

Diabetes Prediction

By : Gloria Pintado

The problem

Company

For Better Health is a new organization and is recruiting members. I have been hired by this nonprofit organization that wants to help communities far from cities that need medical support.

Problem

What are some of the highest indicator that shows to have more possibilities of being at risk of having prediabetes or diabetes? Based on the highest indicators what can we in order to reduce the risk of having prediabetes or diabetes. Getting a solution will reduce not only the risk of having diabetes but also other health problems such as heart disease, vision loss, and other diseases associated with having diabetes.

Data Source

- Diabetes Health Indicator Dataset from Kaggle
- It contains 70692 rows and 22 columns (features).
- It is a balanced dataset.
- Our target variable is the diabetes binary.

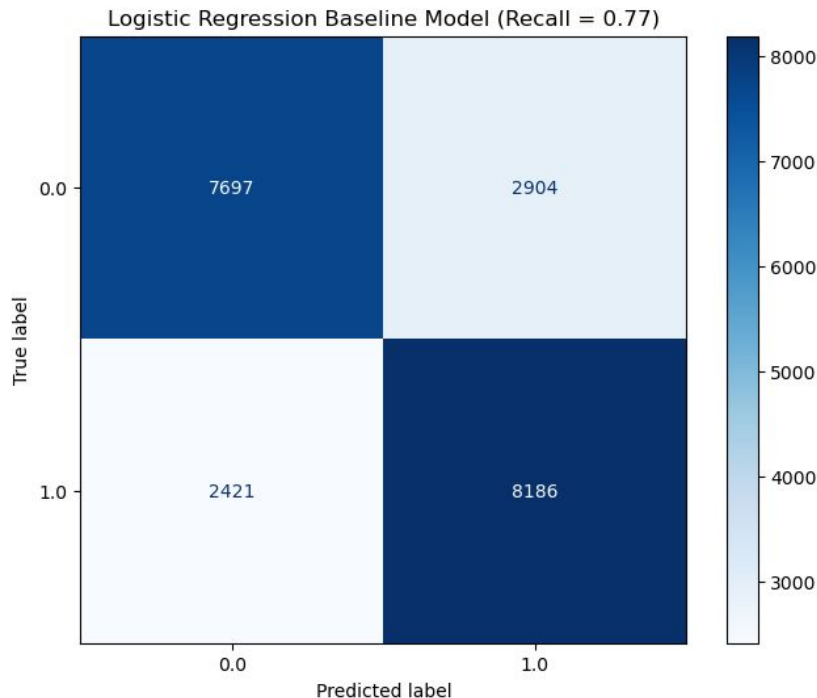
Where :

0 is for no diabetes

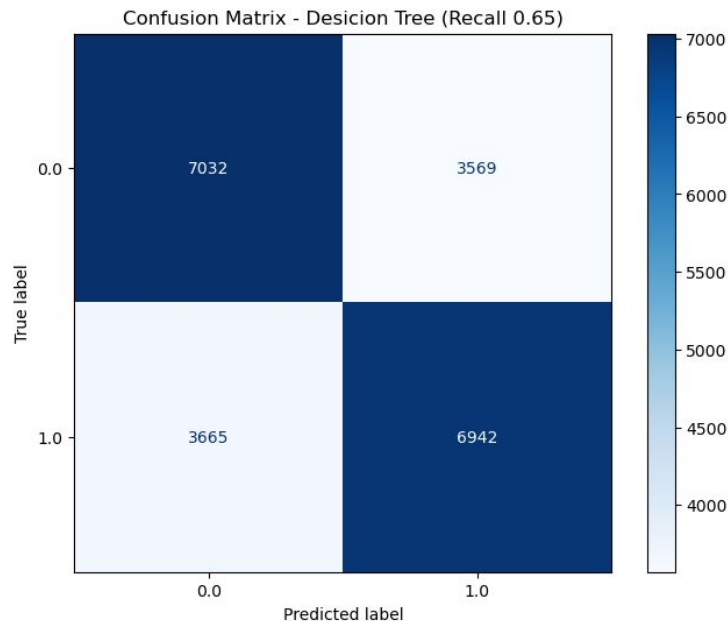
1 is for prediabetes or diabetes

Logistic Regression Baseline Model

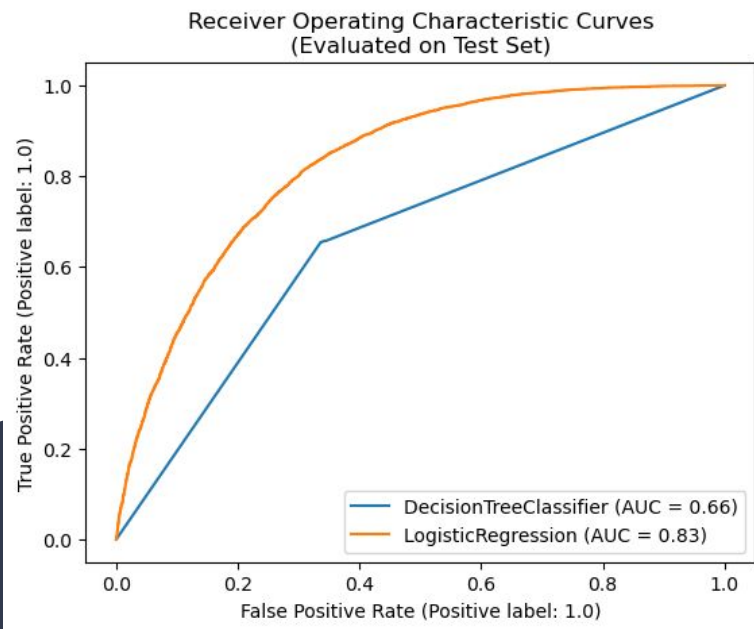
- Our metric for our baseline model is recall. We use recall because we want to reduce our false negatives.
- All features were used.



