Machine Learning Engineer Nanodegree

Capstone Proposal

Name:Manas Mukherjee

Date:May 29th, 2018

Domain Background

Diabetes is a leading chronic disease that affects more than 30 million people in the United States. The economic impact of diabetes is estimated at $105 billion, according to a study published in 2016 and led by researchers from Imperial College London, Harvard T.H. Chan School of Public Health and the World Health Organization.

It is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood sugar. Hyperglycaemia, or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels.

Leading companies, as well as startups, are actively working on this specific healthcare domain and coming up with AI-powered applications on use-cases like Glucose Monitoring Systems, Nutrition Coaching, Early Diagnosis Tools etc.

Motivation

In the last two yearly health checkups, my blood-sugar(glucose) level and BMI were above the normal range. Note -

- Glucose : Reference Range: 65-99 mg/dL
- BMI : Reference Range: 18.5-24.9 (calc)

Since these two factors indicate a pre-diabetes tendency, I would like to take some proactive measures to prevent any future suffering. The motivation behind working on this important domains is to increase self-awareness and apply my newly learned ML skills in a real-world problem.

References - 1. Worldwide trends in diabetes since 1980: https://www.thelancet.com/pdfs/journals/lancet/PIIS0140-6736(16)00618-8.pdf 2. Machine Learning for Managing Diabetes: 5 Current Use Cases: https://www.techemergence.com/machine-learning-managing-diabetes-5-current-use-cases/ 3. How Machine Learning Is Helping Us Predict Heart Disease and Diabetes: https://hbr.org/2017/05/how-machine-learning-is-helping-us-predict-heart-disease-and-diabetes 4. ML and Data Mining Methods in Diabetes Research: https://www.sciencedirect.com/science/article/pii/S2001037016300733

Problem Statement

The goal of this project is to build a machine learning model to predict the onset of diabetes based on some diagnostic measures. The expected outcome of this project is to contribute to the clinical research study related to diabetes.

- Main Objective: Build a mathematical model to predict whether or not a patient has diabetes, based on certain diagnostic measurements.
- Secondary Objective: Identify important features or a set of clusters of features that are highly related to diabetes.

Datasets and Inputs

- Source https://www.kaggle.com/uciml/pima-indians-diabetes-database This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases.
- Content The dataset consists of **8 medical predictors** variables and **one target variable** - 'Outcome' corresponding to the **768 persons**. Here is the list of predictor variables with their brief description.

Feature Name	Description	Datetype
Pregnancies	Number of times pregnant	Numeric
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Numeric
BloodPressure	Diastolic blood pressure (mm Hg)	Numeric
SkinThickness	Triceps skin fold thickness (mm)	Numeric
Insulin	2-Hour serum insulin (mu U/ml)	Numeric
BMI	Body mass index (weight in kg/(height in m)^2)	Numeric
DiabetesPedigreeFunction	Diabetes pedigree function	Numeric
Age	Age (years)	Numeric
Outcome	Class variable (0 or 1)	Numeric

Constraint

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are **females** at least **21 years old** of **Pima Indian** heritage.

Background details and important characteristics of Pima people

The Pima are a group of Native Americans living in an area consisting of what is now central and southern Arizona. The majority population of the surviving two bands of the Akimel O'odham is based on two reservations: the Keli Akimel O'otham on the Gila River Indian Community (GRIC) and the On'k Akimel O'odham on the Salt River Pima-Maricopa Indian Community (SRPMIC).

The Keli Akimel O'odham and the Onk Akimel O'odham have various environmentally based health issues related to the decline of their traditional economy and farming. **They have the highest prevalence of type 2 diabetes in the world, much more than is observed in other U.S. populations.** While they do not have a greater risk than other tribes, **the Pima people have been**

the subject of intensive study of diabetes, in part because they form a homogeneous group.

Reference: Wikipedia - https://en.wikipedia.org/wiki/Pima_people

Solution Statement

The data includes a binary prediction variable **outcome**. Since the model will predict this outcome(true/false), it is considered as a **supervised 'binary classification'** task.

