Predicting Diabetes

Elisabeth Johnson

Abstract

The purpose of this analysis is to build a model that most accurately predicts whether or not someone may have diabetes (or be at risk of diabetes) given several predictors (20 total). Our hope is to help the public gain an understanding of those variables that might be most indicative of diabetes and to also help those at risk of diabetes learn when to seek medical treatment.

The dataset used has over 200K rows.

Design

For this project I used several python libraries (listed below) along with PowerPoint in order to visualize my data and construct various machine learning models. The models I constructed are as follows: KNN, Random Forest, Gradient Boosting, and Logistic Regression.

In order to account for imbalanced data I used over sampling, and in order to account for noise and feature similarity I leveraged Singular Value Decomposition using PCA decomposition in SkLearn.

Libraries Used

- Pandas
- Numpy
- SkLearn
- Seaborn
- MatPlotLib
- ImbLearn

Data

The data I used for this project was obtained from Kaggle and can be found here.

Algorithms

For this project I used Exploratory Data analysis and classical data cleaning techniques.

Communication

For this project I chose to communicate my findings via python visualizations and PowerPoint.