7 = less than $75,000, and 8 = $75,000 or more

Brief Summary of Results

We reviewed and cleaned the dataset, performed exploratory data analysis, and created nine models. Based on this analysis, we were able to discern the best model for our client's needs. Since VFH could better tolerate false positives than false negatives (meaning, VFH is better able to tolerate a model stating that a patient WAS likely to become prediabetic or diabetic when they actually were not versus stating that a patient WAS NOT likely to become prediabetic or diabetic when in actuality they were), we decided to focus on the recall of each model. As a result, we determined that the best model for our client was the XGBoost with random under sampling. The most important features for that model included income, education, age, sex, and diffwalk.

For further information and details please see the notebooks developed for this project. The link can be found here: Diabetes Capstone

Data Wrangling

The first step of this project was to clean the data to make it easier to analyze and understand. Fortunately, our data had only unique values. As a result, we did not have to drop any duplicate records. We also determined that our dataset did not have any missing values.

Next, we wanted to determine what percentage of patients were prediabetic or diabetic versus nondiabetic:

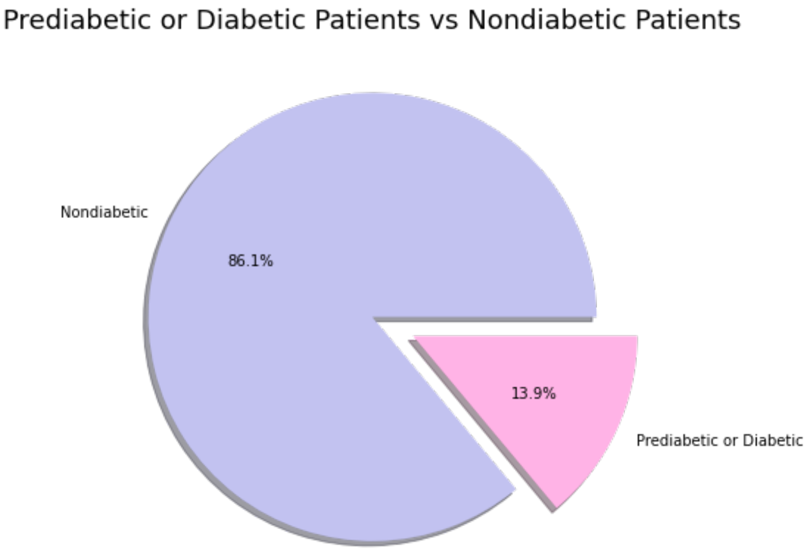
```
Nondiabetic                    218334
Prediabetic or Diabetic        35346
```

As you can see, of the 253,680 patients surveyed, approximately 13.9% were prediabetic or diabetic and 86.1% were nondiabetic.

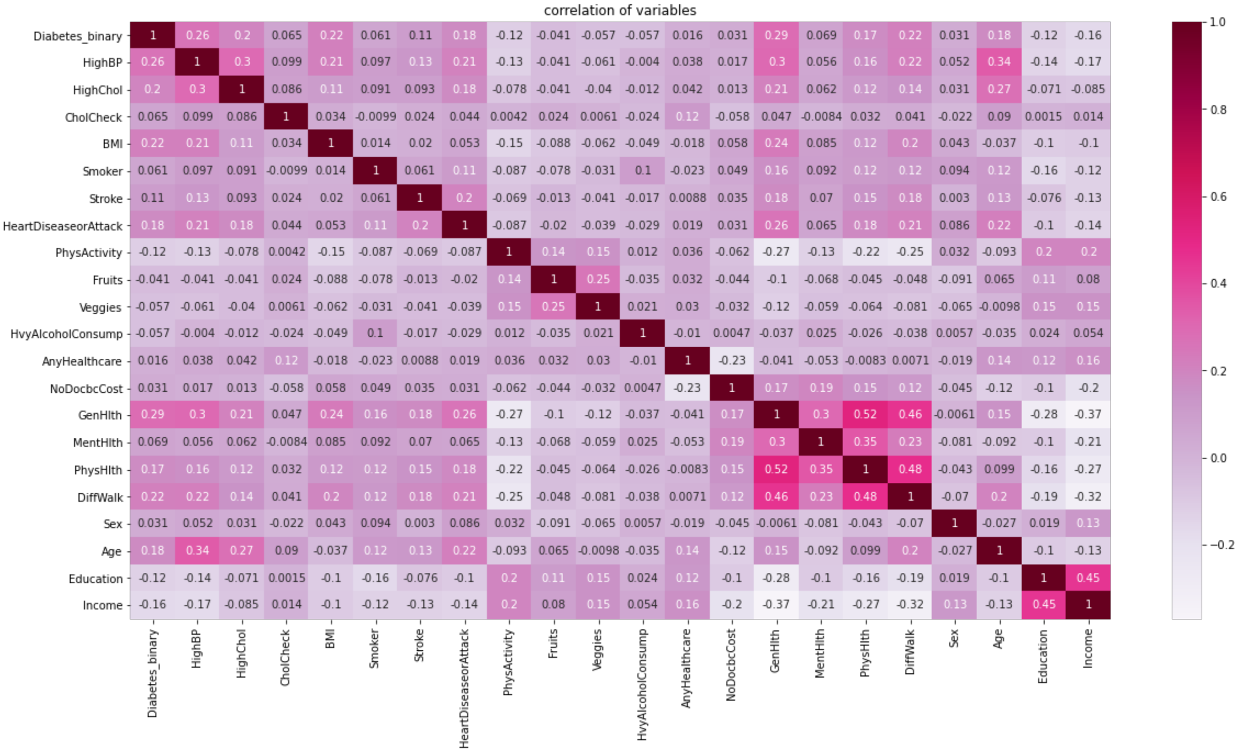
Exploratory Data Analysis

After cleaning and wrangling the dataset, the next step is to perform exploratory data analysis. This means that we want to perform initial investigations on the cleaned data to discover patterns and check hypotheses with the help of statistics and graphical representations.

In particular, we want to conduct initial explorations about the relationship, if any, between the target “prediabetic or diabetic” and the predictor variables.

