Lab 4 - Explainability,
Stat 415, Spring 2023

Wing Yan Yau

26th May 2023

1 Pima Indians Diabetes Database

There are in total 768 data points, with 500 of them are instances without diabetes and 268 of them are diabetics (i.e. around 34.9%). This approximately 2-to-1 ratio suggests the undersampling on diabetic pima indians in this data set. We thus do random-oversampling , i.e. diabetic datapoints from the training data set are selected randomly with replacement, and get a data set now with 1000 data points in total, 50% for each output(diabetic or non-diabetic).
Running a linear model on the balanced dataset with an L1 penalty gives the following results of coefficient.

Pregnancies	Glucose	BloodPressure	SkinThickness
9.14e-3	5.88e-3	-1.68e-3	-1.02e-4
Insulin	BMI	DiabetesPedigreeFunction	Age
-1.36e-4	1.61e-2	6.29e-2	6.24e-3

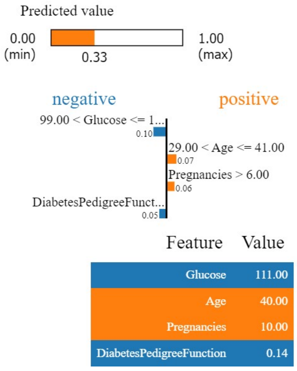
We can see that the 2 major features selected by the model is DiabetesPedigreeFunction and BMI. Pregnancies, Age and Glucose are also some other (though less) important features in predicting if one is diabetic according to their coefficient magnitude in the model.

We then run a random forest on it, the 'feature importance' results are shown in the following table

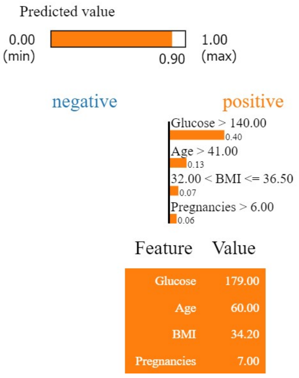
Pregnancies	Glucose	BloodPressure	SkinThickness
5.67e-2	3.67e-1	7.61e-2	5.43e-2
Insulin	BMI	DiabetesPedigreeFunction	Age
4.75e-2	1.54e-1	1.11e-1	1.33e-1

We can see that, for this model, the major features selected are Glucose, BMI, DiabetesPedigreeFunction and Age.

Apart from checking the magnitude of coefficient for different features to get a sense of their relative importance in predicting if one is diabetic or not, we also use LIME on the dataset. 2 of them are presented with details in this report as below. There is a overlapping of important features of Glucose and Age.
Our goal is to find the most significant set of features, so we apply LIME on 200 randomly selected data points, and then observe the distribution of selected top 3 features among them. We



LIME result 1



LIME result 2

found that LIME is quite stable across different datapoints, with Age, Glucose and BMI being dominantly the most important features in general.

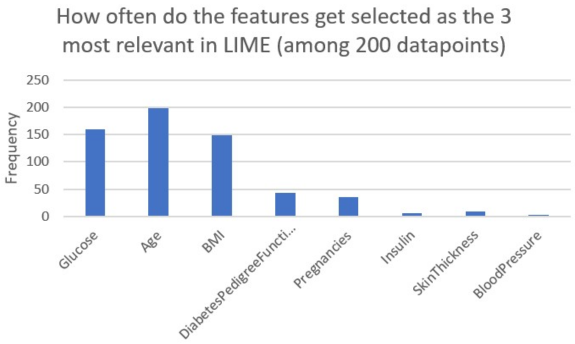


Figure 1: Bar chart on how frequent does each feature get selected as one of the top 3 most important ones among 200 data points

2 Animal Images

2.1 Predictive Modeling

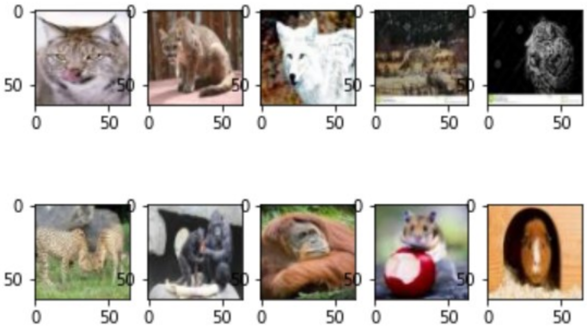


Figure 2: visualization of an example image from each class in the training data set (class 0 to 9, from left to right, then top to bottom)

