Capstone Project Report
Jason Mark – Aug 2019

Definition:

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. The expenses involved in treating diabetes once diagnosed are enormous. It is estimated that in 2012 the cost related to treatment of diagnosed diabetes in the U.S. alone was $245 billion and that there were 252,806 deaths related at least in part to diabetes. [1]

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. [2] Each of these brings with it an increase in other types of risks such as secondary infections. It can reduce both loss of life and cost while simultaneously increasing the quality of living for affected patients.

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 proactively 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.[4] 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
DiabetesPedigreeFunction
Age

Analysis:

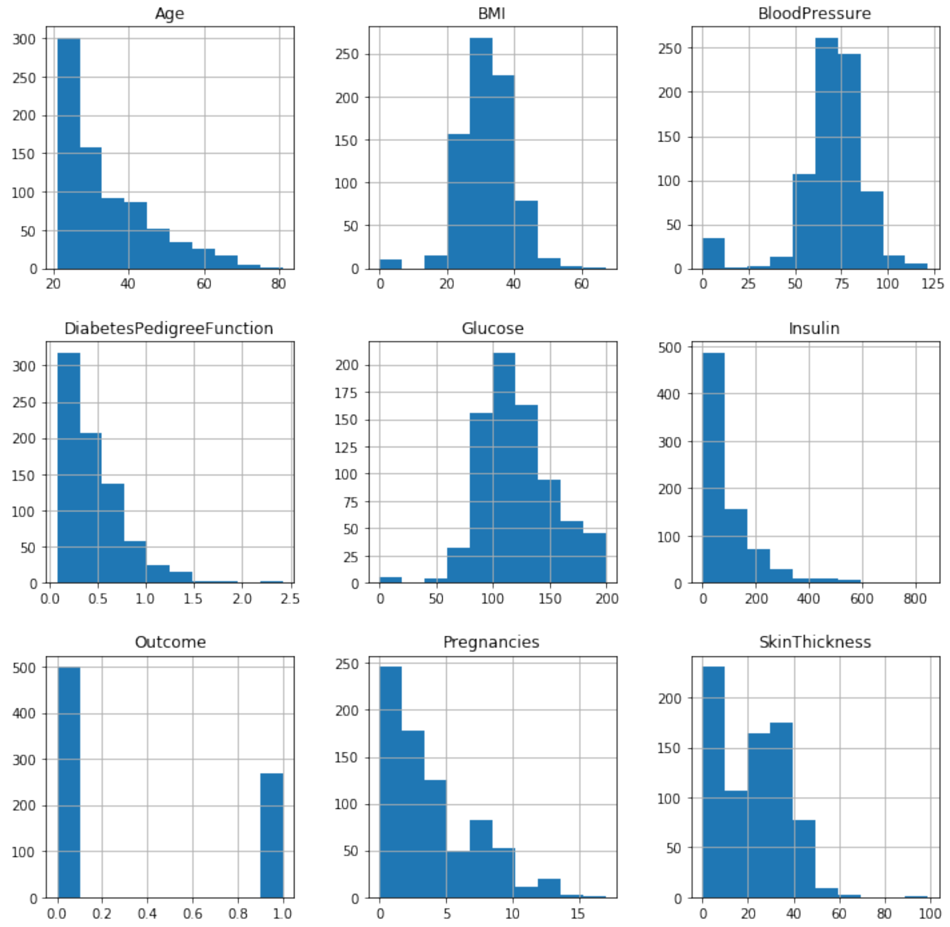
Data Exploration:

Upon further inspection of the Pima Indians Diabetes Dataset, there are 768 patients represented with 268 of these having the diabetes label. Data is not for each feature on each patient observation. Practically this makes sense because unless collected under the same set of controls for each patient, some items may be more difficult to measure than others. An example of a feature that could be particularly problematic for example is the "DiabetesPedigreeFunction". This feature represents the probability of a patient having diabetes given their family history. This could be problematic because it becomes dependent on the accuracy of family history, whether or not diabetes was correctly identified in relatives, and in the case of something like adoption, whether or not family history information is even available on a patient. Because it is likely that at prediction time, all measurements may also not be present, I chose to leave the missing values in the dataset and am using a model which could work around those values. Fortunately, all "missing" values are represented as zeros in the data so there is not a need to replace them explicitly if I am not replacing them with a different value (e.g. median). It is worth pointing out however, that the meaning of a "0" value varies depending on which feature is being considered. In the case of "BloodPressureDiastolic" for instance, a 0 must represent a missing value because an actual blood pressure value of 0 would represent a deceased patient. On the other hand in the case of “PregnanciesNumber”, a 0 represents a valid observational value that a patient has had 0 pregnancies.

One other consideration for this dataset is that it is collected specifically from a study using members of the Pima Indian tribe. It is entirely conceivable that there are genetic or lifestyle differences within this population that make any model built using this data not robust enough or valid for a larger more general population. With this in mind, any resulting model would need to be validated against populations of patients of either their respective population subgroup or against the general population before being widely deployed or used.