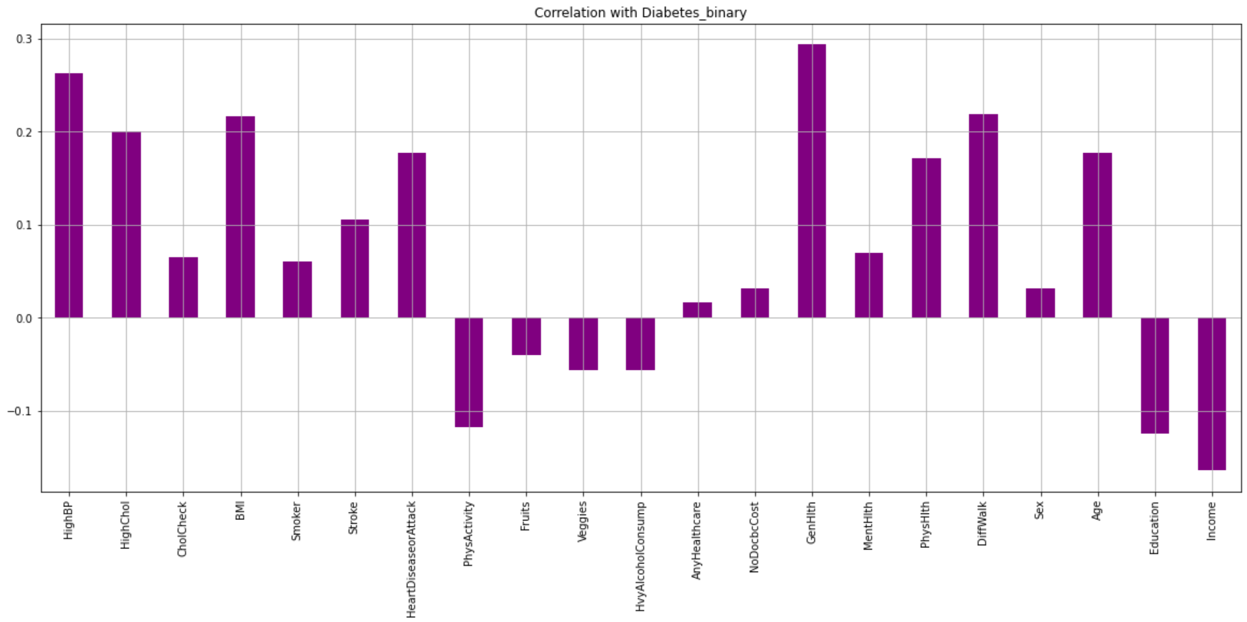


As previously discussed, we have an imbalanced dataset. Of the patients surveyed, 86.1% were nondiabetic and only 13.9% were prediabetic or diabetic. We decided to dig deeper into the data to determine whether there was any correlation between the target variable and predictor variables.



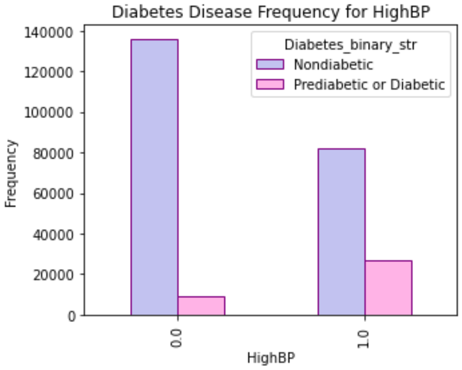
One of the first things we created was a heatmap. Heatmaps are graphical illustrations of data. This heatmap shown above is a correlation heatmap, meaning that it graphically depicts the strength of relationships between features. Darker colors depict a stronger correlation, while lighter colors depict a weaker correlation.

Our heatmap illustrated that there was a very high positive correlation between GenHlth and PhysHlth, as well as GenHlth and DiffWalk. However, there was a high negative correlation between GenHlth and Income. When we look at Diabetes_binary, we see that our target variable had a high positive correlation with GenHlth, HighBP, BMI, HighChol, and HeartDiseaseorAttack.



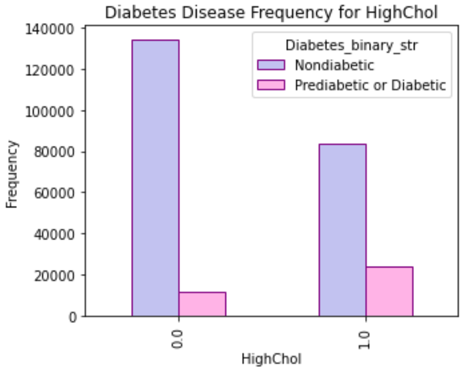
Although the heatmap was good at showing us the correlation between all the variables in the dataset, we want to specifically focus on the correlation with our target variable, "Diabetes_binary." The above plot does just that. We can see that Fruits, AnyHealthcare, NoDocbcCost, and Sex were least correlated with Diabetes_binary. However, HighBP, HighChol, BMI, HeartDiseaseorAttack, PhysActivity, GenHlth, PhysHlth, DiffWalk, Age, Education, and Income were all highly correlated with our target variable. As a result, we will focus on those variables.

First, we created a graph to illustrate the frequency of diabetes among patients with both high blood pressure and normal blood pressure:



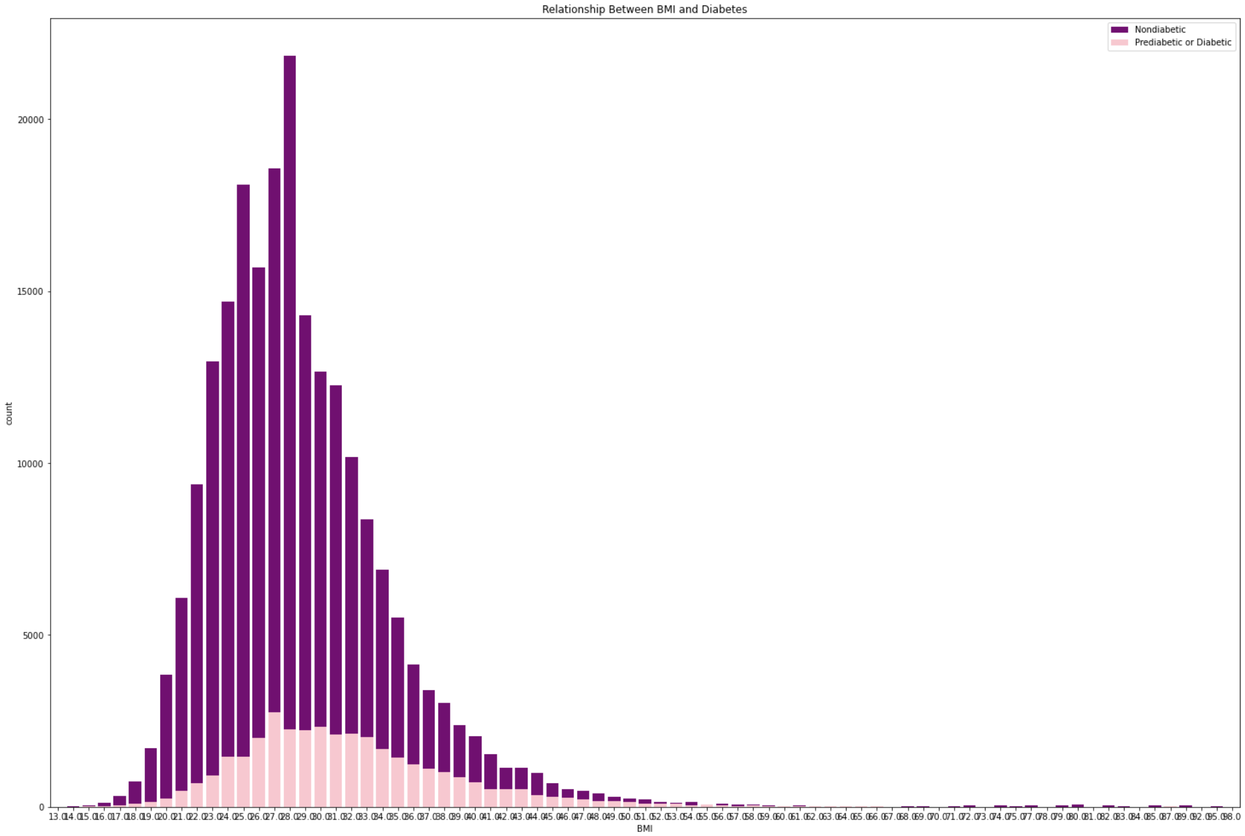
It appears that high blood pressure had a positive correlation with respect to patients who were prediabetic or diabetic. Patients who had high blood pressure also had higher rates of prediabetes or diabetes when compared to patients who did not have high blood pressure.

