415-hw4

Shuo Han

2023-05-27

Introduction

In this project, we will explore several methods for explaining the outputs of predictive models. Specifically, our focus will be on attribution methods that aim to weigh the relative importance of inputs in making predictions. While there are various approaches to consider the interactions between variables, we will concentrate on attribution methods such as LIME, Shapley Values, and SmoothGrad, avoiding the complexities associated with other methods.

Tabular Data

Feature	Type
Pregnancies	int64
Glucose	int64
BloodPressure	int64
SkinThickness	int64
Insulin	int64
BMI	float64
DiabetesPedigreeFunction	float64
Age	int64
Outcome	int64

1

The percentage of patients labeled as diabetic is 34.89%, so the the percentage of patients labels as non-diabetic is 65.11%, which is almost twice the amount of diabetic patients. So the dataset exhibits a class imbalance between the two classes, diabetic and non-diabetic. This indicates that there are more instances of non-diabetic patients in the dataset compared to diabetic patients. Thus, I will apply SMOTE to balance the data.

2 Linear model with L1

Variable	Coefficient
Pregnancies	0.003532
Glucose	0.006139
BloodPressure	-0.002102
SkinThickness	0.000000
Insulin	-0.000113

Variable	Coefficient
BMI	0.014336
DiabetesPedigreeFunction	0.000000
Age	0.005594

MSE (Lasso): 0.16626232800111837

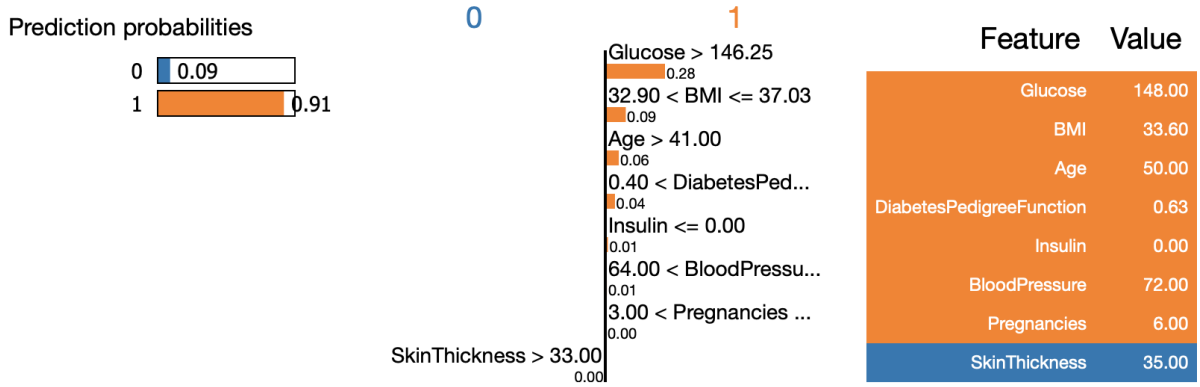
Right here, I have applied SMOTE to deal with the imbalance of the dataset. And then, I have fitted the linear model with L1 penalty here. And the selected features are Pregnancies, Glucose, BloodPressure, Insulin, BMI, and Age with all coefficients listed above. The features "BloodPressure," "SkinThickness," "Insulin," and "DiabetesPedigreeFunction" have negative coefficients; the features "Pregnancies," "Glucose," "BMI," and "Age" have positive coefficients. Based on the provided coefficients, the features that appear to be more important in predicting the outcome are "Glucose," "BMI," and "Age."

3 Random forest

Feature	Coefficient
Pregnancies	0.0714
Glucose	0.2358
BloodPressure	0.0853
SkinThickness	0.0695
Insulin	0.0813
BMI	0.1809
DiabetesPedigreeFunction	0.1311
Age	0.1447

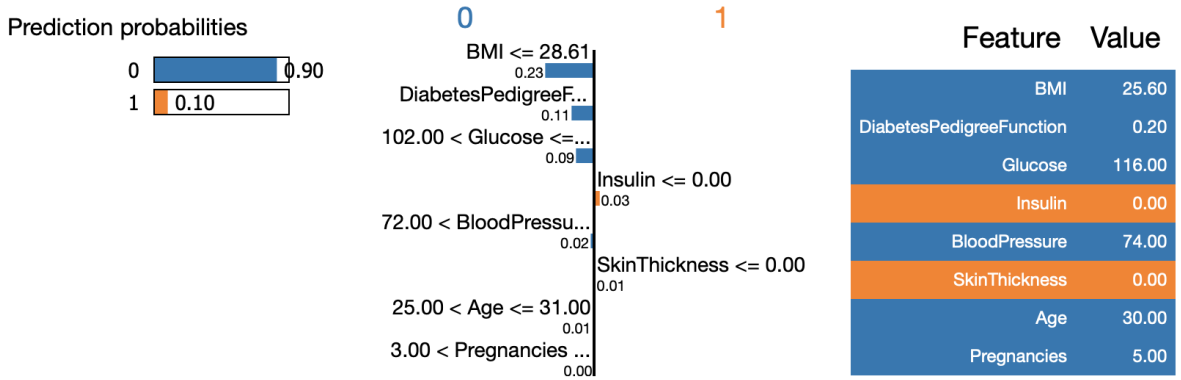
Based on the coefficients,same as the linear model, the features that appear to have a relatively higher importance in predicting diabetes are Glucose, BMI, and Age. These features have larger coefficients, indicating a stronger association with the outcome. But different from the linear model, all eight features are selected here, including Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age.

