Udacity Machine Learning Capstone Project- 2020

Diabetes Prediction with Sagemaker

Amudha Narayanan

Background:

Today world problem is COVID -19, which have 7.3% fatality rate on patients with Diabetes comparing to other patients is 2.3 % [1]. Many complications occur if diabetes remains untreated and unidentified in COVID situation as diabetes is one of the deadliest and chronic diseases which causes an increase in blood sugar end up in multiple organism failure with COVID. With the current situation it is not recommended to visit diagnostic center and hospital multiple times just for diagnosing. Machine learning approaches solves this critical problem at the early stage by analyzing the predicted result with online doctor consultation. As the vaccination for COVID-19 is not still available early detection of diabetes help patients and world health sector to focus on health and infrastructure planning in advance to handle growing number of diabetes patients which is forecasted to be 642 million by 2040.

As my parents have diabetes and as a Machine Learning Engineer working in Research and Technology, I always have intention to build Diabetes predictor, now Udacity ML nanodegree course had now given me confident to explore diabetes dataset and use AWS Sagemaker to come with end to end solution for Diabetes Predictor.

Problem Statement:

The motive of this study is to design a model which can prognosticate the likelihood of diabetes in patients with maximum accuracy just by invoking API with parameters they got from the Glucose Tolerance Test(GTT) result done recently before COVID or from laboratory results done in remote. The model result will be of binary classification The results of this model can then be used to help detect diabetes in an earlier stage to allow individuals to live a healthier lifestyle and be alert in protecting themselves from coronavirus until vaccination is found and safe to inject to all.

Datasets:

The Prima Indian Diabetes Dataset with 768 patient's data has been used in this study, provided by the UCI Machine Learning Repository (https://www.kaggle.com/uciml/pima-indians-diabetes-database). The dataset has been originally collected from the National Institute of Diabetes and Digestive and Kidney Diseases. The number of true cases are 268 (34.90%) and the number of false cases are 500 (65.10%) in this dataset.

If time permits planning to use the dataset of diabetes with 2000 patient's data, taken from the hospital Frankfurt (https://www.kaggle.com/johndasilva/diabetes) also to compare the performance of model with benchmark Prima Indian Diabetes Dataset which also has same proportion of true cases (684 – 34.20% and false cases (1316 – 65.80%) .

Both the dataset consists of following medical distinct measurement variables

Variables	Description	Data type
Pregnancies	Number of times pregnant	int64
Glucose	Plasma glucose concentration a 2 hour in GTT	int64
BloodPressure	Diastolic blood pressure (mm Hg)	int64
SkinThickness	Triceps skin fold thickness (mm)	int64
Insulin	-Hour serum insulin (mu U/ml)	int64
BMI	Body mass index (weight in kg/(height in m)^2)	Float64
DiabetesPedigreeFunction	Likelihood score of diabetes based on family history	Float64
Age	Age of Patient (years)	int64

Solution statement

It is proposed to build a machine learning model with more than one Machine Learning algorithms such as Logistic Regression, XGBoost classifier to predict whether a patient have diabetes or no diabetes based on the measurements captured with Glucose Tolerance Test(GTT). If time permits proposed to host XGBoost classifier model as a Sagemaker endpoint API so that it could be easily integrated into server client web applications or in mobile application developed externally of this Diabetes predictor.

Benchmark model

Many Kaggle projects with Prima Indian Diabetes dataset with 8 attributes able to get the accuracy for Logistic regression in average around 75%, In this SageMaker Diabetes Predictor project targeting to achieve accuracy around 73% to 76% with XGBoost classifier pre-built container of SageMaker by processing the Prima Indian Diabetes dataset and by hyper parameter tuning.

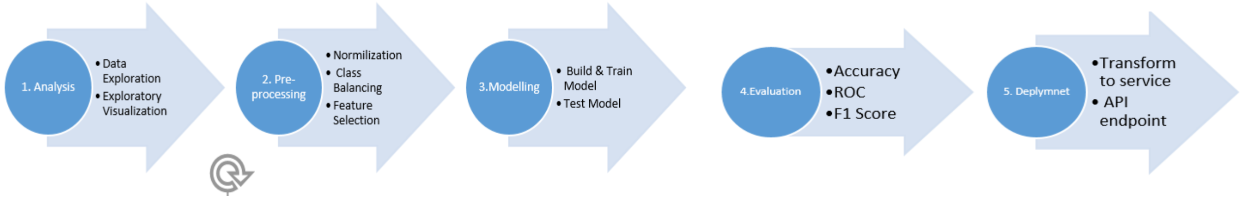
Evaluation metrics

The model will primarily be evaluated based on overall accuracy, since the real dataset doesn’t have equal samples belong to each class, it is always good to evaluate with more than one classification metrics as it important to consider the percentage of false positives and false negatives. Since our model predicts a disease state, 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 which defeat goal of model. So planning to evaluate model by

1. Calculating classification **Accuracy** where Accuracy is the ratio of number of correct predictions to the total number of input samples.
2. Plotting Receiver Operating Characteristic(**ROC**) between True Positive Rate and False Positive Rate, the larger the Area Under Curve(**AUC**), the better is the model
3. Calculating **F1 score** which is the Harmonic Mean between precision and recall, greater the F1 Score, the better is the performance of the model

Project Design

The basic workflow of the proposed project is given below



1. **Analyze** the problem through visualizations and data exploration to have a better understanding of data and features that are appropriate for solving it by plotting correlation matrix.
2. **Pre-Processing** of data is important as it play an important role in improving performance of the model for small dataset such as by replacing null values with mean, balance imbalance class and select important features through feature selection
   Split the processed dataset into Train and Test dataset (80: 20 ratio), load Training, Test and Validation data in S3 Bucket
3. Implement **Modelling** in SageMaker by selecting Algorithm Container Registry Path, Configure Estimator for training - Specify Algorithm container, instance count, instance type, model output location to train model
4. Perform **Evaluation** of model as explained in Evaluation Metric section of this proposal by computing Accuracy Score, F1 score and ROC-AUC score.
5. Finally **Deploy** model to run prediction through API by specifying instance count, instance type and endpoint name in Sagemaker

References

1. Wu Z, McGoogan JM. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. JAMA. 2020. https://doi.org/10.1001/jama.2020.2648.
2. The Pima Indian Diabetes Dataset used originally came from this paper: ** Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press. Now available for download via Kaggle here.
3. N.Sneha ,Tarun Gangil Analysis of diabetes mellitus for early prediction using optimal features selection
4. Deepti Sisodia , Dilip Singh Sisodia. Prediction of Diabetes using Classification Algorithms International Conference on Computational Intelligence and Data Science (ICCIDS 2018)
5. https://www.kaggle.com/vipulgandhi/how-to-choose-right-metric-for-evaluating-ml-model