We then created a graph to illustrate the frequency of diabetes among patients with high cholesterol and normal cholesterol levels:

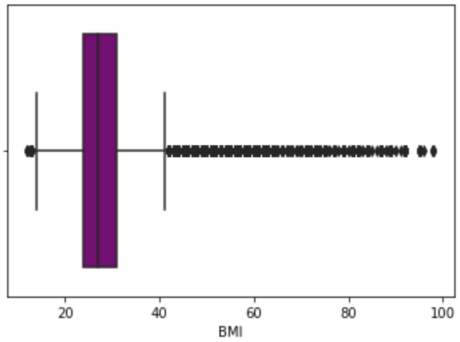


Similar to high blood pressure, high cholesterol also had an impact on whether a patient had prediabetes or diabetes. Patients who had high cholesterol were more likely to also have prediabetes or diabetes when compared to patients who did not have high cholesterol.

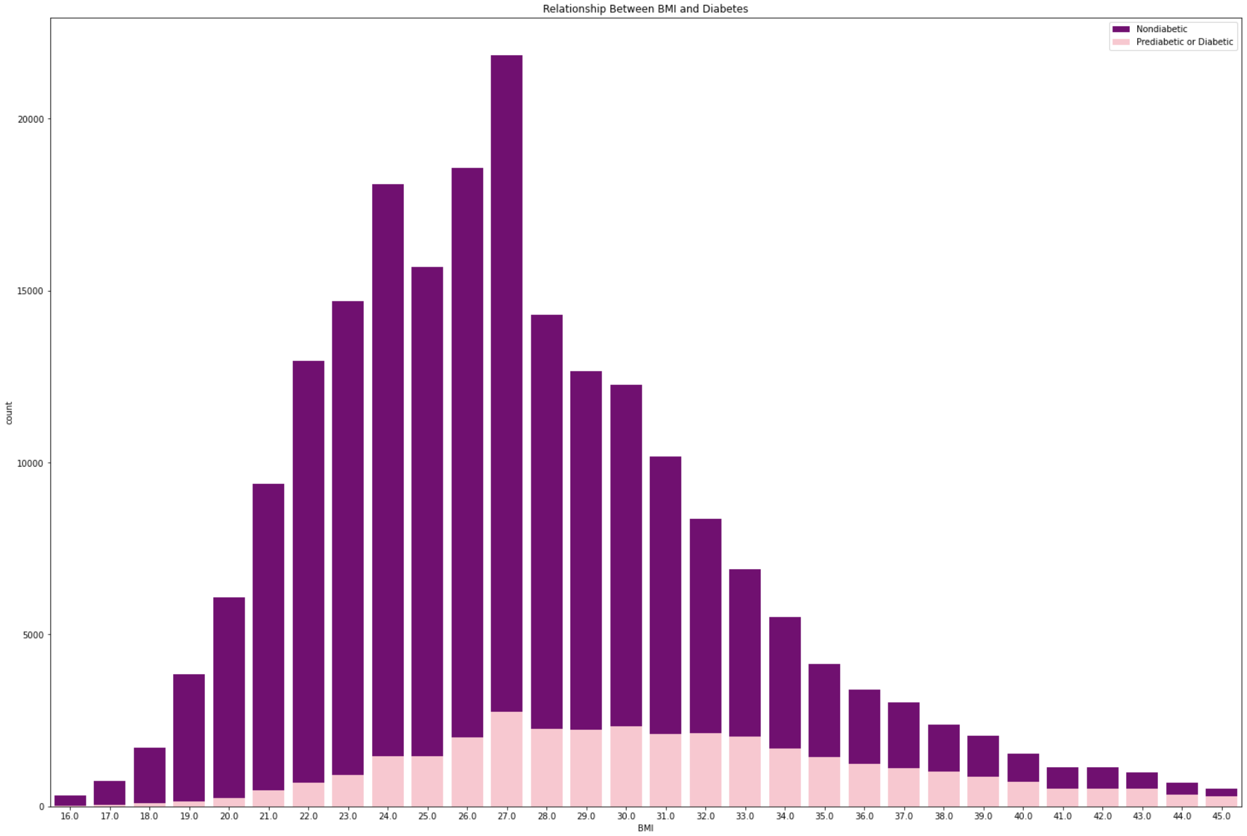
We then created a graph to illustrate the relationship between BMI and diabetes:



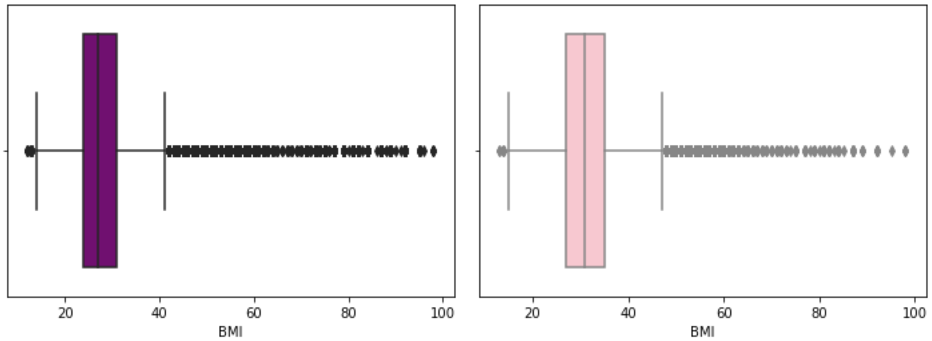
The above graph was difficult to read due to the fact that some patients had a very high BMI. So we created a boxplot to determine which BMIs we could focus our inquiry:



Based on the above boxplot, we can see that the vast majority of patient BMIs was between 15 and 45. We focused on those BMIs to make the graph easier to read:

