Machine Learning Capstone Project

‘Pima Female Indians Diabetes Prediction Algorithm’

Siddhant Tandon

August 11th, 2019

Domain Background:

Today we have the privilege of enjoying many benefits from the comfort of our homes, which has been changing our lifestyle preferences as we grow old. Although these lifestyle changes have improved quality and experience yet has also given birth to diseases which are becoming widespread. Diabetes mellitus is a metabloic diseases that is affecting people across the world due to their lifestyle. USA being one of the developed economies has also started seeing a drastic rise in diabetic population over the last decade. This is quite alarming and thus detection of diabetes in its early stage has become a necessity to avoid severe health problems in both men and women.

The main cause for diabetes as discussed by NIH is when blood glucose level is too high and insulin, a hormone made by the pancreas isn’t enough to turn sugar into energy for our cells. Glucose then stays in blood overtime and doesn’t reach your cells. This is the reason why many organs get affected and people with diabetes start having health problems.

My family has had diabetic history in recent generations from both my parents’ sides. This makes me more susceptible to have diabetes later in life. Thus, my motivation to do this study is to help advance the detection of diabetes in humans to help treat it in its early stages.

This report will cover the data on the Pima Indian population near Phoenix, Arizona. Related research work paper is:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245318/pdf/procascamc00018-0276.pdf

Thus, it has become important to use the help of machine learning algorithms to help detect the possibility of diabetes (in females for this project) so that proper steps can be taken in order to avoid the onset of health problems to those affected by this disease.

Problem Statement:

Build a ML model to accurately predict whether the patients (females of age 21 and older of Pima Indian heritage) in the dataset have diabetes or not.

Datasets and Inputs:

The dataset is obtained from Kaggle.com under the tab ‘datasets’ for public use. It was posted from the UCI Machine Learning repository. The dataset link is below.

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

The dataset has 768 instances and 9 columns. There are 8 medical predictor variables (independent) and one target (dependable) variable labelled ‘outcome’. The independent variables are listed below:

- Pregnancies (Number of times pregnant)
- Glucose (Plasma glucose concentration: a 2 hours in an oral glucose tolerance test)
- Blood Pressure (Diastolic blood pressure (mm Hg))
- Skin Thickness (Triceps skin fold thickness (mm))
- Insulin (2-hour serum insulin (mm U/ml))
- BMI (Body mass index (weight in kg/ height in m^2))
- Diabetes Pedigree function
- Age (years)

The dataset was recorded by National Institute of Diabetes and Digestive and Kidney Diseases.

Solution Statement:

This is a binary classification problem and its solution can be from our lessons under Supervised learning. The most common solution to approach such prediction problems, which are multivariate time series, is to test and compare the following options:

- Logistic regression
- Decision Trees
- Naive Bayes
- Random Forest vs XGBoost vs LightGBM
- SVC
- K-Nearest Neighbors

The idea is to implement ensemble methods in the end to optimize the output from above to produce the best prediction algorithm.

Benchmark Model:

According to the dataset given, there are 268 females with diabetes and 500 without diabetes. Thus, the rate of diabetes in this sample set is close to 34.9%.

The benchmark model chosen is from the larger class in the dataset, which is the negative result (or samples with no diabetes). That gives us the model benchmark of 65.1%.

Aim is to create a model which can detect whether a person has diabetes or not with greater accuracy than 65.1%.

Evaluation Metrics:

A good metric based on the above discussion of binary problems is the F1 score. It measures the test's accuracy. The formula for F1 score is:

F1 = 2* (precision*recall)/(precision + recall)

It can be noted that the F1 score is a weighted average of precision and recall, where a F1 score reaches its best value at 1 and worst at 0.

Project Design:

- Data Visualization – Visual representation of data to find the degree of correlations between predictors and target variable.
- Data Preprocessing – Finding missing numbers, scaling and normalizing operations on data
- Feature Engineering – Finding relevant features, engineer new features
- Model Selection – Fit the data to most appropriate learning algorithms and ranking them according to scores
- Model Tuning – Fine tune the selected algorithm to increase performance
- Final Model – Data visualization and results using the optimized learning algorithm