Supervised: The given dataset has diagnostic features of several individuals and a corresponding label that indicates presence or absence of the diabetes disease. In this task, we will train the model with a subset of the given data and use the other part of the data for testing.

Classification: The goal of this classification function is to predict the outcome given the diagnostic values. As part of the process, we will try different binary classification models using standard algorithms like SVM, Decision Tree, Logistic Regression etc and will find out the best model after doing a comparative analysis. We will also experiment with different model specific hyper-parameters to find the optimized model.

Furthermore, we will do a thorough 'feature importance analysis', clustering using PCA to understand the key feature or set of features that drive the outcome.

Benchmark Model

- **Benchmark-1 : Naive Model**

Since 500 out of 768 individuals don't have diabetes, so we would consider this an unbalanced dataset. A naive benchmark model for this problem would be to predict that none of the patients have diabetes.

- **Benchmark-2 : Domain Knowledge-based Model**According to the Wikipedia "They have the highest prevalence of type 2 diabetes in the world, much more than is observed in other U.S. populations".Type 2 diabetes (formerly called non-insulin-dependent, or adult-onset) results from the body's ineffective use of insulin. **Type 2 diabetes comprises the majority of people with diabetes around the world, and is largely the result of excess body weight and physical inactivity.**According to several legitimate sources, Type-2 diabetes is correlated with obesity and/or BMI. Following article suggests that **Overweight was defined as having a body mass index between 25 and 29.9**.
  http://www.diabetesincontrol.com/body-mass-index-and-type-2-diabetes-risk/We'll consider the **average(27.5)** of the above BMI range(25 to 29.9) as a threshold of this domain knowledge-based model. If a patient's BMI is above 27.5 then this model will diagnosis 'diabetes'.

Evaluation Metrics

While evaluating the performance of the benchmark models and best solution model, I will consider the 'model accuracy' using train/test split or K-fold cross-validation. The input dataset is bit unbalanced(~35%:65%) w.r.t the prediction variable 'outcome', so I will leverage the 'recall' factor

and use 'confusion matrix' to measure different types of prediction. Since this model will be used to disease diagnosis, we will try to minimize the 'false negative' type error (i.e A result that appears negative when it should not).

In this case, we would use F2 Score since it favors 'recall' over 'precision'.

- Related formula**Fβ score = (1+β2) * (precision□recall / (β2□precision)+recall)**
  - In this cases, β = 2.
- **Precision = [True Positives/(True Positives + False Positives)]Recall = [True Positives/(True Positives + False Negatives)]**

Project Design

I'll follow the general machine learning workflow while working on the solution.

1. Identify Environment and Libraries
   - Language and Version: Python(Conda distribution) - 2.7
   - Libraries: Pandas, Scikit-learn, Matplotlib, Seaborn etc. Others libs will be added if required.
2. Exploratory Analysis and Data Preprocessing
   - Clean missing values(if any)
   - Visualize individual features to understand the distribution and skewness
   - Visualize feature correlations
   - Apply log transformation to reduce the effect of outlier
   - Normalize numerical features to ensure each feature is treated equally when applying supervised learners
3. Experiment with different ML algorithm and find the best one using the above performance metric.
   - Evaluate FB scores using the following supervised algorithms
     - Logistic Regression
     - K-Neighbors Classifier
     - Support Vector Classifier
     - Decision Tree
     - Gaussian Naive Bayes
     - Ensemble methods
     - We might use the MultiLayer Perceptron(DL) etc
   - Select top 3 models and perform hyper-parameter tuning
   - Evaluate the best model on the testing set
4. Feature Importance and Insights Find top feature or set of features that are essential for this model.
   - Evaluate the feature importance
   - Identify feature cluster(if any) using PCA

References :

1. https://pandas.pydata.org/pandas-docs/stable/visualization.html
2. https://www.analyticsvidhya.com/blog/2016/02/7-important-model-evaluation-error-metrics/
3. https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6

4. Smith-Morris, Carolyn. "Diabetes Among the Pima: Stories of Survival". Summary of the above book - https://muse.jhu.edu/article/450458/summary