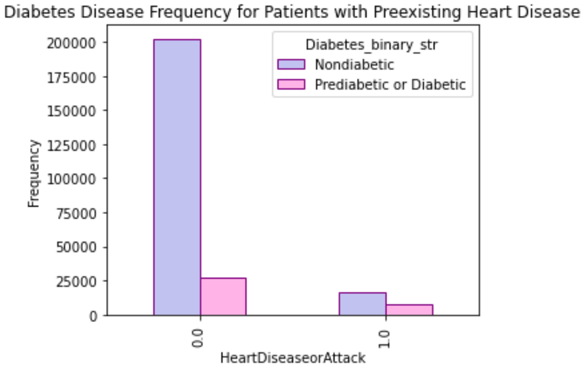


The above illustrated that, for the nondiabetic patients, the majority of their BMIs ranged from 22 to 30. Additionally, for the prediabetic or diabetic patients, the majority of their BMIs ranged from 24 to 37. We created two more boxplots to determine the median BMIs for the nondiabetic and prediabetic or diabetic patients:



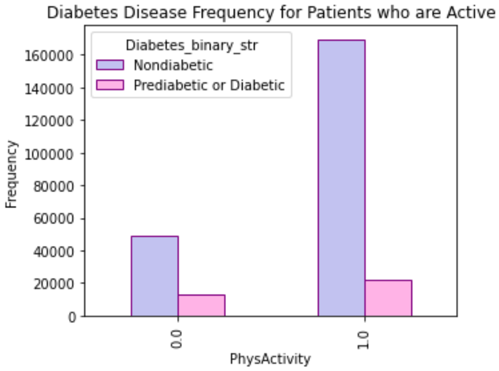
The median BMI for nondiabetic patients was approximately 27. However, the median BMI for prediabetic or diabetic patients was approximately 30. Thus, the median BMI for prediabetic or diabetic patients was larger than the median BMI for nondiabetic patients.

We then created a graph to illustrate the frequency of diabetes among patients with pre-existing heart disease:



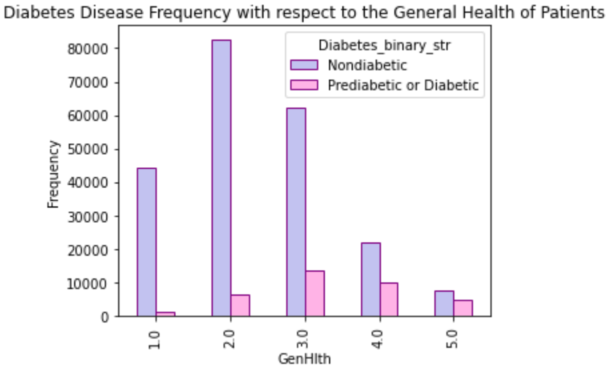
The above graph illustrated that when the rates of heart disease or attacks increase, the rates of prediabetes or diabetes also increase.

Then we created a graph to illustrate the frequency of diabetes among patients who are active.



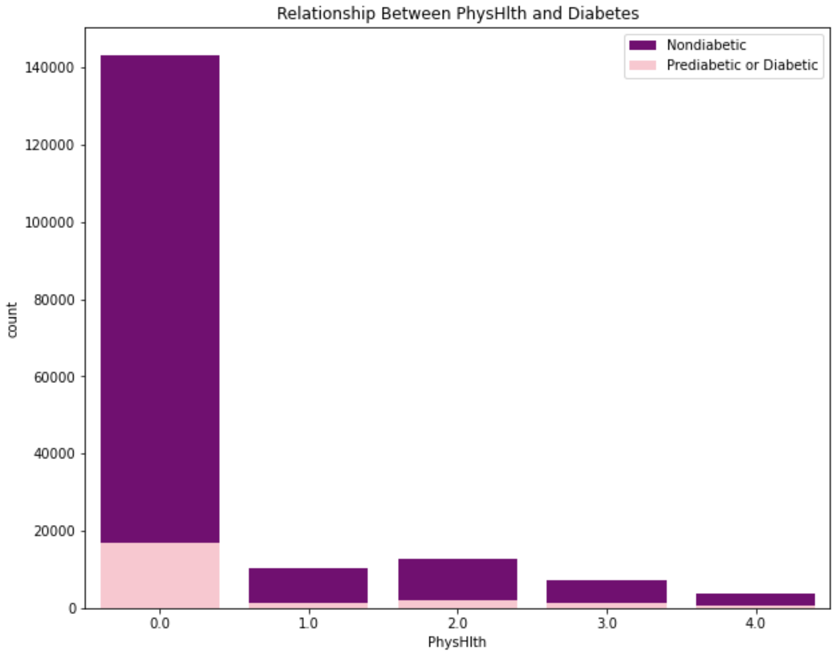
The above graph illustrated that when patients are more active, they were less likely to suffer from prediabetes or diabetes.

We also created a graph to illustrate the frequency of diabetes with respect to the general health of patients (on a 1 to 5 scale):



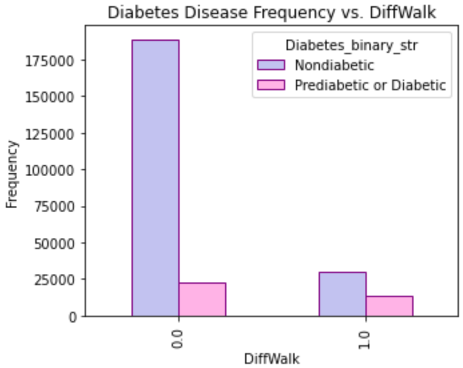
When the general health of a patient was good (closer to a 1), the rate of prediabetes or diabetes decreased.

We then created a graph to illustrate the relationship between a patient's physical health and diabetes:



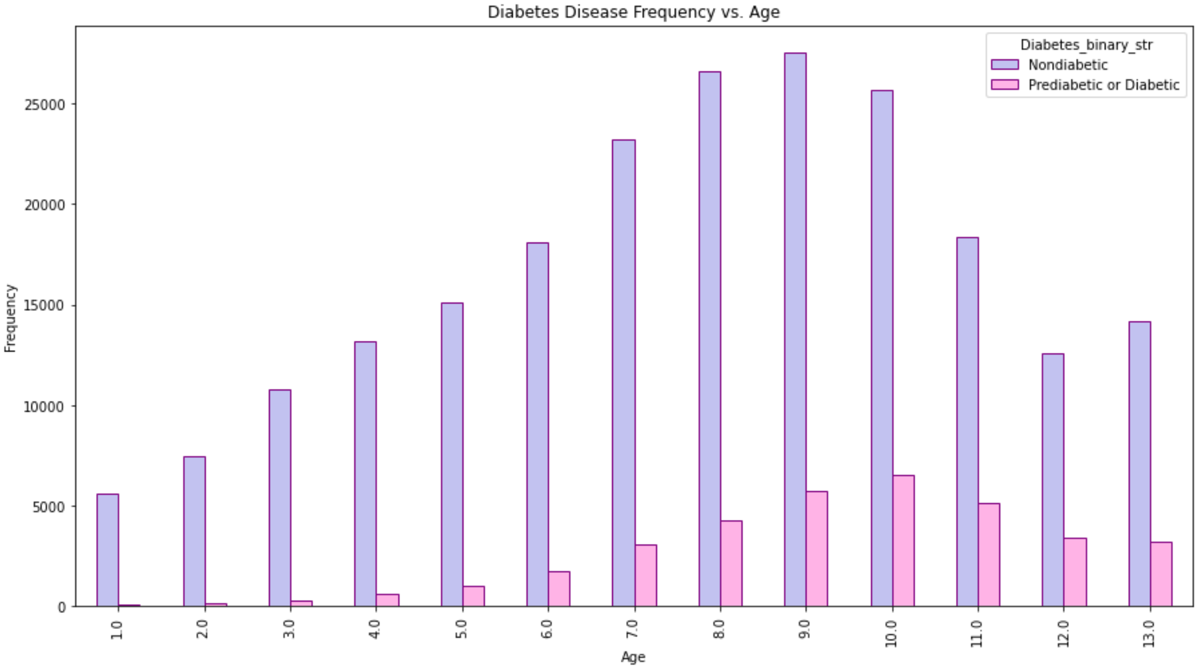
According to the above graph, it appears that the vast majority of patients experienced 0 days of illness or injury in the last 30 days. The median number of days where they experienced illness or injury was 0 for nondiabetic patients and 1 for prediabetic or diabetic patients.

