MS Data Science

Interpreting Model Prediction using SHAP

Atharva Sherekar. Ayush Oturkar. Keshvi Gupta. Yuyue Sun

Summary / Abstract

This report implements the research paper "A unified approach to Interpreting Model Predictions" to demonstrate the effectiveness of the SHAP method in improving model interpretability. Using the Diabetes Health Indicator Dataset, SHAP values are assigned to each feature to identify the most important factors driving the prediction and understand how various features impact the prediction of diabetes. SHAP method is shown to be a powerful tool in enhancing the transparency and interpretability of complex models in various domains.

About SHAP

The highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models. To address this problem, Lundberg and Lee (2017) presented a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations), which assigns each feature an importance value for a particular prediction.

The premise of the paper and Shapley values comes from approaches in game theory. In a group of n players with different skillsets, how do we divide a prize so that everyone gets a fair share based on their skill set? Similarly, given a prediction, how do we most accurately measure each feature's contribution? SHAP values can improve model interpretability both globally and locally.

SHAP Mathematical formula

Shapley Value Estimation:

This method get all subsets of features S ⊆ F, where F is the set of all features, that do not contain feature $X_{\{i\}}$. Compute effect on predictions of adding $X_{\{i\}}$ to all those subsets. Aggregate all contributions to compute the marginal contribution of the feature. Thus, the Shapley values are a weighted average of all possible differences:

$$\phi_i = \sum_{S \subseteq F\{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S(x_S)\right]$$

SHAP values are a unified measure of feature importance, which are the Shapley values of a conditional expectation function of the original model.

Steps to calculate Shapley values

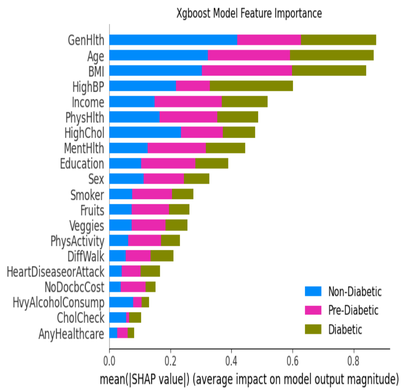
1. Create the set of all possible feature combinations (called coalitions)
2. Calculate the average model prediction
3. For each coalition, calculate the difference between the model's prediction *without* F and the average prediction.
4. For each coalition, calculate the difference between the model's prediction *with* F and the average prediction.
5. For each coalition, calculate how much F changed the model's prediction from the average (i.e., step 4 – step 3) – this is the marginal contribution of F.
6. Shapley value = the weighted average of all the values calculated in step 5 (i.e., the average of F's marginal contributions)

Data

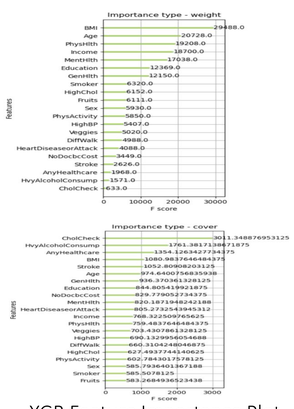
The Diabetes Health Indicators Dataset is a cleaned dataset with 253,680 survey responses. The target variable Diabetes_012 has 3 classes. 0 is for no diabetes or only during pregnancy (~83%), 1 is for pre-diabetes (~2%), and 2 is for diabetes (~15%). This dataset has 21 features capturing physical and mental health factors, medical history, eating & drinking habits, age and fitness levels. The dataset has 23,899 duplicate rows that were dropped before building XGBoost model on 80% training data.

Findings and Evaluation

Global Interpretability



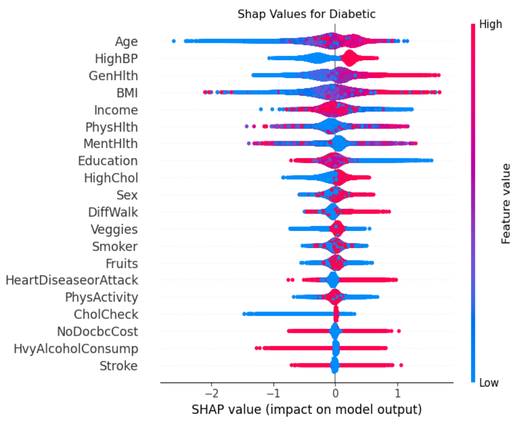
SHAP Summary Plot



XGB Feature Importance Plot

The model gives highest importance to self-reported health followed by age, BMI, HighBP in distinguishing between the three classes. Education, mental health, smoking, fruit and vegetable consumption, and physical activity significantly impact pre-diabetes prediction, while features like difficulty walking, heavy alcohol consumption & history of heart attack are not given much importance.

Also, the feature importance of SHAP is more consistent in comparison to the variability in the XGBoost interpretation plots.

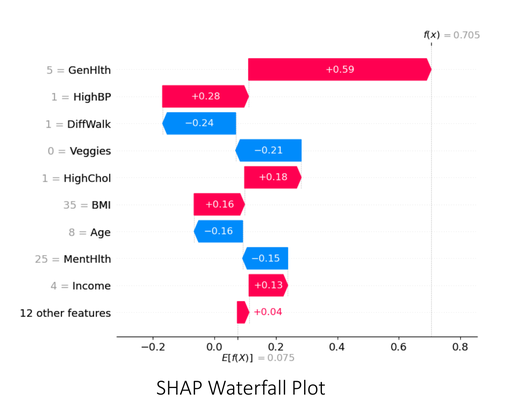


SHAP Summary Plot

- Plot shows contribution of the features for each instance (row of data)
- Age, High BP, High GenHlth, High BMI, Difficulty in walking, and history of heart attack increase SHAP values and predict a person as diabetic.
- Lower income and lower education also increase SHAP values and predict a person as diabetic, while lower age significantly reduces the SHAP value and restricts the prediction of type-2 Diabetes.

Local Interpretability

SHAP Force Plot



SHAP Waterfall Plot

- Plots shows contribution of the features for a true type-2 diabetic prediction.
- Poor self-reported health, High BP, High Cholestrol levels and BMI are pushing towards the higher risk of type-2 diabetes above base value.
- Difficulty in walking, No consumption of veggies and low age is lowering the impact towards having type-2 diabetes. This could indicate it being pushed to the pre-diabetes class.

Conclusions

The lack of a consistent approach to understanding model outputs can lead to confusion and errors, especially with the increasing number of statistical models in use today. SHAP provides a common interpretation approach for complex models, preserving consistency, and also offering insight on global and local level. Compared to other interpretation techniques like LIME, SHAP has shown better consistency in interpretation across models and data points.

However, SHAP has limitations such as computational expense, additivity assumption, potential accuracy decrease, and tabular data restriction.