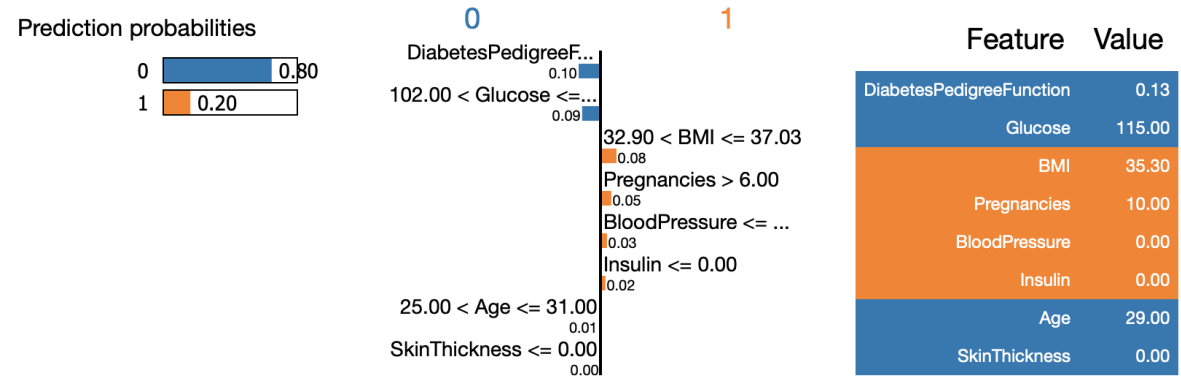
4



Right here I have chosen to use LIME on the first data point here. And we cans see that the probability of the patient being diabetic is 0.09, and the probability of the patient not being diabetic is 0.91. Glucose, BMI, and Age are 3 features of most importance for the prediction here, while SkinThickness is the least important for the prediction here.

5





Right here I have chosen to use LIME on the 6th data point here. And we cans see that the probability of the patient not being diabetic is 0.09, and the probability of the patient being diabetic is 0.1, which is an opposite result from the first datapoint. Glucose, BMI, and DiabetesPedigreeFunction are 3 features of most importance for the prediction here, while SkinThickness is the least important for the prediction here. So the features are really similar to the features for the first dataset selected. Also, I have tried one more datapoint here, it shows the same rule, so we can see LIME is fairly stable across different datapoints.

6

LIME, Forest importance, and linear model weighting are different methods for explaining feature importance. LIME generates local explanations for a specific prediction by assigning feature weights based on their contribution to the prediction. Forest importance calculates feature importance by averaging their impact across all trees in a random forest model. Linear model weighting refers to the coefficients assigned to features in a linear model. Based on the result above, in linear model, the most important features in predicting the outcome are "Glucose," "BMI," and "Age"; Glucose, BMI, and Age for random forest; and Glucose, BMI, Age, and DiabetesPedigreeFunction in LIME. Thus, we can see that all these three methods select similar features.

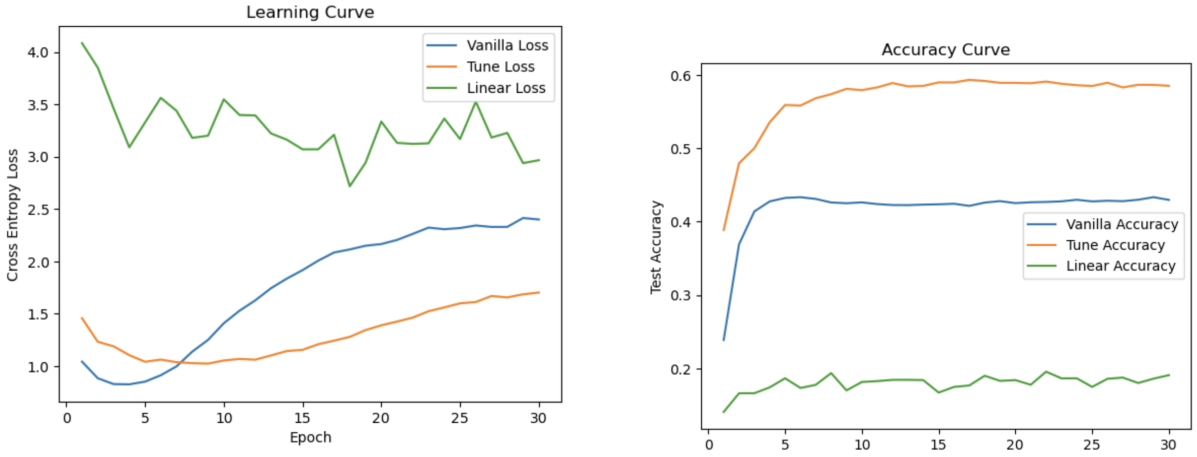
Predictive Modeling on Animal Images

7

Shape: torch.Size([45000, 3, 64, 64]) torch.Size([45000]) torch.Size([5000, 3, 64, 64]) torch.Size([5000])