We also created a graph to illustrate the frequency of diabetes among patients who have difficulty walking and do not have difficulty walking:



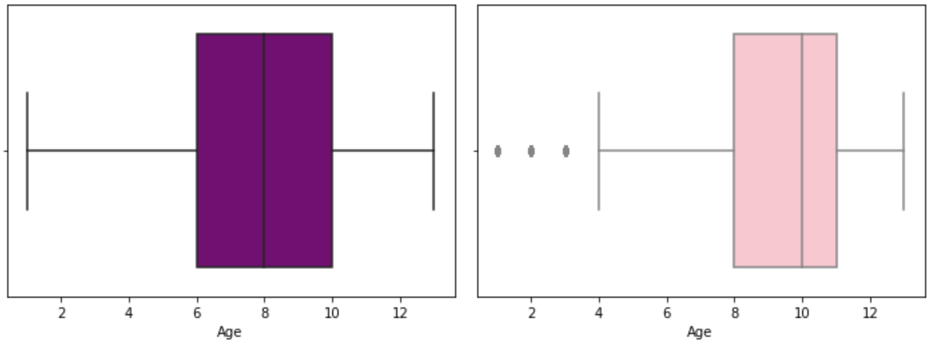
When patients have difficulty walking, they had a higher likelihood of having prediabetes or diabetes than those patients who did not have difficulty walking.

We then created a graph to illustrate the frequency of diabetes amongst various age categories:



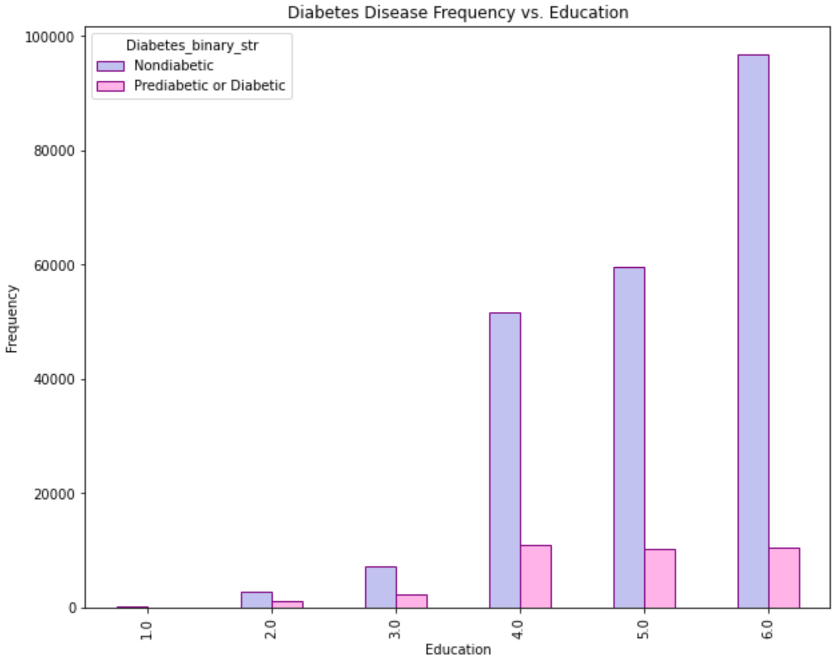
It appears that as the age of a patient increased, the likelihood of developing prediabetes or diabetes also increased.

We also wanted to create two boxplots to determine the median age category for nondiabetic and prediabetic or diabetic patients:

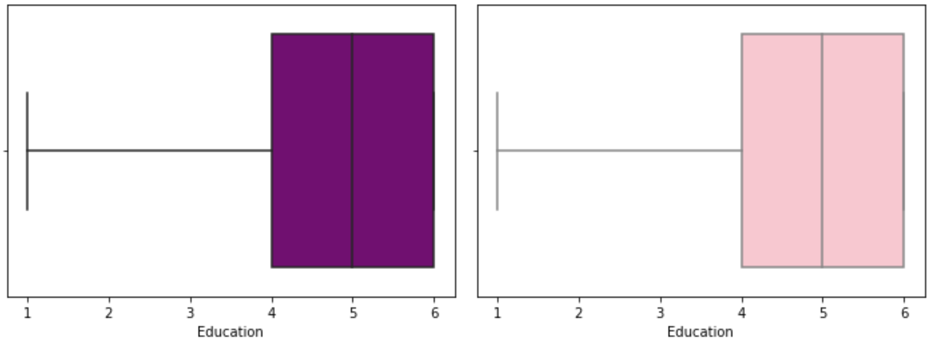


The above boxplots show us that the median age category for nondiabetic patients was an 8 (which equates to 55 to 59), while the median age category for prediabetic or diabetic patients was a 10 (which equates to 65 to 69). So older patients are more likely to be diabetic than younger patients.