For this data set, we trained 3 models to try classifying animals from an image of them. Among them there are a linear model, a vanilla convolution neural network(CNN) and a tuned CNN. The vanilla CNN consists of only 1 convolution layer, followed by a flatten layer and linear layer. For the tuned CNN, there are 4 main layers, each consists of a convolution layer, a batchnorm layer and a activation layer (we chose ReLU in this project), with the last one as a maxpool layer added. At the end there is a flatten layer and a linear layer. Below is the learning curves of the 3 models.

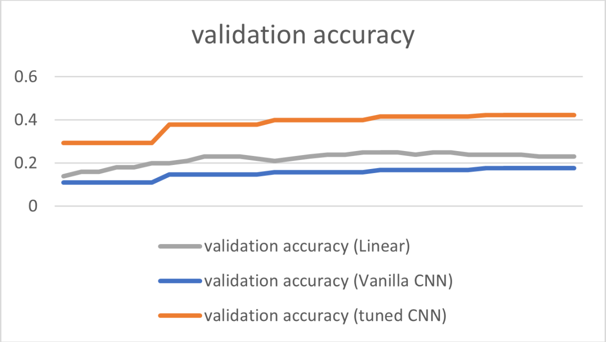


Figure 3: validation accuracy against number of epoches

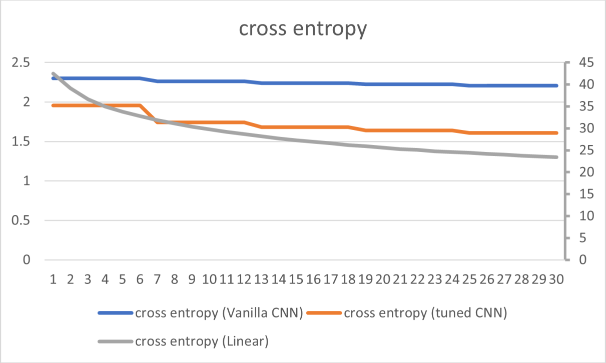


Figure 4: cross entropy against number of epoches (secondary axis on the right for the linear model)

Learning rate is one of the most important hyper-parameters, the model's performance would be severely affected if we did not choose an appropriate one. We first use 1e-9 as the learning rate as it is approximately the largest possible value to set without resulting in gradient explosion (all the outputs become nan). Yet the model barely learn anything, so we increased it to 1e-5 and the performance, in terms of validation accuracy, shows a significant improvement. With learning rate equals to 1e-9, the tuned CNN shows an increase of best validation accuracy less than 0.01 after 5 epoches. With learning rate equals to 1e-7, the tuned CNN shows an increase of best validation accuracy around 0.21 after the same amount of training. When we finally increased it to 1e-4, we get an improvement of around 0.3 in 5 epoches of training (from 0.242 to 0.433).

Adequate amount of epoches are also essential.

2.2 Feature Atribution

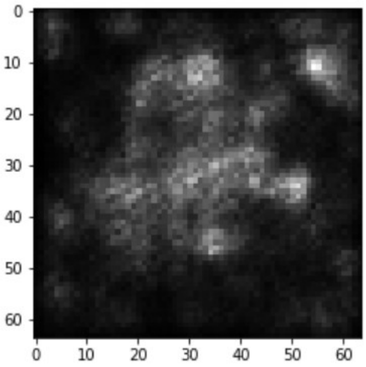


Figure 5: A heat map of the gradient on the image demonstrating what pixels SmoothGrad is selecting on

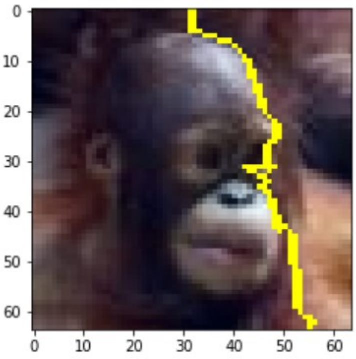


Figure 6: LIME explanation

Comparing the features selected on by LIME and SmoothGrad, the curvy forehead might be a common features for Orangutan.