With the starter code, I have loaded the data here with normalization and train_test_split. Also, the an example image from each class in the training dataset is shown above.

model visualization



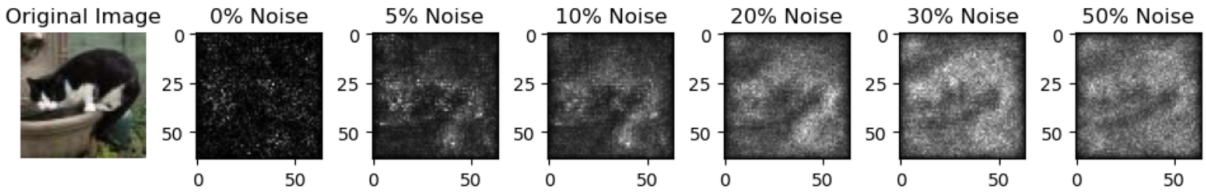
First, I have trained a linear model on the animal clasification dataset with PyTorch with the starter code. Also, I trained a vanilla deep convolutional network on the animal classifcation dataset with 2 concolutional

layers with maxpooling layers, 2 fully connected layers, and ReLU activations in 30 epochs. And I have tuned my neural network by adding batchnorm, more layers, changing the batch size and learning rate. The learning curves above illustrate the performance of different models in terms of cross entropy loss and accuracy. It is evident that the linear model exhibits the lowest accuracy and the highest validation loss. In contrast, the tuned deep convolutional network achieves the highest accuracy and the lowest validation loss. Regarding the importance of hyperparameters, it is apparent that the number of layers and learning rate play a significant role in my tuning process. These factors heavily influence the model's performance and contribute to the observed differences in accuracy and validation loss.

Feature Atribution on Animal Images

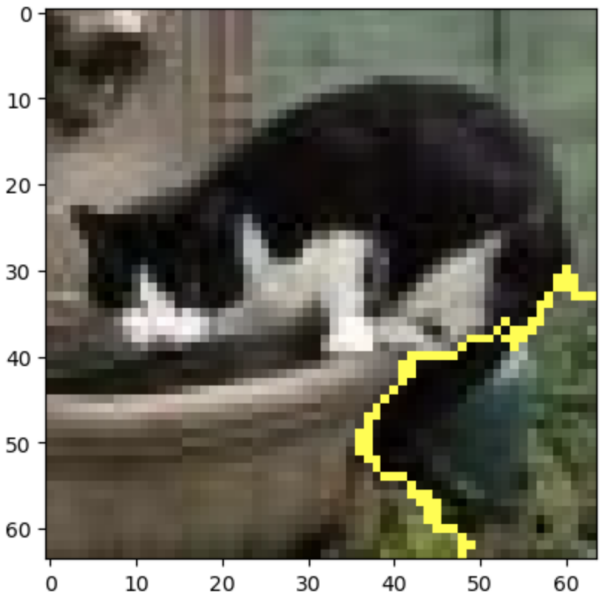
In this section, we will delve into the utilization of two interpretability methods, SmoothGrad and LIME, applied to a convolutional model. Our aim is to gain a deeper understanding of the features that the model considers significant and to identify the common features associated with specific animals. By employing these techniques, we can visually perceive the significance of individual pixels and generate explanations that shed light on how the model makes its decisions. Through this analysis, we will uncover valuable insights into the model's feature selection process and obtain meaningful interpretations of its decision-making.

11



SmoothGrad is an approach that generates a heatmap showcasing the gradients on the image, emphasizing the pixels that have the greatest impact on the model's predictions. Through the visualization of these gradients, we can discern the specific areas of the image that the model prioritizes when making its predictions. This technique aids in comprehending the pixel-level significance for the model's decision-making process, offering valuable insights into the model's behavior and reasoning.

12



Likewise, we will employ LIME to generate explanations for the features chosen by our convolutional network. LIME is a method that involves modifying the input image and observing the resulting changes in the model's output. By analyzing these modifications and their effects on the predictions, LIME identifies the features that the model relies on to make its decisions. This enables us to gain insights into the specific attributes or characteristics of the animals that hold the most significance for the model.

These two methods both marks the tail parts with sharp color changes to mark section differences indicating that the tail is an important feature for distinguishing the pixels in the images. The difference is Smooth-Grad provides more selcted features to distinguish pixels, so SmoothGrad can provide a more detailed and comprehensive understanding of the model's decision-making process here. Thus, SmoothGrad captures a wide range of features and regions, providing insights into the model's reasoning beyond the tail. LIME focuses on localized explanations but may overlook other relevant features. Combining both methods offers a comprehensive understanding: SmoothGrad highlights multiple features, while LIME emphasizes the tail's impact. Considering both outputs reveals the tail's significance and other image features.