Then we created a graph to illustrate the frequency of diabetes amongst a range of education categories:

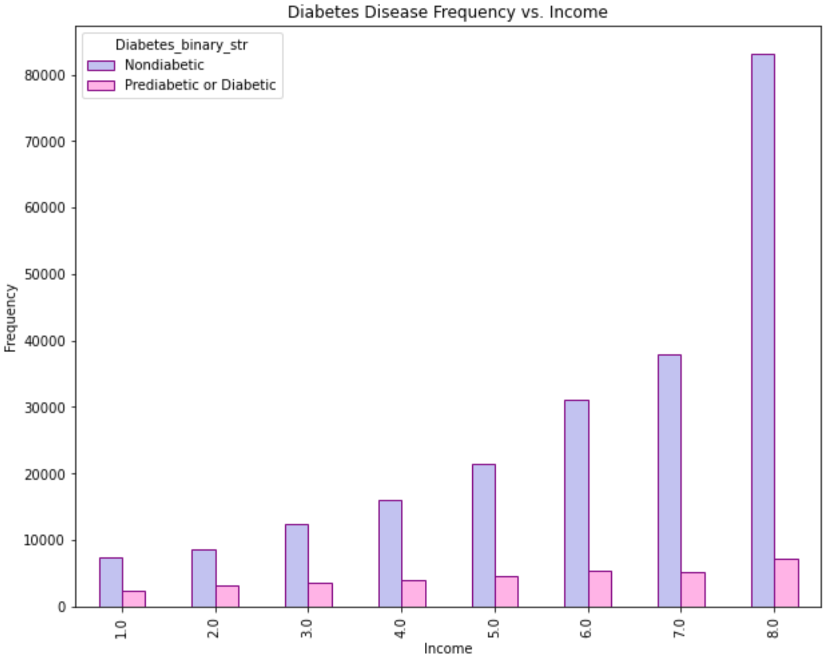


It appears that the likelihood of prediabetes or diabetes decreased as education increased. However, like before, we wanted to create boxplots to see if there was a difference in the median amount of education between nondiabetic and prediabetic or diabetic patients:

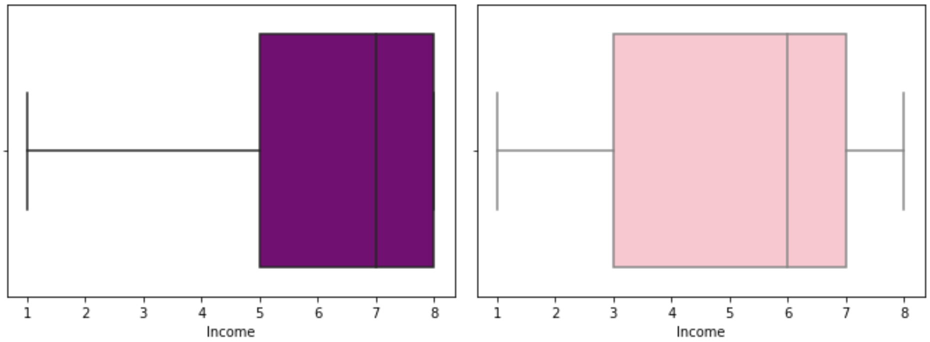


There was no difference in the amount of education for nondiabetic versus prediabetic or diabetic patients. In both cases, the median education category was a 5 (which equates to 1 to 3 years of college).

Finally, we created a graph to illustrate the frequency of diabetes amongst varying income categories:



The above indicated that as the income of a patient increased, the rate of prediabetes or diabetes decreased. We then looked at boxplots to determine if there was a difference in the median income:



The median income category for nondiabetic patients was a 7 (which equates to between $50,000 and $75,000). However, the median income category for prediabetic or diabetic patients was a 6 (which equates to between $35,000 and $50,000).

Model Selection

Pre-Processing and Training the Data

Before we could select the best model for our client's needs, we must first pre-process and train the data.

To use categorical variables in a machine learning model, we first need to represent them in a quantitative way. To do this, we would use dummy variables. However, in this particular dataset we did not need to use dummy variables as all of the data had already been represented quantitatively.

We then performed a stratified train/test split of the dataset. Stratification is done for classification machine learning problems to avoid model bias, as it helps to ensure that the target variable (in this case, "prediabetic or diabetic") will have the same (or close to the same) class proportions in our training and testing datasets.

We then built a Logistic Regression model on the training set and evaluated its performance with the testing set. After doing so, we determined the accuracy score of each:

```
[Test] Accuracy score: (y_test, y_predict_test) 0.8555792074003995


[Training] Accuracy score: (y_train, y_predict_training) 0.8560165788169573
```

Both the training accuracy and testing accuracy were very close, meaning that there was no "variance." However, the model's training accuracy was below 100%, indicating that there was some bias in this model.

We also created a classification report for both the training and testing sets:

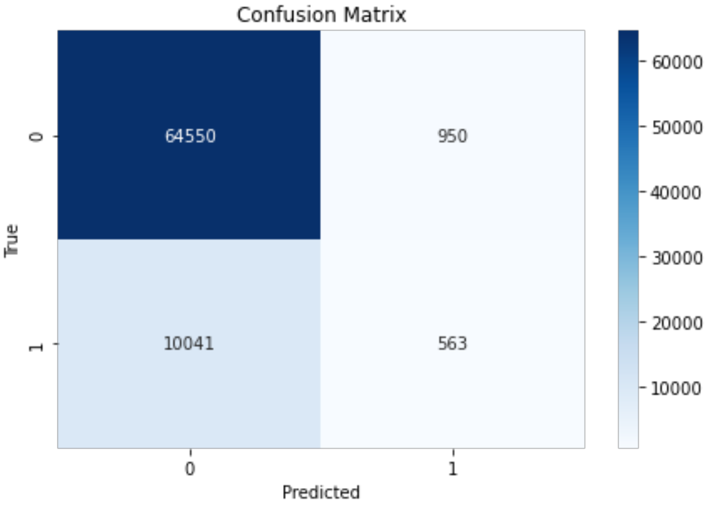
```
[Training Classification Report]
              precision    recall  f1-score   support

         0.0       0.87      0.99      0.92    152834
         1.0       0.38      0.06      0.10     24742

    accuracy                           0.86    177576
   macro avg       0.62      0.52      0.51    177576
weighted avg       0.80      0.86      0.81    177576

[Test Classification Report]
              precision    recall  f1-score   support

         0.0       0.87      0.99      0.92     65500
         1.0       0.37      0.05      0.09     10604

    accuracy                           0.86     76104
   macro avg       0.62      0.52      0.51     76104
weighted avg       0.80      0.86      0.81     76104
```

The classification report details the precision, recall, and f1-scores for both the training and testing datasets, and for both classes in each set. Precision refers to the number of class predictions that belong to the specific class, while recall refers to the number of class predictions made out of all the specific examples in the entire dataset. The f1-score refers to the harmonic mean between precision and recall, and generally speaking, higher f1-scores are better.

So what does the above illustrate? The precision between both datasets is the same for 0 (which equates to nondiabetic). However, the precision, recall, and f1-score are slightly higher for 1 (which equates to prediabetic or diabetic) for the training set.

This report also shows how much support there is for each variable. As expected, the training dataset is much larger than the testing dataset.

Next, we created a confusion matrix using the testing set:



Confusion matrices are useful for measuring recall, precision, and accuracy. It is a visual representation of actual versus predicted values. There are four elements of a confusion matrix:

True Positive – the values which were actually positive and predicted positive
False Positive – the values which were actually negative but falsely predicted as positive, also known as a Type I Error
False Negative – the values which were actually positive but falsely predicted as negative, also known as a Type II Error
True Negative – the values which were actually negative and were predicted negative

So knowing that, the above confusion matrix illustrates the following:

True Positives = 563
False Positives = 950
False Negatives = 10,041
True Negatives = 64,550

The accuracy of the confusion matrix is approximately 85.6%.

