Capstone Project Proposal
Jason Mark – Aug 2019

Background: Diabetes is one of the most common and most expensive chronic diseases worldwide. In 2004 it was estimated that in the US alone, approximately 5 million people unknowingly had the disease while another 13 million were aware of their diagnosis.

Problem Statement: Early detection of the disease can help reduce the risk of serious life changing complications such as premature heart disease, stroke, blindness, limb amputations, and kidney failure. Models that can help predict an individual with diabetes could be a useful tool to support a physician's decision-making process when working with patients. It could also be leveraged to screen populations of patient data to identify patients most likely to have undiagnosed diabetes and intervene with further testing and monitoring. This can be framed as a binary classification problem to separate those who will vs. those who will not develop diabetes.

Datsets and Inputs: For purposes of this project, I will be using the Pima Indians Diabetes Dataset. This dataset originates from the National Institute of Diabetes and Digestive and Kidney Diseases and is now hosted on Kaggle by the UCI Machine Learning Repository. It contains 8 measurement values for 768 patients of which 268 were determined to have diabetes as indicated in an 'Outcome' variable. The 8 included measurement variables are:

Measurement
PregnanciesNumber
BloodPressureDiastolic
GlucosePlasma
SkinThicknessTriceps
Insulin-2Hour
BMIBody
DiabetesPedigreeFunctionDiabetes
Age

Proposed Solution: It is proposed to build a machine learning model to predict a patient having diabetes based on the measurements present in the dataset. It is further proposed to host this model as a Sagemaker endpoint which could be easily integrated into other applications to enhance them with this functionality.

Benchmark Model: Existing models on Kaggle have achieved accuracy in the range of 70-80%. The original paper published using this dataset obtained 76% accuracy and balanced out sensitivity and specificity. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245318/?page=5) My expectation is to be able to build a model achieving this same level of performance.

Evaluation Metrics: The model will primarily be evaluated based on overall accuracy, although consideration also needs to be given to the percentage of false positives and false negatives. Because the model predicts a disease state, it may be more important depending on the particular case where the model is used to either be more certain of a positive diagnosis or more certain of a negative diagnosis. Incorrectly predicting diabetes for a patient could be upsetting to the patient and lead to unnecessary actions or tests while incorrectly missing a diagnosis could lead to a patient developing the very complications that the model hopes to prevent.

Outline of Project Design:

The basic outline of the proposed project is:

- Preprocess and evaluate the available data.
- Develop an understanding of any gaps in the available data and how they may impact model performance.
- Split the data into appropriate test/train groups
- Feature engineering including normalization of values and determination of features most useful for modeling.
- Build, train, and deploy model as SageMaker endpoint.

    - I intend to build a model using the xgboost algorithm based on the success this algorithm tends to currently have on a variety of tasks.
    - It may also be useful however to consider a Decision Tree algorithm in this case because the logic contained in the decision tree cuts may yield good guidelines for interpreting the data that can easily be understood by physicians. Given that we are dealing with a health problem, interpretability behind a prediction is important.
- Evaluate resulting model using held out test set data.

References:

Diabetes: A National Plan for Action. The Importance of Early Diabetes Detection. https://aspe.hhs.gov/report/diabetes-national-plan-action/importance-early-diabetes-detection

Pima Indians Diabetes Dataset
https://www.kaggle.com/uciml/pima-indians-diabetes-database/activity

Using the ADAP Learning Algorithm for Forecast the Onset of Diabetes Mellitus
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245318/