CM Decision Tree



Comparing ROC Curve



As we can see Logistic Regression was better than Decision Tree. Our next step will be tuning our hyperparameter in our Logistic Regression Model to see if it can change for better. Not doing the hyperparameter tuning in our decision tree since it did a lot worse.

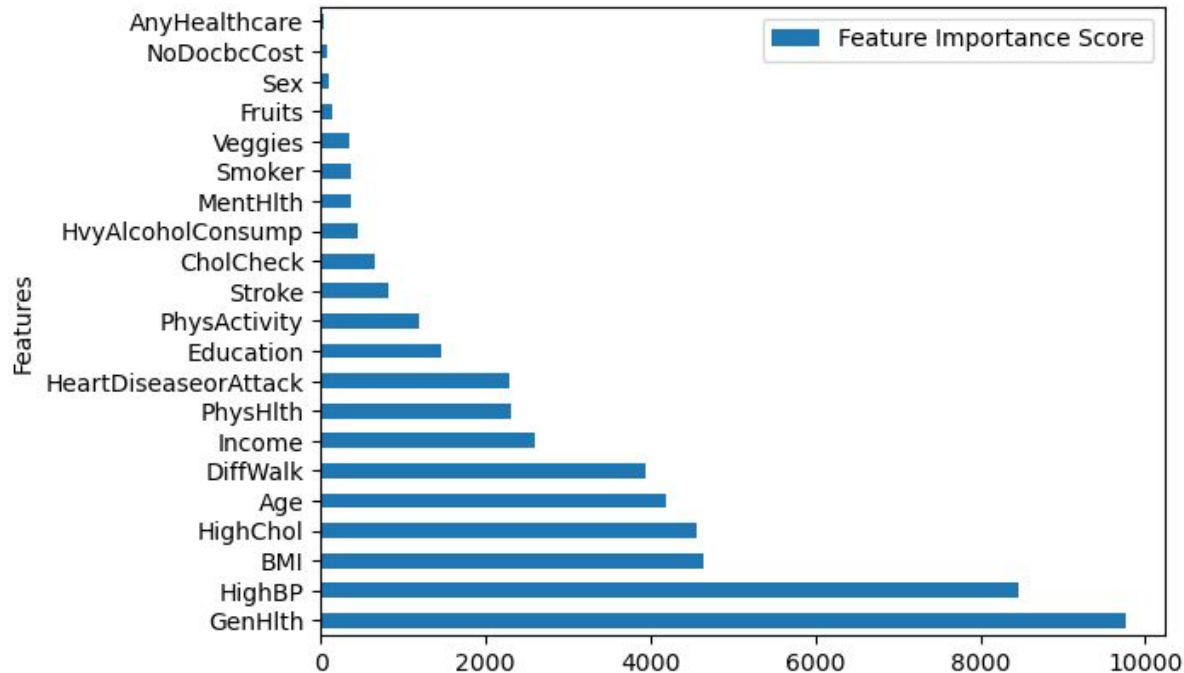
Tuning Hyperparameters for our Logistic Regression Baseline Model

```
C_list = [0.001, 0.01, 0.1, 1, 10, 100, 1000]
solver_list = ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']
solvers = []
for C in C_list:
    for solver in solver_list:
        logreg = LogisticRegression(C=C, random_state=42, solver=solver)
        logreg.fit(X_train_scaled, y_train)
```

Output: [(0.1, 'newton-cg', 0.767209668943773, 0.7717545017441312),
(0.1, 'lbfgs', 0.767209668943773, 0.7717545017441312),
(0.1, 'liblinear', 0.767209668943773, 0.7717545017441312)]

Tuning our Baseline model with the Penalties, C's, and Solvers did not have a significant impact in our model's performance

Features Importance



Solutions

- Our top ten features that should be more aware to not be at risk of having prediabetes or diabetes are GenHlth, High BP, BMI, HighChol, Age, DiffWalk, Income, PhysHlth, HeartDiseaseorAttack.
- Promote knowledge about diabetes in promotion and prevention campaigns, providing the information necessary to recognize the disease (education and income).
- Create new strategies to dissolve a sedentary lifestyle and promote new healthy habits (DiffWalk, PhysHealth).
- Do medical campaigns where they carry out checkups, such as taking blood pressure, and laboratory tests to know the cholesterol levels in the blood. Consult with nutritionists to change your diet since each person is different and must follow a special diet and know their body mass index (High BP, BMI, and high cholesterol).
- Mostly aimed at people between the ages of 35 and 40 who are the most prone to having this disease (age).

Next Steps

- Look for outliers since logistic regression is sensitive to outliers in the dataset

Contact

Gloria Pintado

Github: <https://github.com/gloriapintado>

LinkedIn : www.linkedin.com/in/gloriapintado

Email : karmely.1999@gmail.com