Baseline / Benchmark: I am interested in comparing my results to that of the original paper that was written about the Pima Indian Dataset. [3] In this paper, the sensitivity and specificity achieved was 76%. The original paper also shows the ROC curve which is what I'm intending to compare to my results.

Exploratory Visualization:



Visual inspection of a summary of the data shows that there are a number of features with minimum values of 0. These are worth exploring in more detail. It does not look like the data contains any NaN / NA values.

Review of missing feature values:

There are missing feature values in the dataset although they're not all immediately apparent because they are coded as zeros rather than NaN or NA values. Because zero can be a valid measurement for some of the variables, we'll need to consider them one by one:

Feature	Whether or not 0 is valid
Pregnancies	0 can be a valid measurement
Glucose	0 is unlikely to be a valid measurement
Blood Pressure	0 is unlikely to be a valid measurement
Skin Thickness	0 is unlikely to be a valid measurement
Insulin	0 is unlikely to be a valid measurement
BMI	0 is unlikely to be a valid measurement
DiabetesPedigreeFunction	0 can be a valid measurement

Algorithms and Techniques:

Initially I tried using simpler linear models and the DecisionTreeClassifier from the scikit-learn package to build models and make predictions from the data. My rationale for using these approaches was that they could also offer pretty clear guidance on how the cuts were made which could be more straightforward for a human to understand. Given that this problem is within the medical domain and could be used to guide care for patients, I think that an easily explainable model would be favored over a more opaque or “black box” type model. In applying these models however, I didn’t get results much different from the original paper. Given that my goal was to try to improve on the original paper, I set these models aside.

I finally landed on using XGBoost algorithm to predict those cases where diabetes onset will occur. The XGBoost algorithm is currently a widely successful algorithm on a variety of problems. In many cases it can yield performance close to or on par with deep learning models. It is also robust enough to handle the class imbalance in the dataset. Furthermore, AWS SageMaker has a pre-built container for this algorithm which simplifies the development of an initial model without needing to implement a custom container for a deep learning approach like TensorFlow or PyTorch. I'm favoring the most straightforward approach first and can come back and explore those models later if needed.

Methodology:

Data Preprocessing:

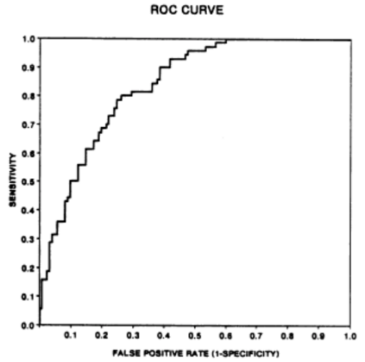
- Using scikit-learn MinMaxScaler class, features were scaled to values between 0 – 1 to help improve model performance.
- Using scikit-learn train_test_split method, data was split into:
    - 80% Training set
    - 10% Validation set
    - 10% Testing set
- Prepared data was output to csv format and loaded to S3 for use with SageMaker

Implementation / Refinement:

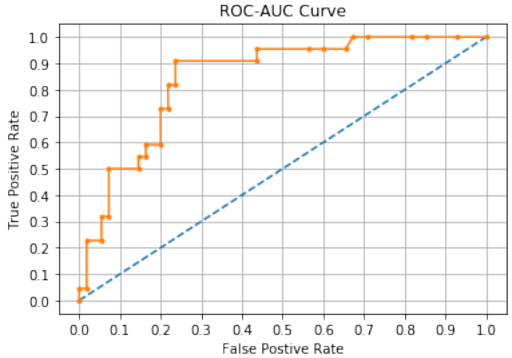
- Standard SageMaker xgboost container was used.
- Created an xgboost Estimator.
- Created a HyperParameter Tuner object.
- Ran HyperParamter Tuner job to search across up to 10 models for best performance based on optimizing for the "validation:auc" metric.
- Output hyperparameters of best performing model to simplify future reproducibility.
- Best performing model was deployed as SageMaker endpoint.
- Holdout test set passed to endpoint to evaluate model performance.

Results:

Model Evaluation and Validation:



ROC from original paper



ROC from project model

Based on the ROC-AUC curve, my model’s performance is similar to very slightly better than the original paper model. In comparing to other results using the same dataset, they have similar results using a variety of different models. I think that the similarity in results from a variety of researchers over a span of many years indicates that we’ve likely reached the best obtainable results on this very limited dataset with current methods. Diabetes and health in general are complicated problems that are influenced by a wide variety of factors. I would say that the ability of the models to have some positive predictive power supports the idea that better models could be produced given a larger dataset or a dataset with more available features per observation.

References:

[1] CDC – Deaths and Cost related to Diabetes.
https://www.cdc.gov/diabetes/data/statistics-report/deaths-cost.html

[2] 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

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

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