Fighting Diabetes:

Low-cost Solutions to Health Care Inequity in Indigenous Communities

---

Abstract

The goal of this project is to build a machine learning classification algorithm that uses baseline health measures to predict if someone has diabetes. This model is built to detect diabetes for females of Native American descent, as Indigenous (Native American/Alaskan) communities in the U.S. not only face significantly higher instance of diabetes, but also lower access to health insurance and preventative health care. Using a 1988 dataset from The National Institute of Diabetes and Digestive and Kidney Diseases containing health profiles of females of Pima Indian descent, I built, cross-validated and tested five classification models, ultimately selecting a tuned Random Forest classifier with a F-2 of 90.2% and recall score of 89.2%.

Design

This binary classification model predicts instance of diabetes based on a person's baseline health characteristics, including Blood Pressure, Glucose Levels (two hours after consuming a sugary drink), Age, and BMI. The data were trained on multiple models, the two top-performing models were tuned using Random Search CV, and the final model was selected via K-fold cross-validation, and the scoring metric used was F2-Score and Recall. These metrics were chosen to ensure a minimal number of false negatives, as false negative diagnosis of diabetes could prove extremely dangerous to the health of patients.

Data & Algorithms

The dataset contains 768 observations. All observations are health profiles taken from Pima Indian people who are biologically female, and over the age of 21. The target is a binary variable that indicated a negative (0), or positive (1) status. The features included are: Age, Diastolic blood pressure, BMI, Number of Pregnancies, Insulin, Glucose, Skin Thickness, and a diabetes predisposition function. This function factors in genetic disposition, i.e. family history of diabetes to calculate a value from 0-2.

- Top line EDA revealed near-normal distributions for most features, with the exception of age, and high separability between classes for the glucose and insulin distributions. That said, nearly half of the observations had missing values for insulin.
    - ✳ Left-skewed Age distribution, mostly patients aged 20-30.
    - ✳ Strong correlations between BMI & Skin Thickness, Age and Pregnancy, Glucose and Insulin.
- I fit a host of baseline models to determine which models to tune
    - ✳ Random Forest and Logistic Regression had the strongest baseline Recall scores,
    - ✳ Strong correlations between BMI & Skin Thickness, Age and Pregnancy, Glucose and Insulin.
- I add engineered features to training sets and cross validate scores for the two highest-performing models
    - ✳ Adding a >7.5 pregnancies variable increases Logit performance.
    - ✳ Dropping Insulin and Skin Thickness increase RF performance.
- Tune Hyperparameters of Logit and RF models using Random Search CV
    - ✳ Adding a >7.5 pregnancies variable increases Logit performance.
    - ✳ Dropping Insulin and Skin Thickness increase RF performance.
- Final Model Selection:
    - ✳ RF with >7.5 pregnancies, and drop insulin and skin thickness returned a F2 score 95.8% and a recall score of 95.1%.
    - ✳ This model was incorporated into a simple app interface using Flask.

Tools

- Numpy & Pandas for Data Cleaning & EDA
- Scikit Learn library for building, fitting, and cross-validating Classification Models
- Seaborne and MatPlotLib for visualizations
- Flask for App Development

Communication

- The algorithm has been incorporated into a simple app, which I created using Flask.