Pima Indian Diabetes Discussion

Dataset

The dataset consists of only female individuals whom are at least 21 years old and are of Pima India descent. The data originated from the National Institute of Diabetes and Digestive and Kidney Diseases; our target is the 'Outcome' column which shows if the individual has diabetes or not.

Source: https://www.kaggle.com/uciml/pima-indians-diabetes-database

Data Analysis

There was little correlation between the columns, aside from slight correlation in the following areas: Age+Pregnancies, Glucose+Outcome, Insulin+SkinThickness.

The Age and Pregnancies correlation is somewhat clear, you are less likely to have children when you are younger due to simply being young and having to go through education and at least the start of a career.

Blood glucose levels more directly correlate with the diabetes 'Outcome' as it is generally a significant factor in controlling diabetes – hence why insulin is prescribed, to control these levels. Interestingly, the blood glucose levels correlation with diabetes is not higher, suggesting that further underlying factors are at play when determining diabetes. This provides a meaningful area to use a neural network, to find these hidden correlations.

Models

4 deep learning models were created, the initial being a tensorflow base model to compare the others with. Note that in this instance different models perform significantly differently in areas of recall and precision when predicting HAS NO diabetes and HAS diabetes. What would be the most important when it comes to determining diabetes? We do not primarily need to focus on who does not have diabetes – we want to know who *does* have it, so that they may go on to further critical testing. In this case, we are looking for a high recall score for 'Outcome'=1 (HAS diabetes).

In overall score, the logistic regression model with GridSearch performed the best. Although, in the metric we are observing, no model came close to the original base model's impressive recall of 97% when predicting HAS diabetes – but it suffered at 30% in predicting HAS NO diabetes.

This original sequential model should be used for our initial goal of finding individuals whom likely have diabetes before their official diagnosis. Such an effective model can be used very quickly once the patient's information has been recorded – this will be of particular use when the doctor may not diagnose diabetes, but the model does, prompting further testing.

Further Investigation

Key insights can be gained by removing different columns and recreating the model – comparing the efficacy of the models with some factors missing. An overall study of this would shed light on the weighting of each factor – which ones are more important than others when it comes to diabetes.