Diabetes prediction - Final report

Katarzyna Rogalska, Nazarii Bihniak

April 23, 2024

1 Introduction

Treating diseases is significant, but what comes before that is spotting them. That is why our project aimed to develop a model to assist healthcare workers in predicting the patient's risk of diabetes, enabling faster diagnosis and implementation of preventive measures. We used data from an annual survey conducted in the USA in 2015, in which individuals were asked multiple questions about their health.

2 Best model development

The first step in our project was to split our dataset into 3 parts: training, validation, and test set to ensure the possibility of independent validation. Our work was validated by another group within 3 milestones: Data Analysis, Feature engineering, and Final Model Testing. At every step of our work, we received relevant feedback to help us improve our performance. Eventually, the model we selected was tested by the validation group with the test data frame, which has not been used by the development team (us).

3 Model selection and interpretation

The model we are using is called the Light Gradient Boosting Machine. LGBM is fast, memory efficient, and accurate. We have chosen it due to the best measure score.

The measure we have selected for assessment is ROC AUC score, which is based on prediction percentage, and is totally not useful when it is 0, and has the best prediction when it equals one. Testing on the validation test has shown us an 82.33% score, and the test set has shown an 84.69% score, which is considered quite good.

Also, an analysis of the most important features for prediction has been conducted to make sure the model is interpretable. Those include the patient's general health, blood pressure, cholesterol, BMI score, age, and income. Their evaluation was based on the Chi-squared test, and random forest feature selection.

4 Output interpretation for a single patient

We will quickly discuss what features had the biggest impact on patient's number x diabetes risk prediction.

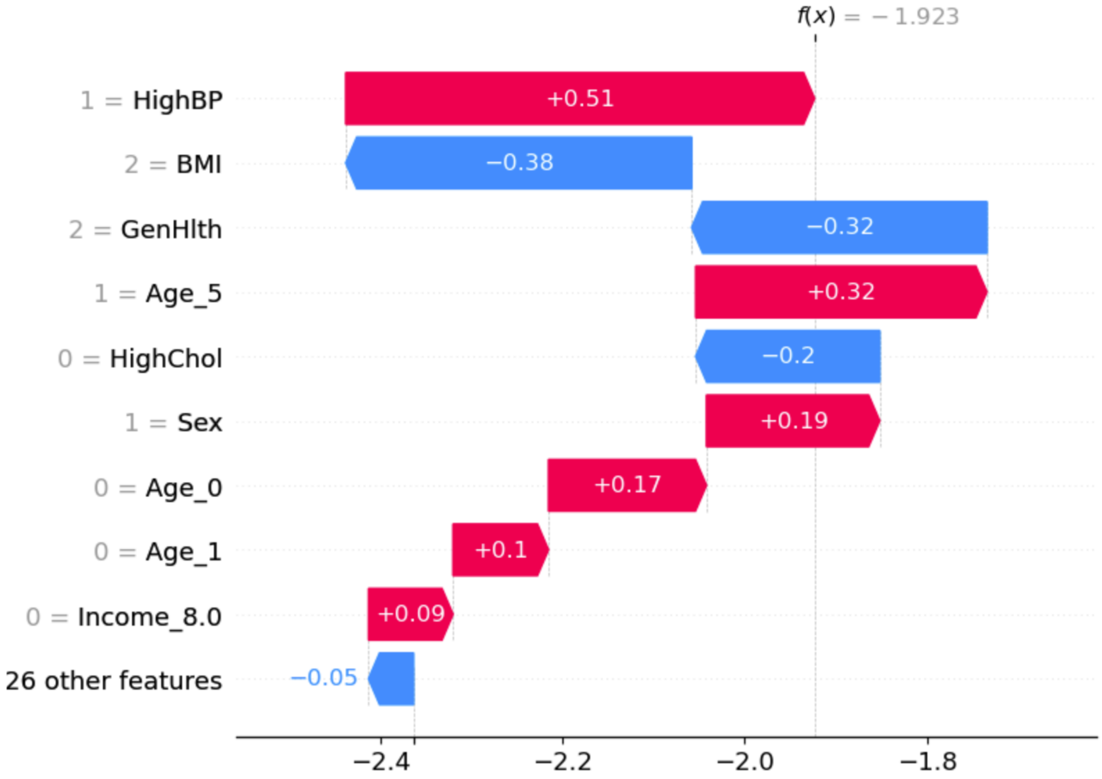

Figure 1: Model's output for patient x

From this plot, we can see that this particular patient has high blood pressure, is a male, is in 5-age group, which means he is 70-74 years old. He also has a low BMI and good general health. Features that increase the risk of diabetes are high blood pressure, age, and sex. On the other hand, features that decrease the risk are low BMI, good general health, and good cholesterol.

5 Conclusion and perspectives

The model created has shown good results, which were double-checked by an independent and impartial validation team. Furthermore, the model is possible to interpret. Thus, it can be used in healthcare institutions all over the world, and help identify diabetes as soon as possible, making people's lives longer.