Creating Models

In total, we created nine models: (1) a Logistic Regression model using random under sampling, (2) a Logistic Regression model using SMOTE, (3) a Logistic Regression model using balanced weights of classes, (4) a Random Forest model using random under sampling, (5) a Random Forest model using SMOTE, (6) a Random Forest model using balanced weights of classes, (7) an XGBoost model using random under sampling, (8) an XGBoost model using SMOTE, and (9) an XGBoost model using balanced weights of classes.

In order to help summarize the nine models that were created, we created a table for the test set. The datasets were the same as the one used to build the first Logistic Regression model. Please note, that RUS stands for random under sampling while SMOTE stands for synthetic minority oversampling technique:

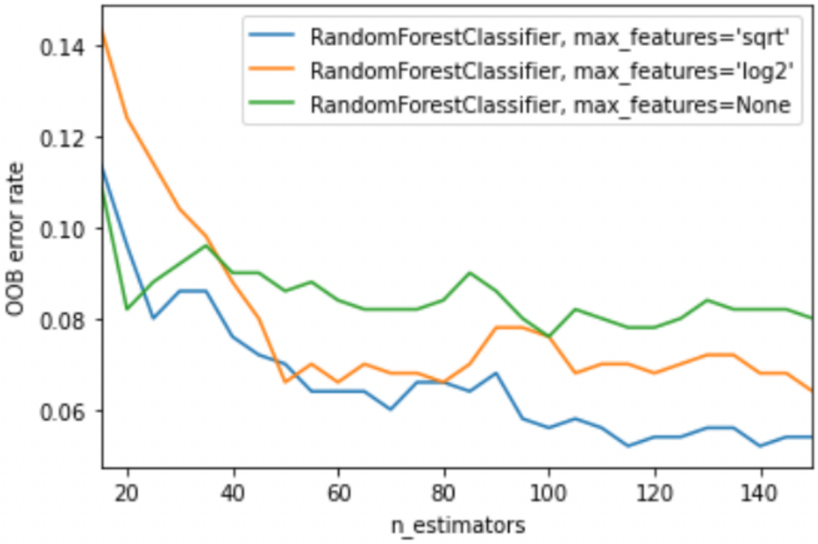
```
 Model                           |Test Precision |   Test Recall |   Test F1-Score
-----------------------------+------------------+----------------+-----------------
 Logistic Regression - RUS       |          0.31 |          0.77 |            0.44
 Logistic Regression - SMOTE     |          0.31 |          0.76 |            0.44
 Logistic Regression - Balanced |          0.31 |          0.77 |            0.44
 Random Forest - RUS             |          0.29 |          0.79 |            0.42
 Random Forest - SMOTE           |          0.47 |          0.20 |            0.28
 Random Forest - Balanced        |          0.46 |          0.16 |            0.23
 XGBoost - RUS                   |          0.30 |          0.79 |            0.43
 XGBoost - SMOTE                 |          0.55 |          0.18 |            0.28
 XGBoost - Balanced              |          0.31 |          0.77 |            0.44
```

The above table illustrates the nine models that were created and their accompanying precision, recall, and f1-scores. For the needs of our client, we decided it would be ideal to optimize recall, as our client can best tolerate false positives than false negatives. Based on the above, we see that the two best models were the Random Forest using random under sampling and the XGBoost using random under sampling, both of which had a recall of 0.79.

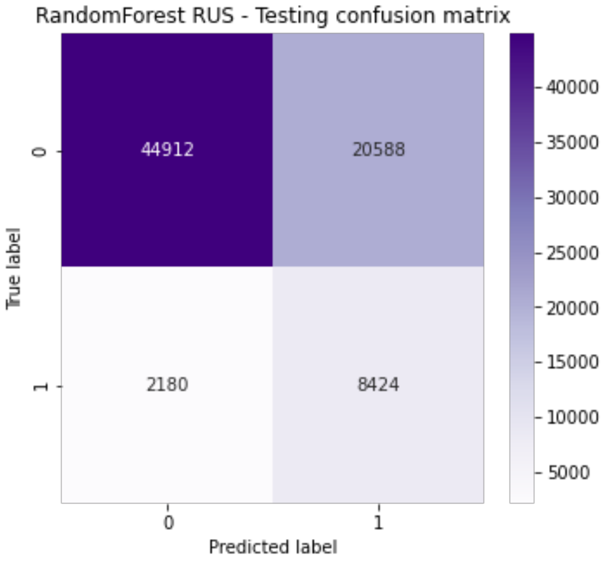
Hyperparameter Tuning

We now want to apply hyperparameter tuning to the top two models to see if their performance could be improved.

First, we worked on hyperparameter tuning the best performing Random Forest model:



Per the above graph, it appeared that we should use n_estimators = 118 and max_features = sqrt. We did so for the best performing Random Forest model, and obtained the following confusion matrix:



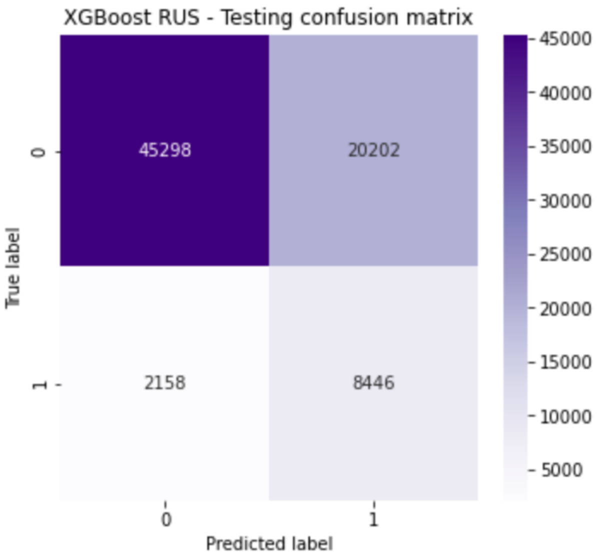
We also obtained the following classification report:

```
[Test Classification Report]
                         precision    recall  f1-score   support

            Nondiabetic       0.95      0.69      0.80     65500
Prediabetic or Diabetic       0.29      0.79      0.43     10604

               accuracy                           0.70     76104
              macro avg       0.62      0.74      0.61     76104
           weighted avg       0.86      0.70      0.75     76104
```

We then worked on hyperparameter tuning the best XGBoost model:

```
In [73]:       xg_rus_cv.best_params_

Out [73]:      {'subsample': 0.7,
               'n_estimators': 1000,
               'max_depth': 15,
               'learning_rate': 0.01,
               'colsample_bytree': 0.6,
               'colsample_bylevel': 0.7}
```

This told us which were the optimal parameters to use for our XGBoost model. After applying the above parameters, we obtained the following confusion matrix:



We also obtained the following classification report:

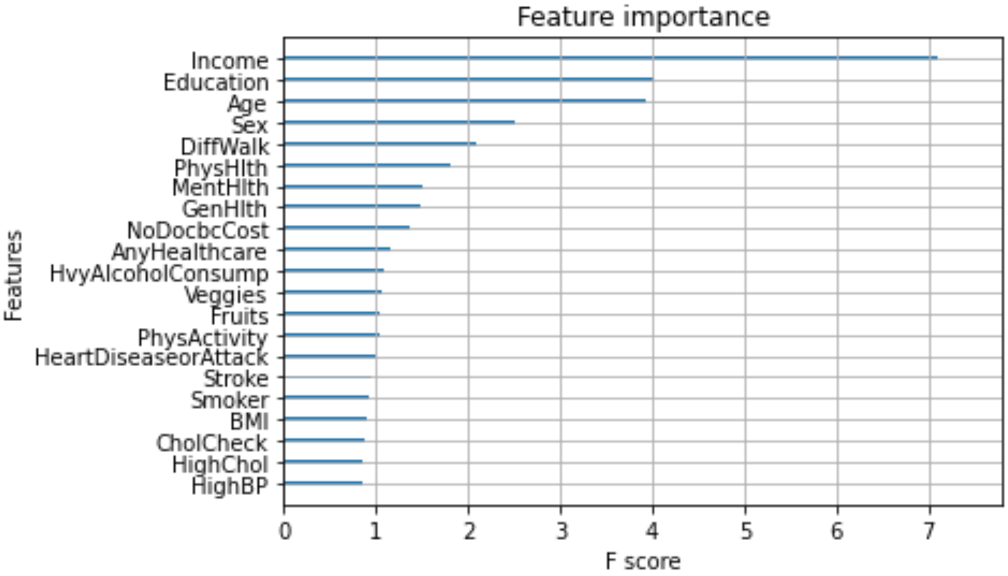
```
[Test Classification Report]
                         precision    recall  f1-score   support

            Nondiabetic       0.95      0.69      0.80     65500
Prediabetic or Diabetic       0.29      0.80      0.43     10604

               accuracy                           0.71     76104
              macro avg       0.62      0.74      0.62     76104
           weighted avg       0.86      0.71      0.75     76104
```

After hyperparameter tuning the two best models, we created another chart to show each model with their associated precision, recall, and f1-scores:

```
Model                          |Test Precision |   Test Recall |   Test F1-Score
-----------------------------+------------------+----------------+-----------------
 Logistic Regression - RUS      |            0.31 |            0.77 |            0.44
 Logistic Regression - SMOTE    |            0.31 |            0.76 |            0.44
 Logistic Regression - Balanced |            0.31 |            0.77 |            0.44
 Random Forest - RUS            |            0.29 |            0.79 |            0.43
 Random Forest - SMOTE          |            0.47 |            0.20 |            0.28
 Random Forest - Balanced       |            0.46 |            0.16 |            0.23
 XGBoost - RUS                  |            0.29 |            0.80 |            0.43
 XGBoost - SMOTE                |            0.55 |            0.18 |            0.28
 XGBoost - Balanced             |            0.31 |            0.77 |            0.44
```

Although we performed hyperparameter tuning on the two best models, we can see that this did not change their performance metrics considerably. However, the XGBoost model using random under sampling slightly outperformed the Random Forest model using random under sampling. As a result, we have determined that the best model for the needs of our client was the XGBoost model using random under sampling. This model had a test precision of 0.29, a test recall of 0.80, and a test f1-score of 0.43.

Since we identified the best model for our client, we were then able to use that model to determine which features were the most important for that model.



From the above, we can see that the five most important features (in descending order) for our chosen model were as follows:

- Income
- Education
- Age
- Sex
- DiffWalk

Conclusions & Future Work

In conclusion, our client wanted to know whether it would be possible to predict which of their patients were likely to become prediabetic or diabetic in the future. In order to make that determination, we cleaned the provided dataset, performed exploratory data analysis, and created nine models.

During exploratory data analysis, we made several discoveries:

- Our dataset was imbalanced. 86.1% of the surveyed patients were nondiabetic, and 13.9% were prediabetic or diabetic.
- High blood pressure had a positive correlation with respect to patients who were prediabetic or diabetic. Additionally, patients who had high blood pressure also had higher rates of prediabetes or diabetes when compared to patients who did not have high blood pressure.
- Similarly, high cholesterol also had an impact on whether a patient had prediabetes or diabetes. Patients who had high cholesterol were more likely to also have prediabetes or diabetes when compared to patients who did not have high cholesterol.

- Of the patients surveyed, most had a BMI between 15 and 45. Of the nondiabetic patients, the median BMI was approximately 27, while of the prediabetic or diabetic patients the median BMI was approximately 30. Thus, the median BMI was higher for prediabetic or diabetic patients than nondiabetic patients.
- When the rate of heart disease or attacks increased, the rates of prediabetes or diabetes also increased.
- Additionally, when patients were more active, they were less likely to suffer from prediabetes or diabetes.
- When the perceived general heath of a patient was good (closer to a 1 versus a 5), the rate of prediabetes or diabetes decreased.
- For nondiabetic patients, the median number of days where they experienced illness or injury in the last 30 days was 0. However, for prediabetic or diabetic patients the median was 1.
- When patients had difficulty walking they had a higher likelihood of having prediabetes or diabetes.
- The median age category for nondiabetic patients was an 8 (which equates to 55 to 59), while the median age category for prediabetic or diabetic patients was a 10 (which equates to 65 to 69). Thus, older patients were more likely to be prediabetic or diabetic than younger patients.
- As the education of a patient increased, the rate of prediabetes or diabetes decreased.
- Similarly, as the income of a patient increased, the rate of prediabetes or diabetes decreased.

After creating models with the dataset, and performing hyperparameter tuning, we were able to discern the best model for our client's needs. Since VFH can better tolerate false positives than false negatives (meaning, VFH is better able to tolerate a model stating that a patient WAS likely to become prediabetic or diabetic when they actually were not versus stating that a patient WAS NOT likely to become prediabetic or diabetic when in actuality they were), we decided to focus on the recall of each model. As a result, we determined that the best model for our client was the XGBoost with random under sampling. The most important features for that model included income, education, age, sex, and diffwalk.

Since VFH now has a way to predict whether a patient is likely to develop prediabetes or diabetes in the future, they can also use that knowledge to help their patients avoid the "inevitable." As a result, we recommend the following for future work:

- If a patient falls into a risk category based on lifestyle, encourage those patients to adopt a healthier lifestyle.
- VFH could also perform similar analysis on patients after they have enacted lifestyle changes to determine whether those changes sufficiently decreased their likelihood of developing prediabetes or diabetes.
- VFH could also request another survey in approximately 10 years to determine whether the risk categories have changed.