

Diabetes Prediction

Alexander Wen, Raymond Wong, Michael Eirikson

Table of contents

Summary	1
Introduction	2
Methods	2
EDA	2
Modeling	2
Discussion	2
References	2

Summary

In this project we attempt to build a model to predict diabetes disease. We compared a decision tree model and naive bayes model and found the decision tree is stronger in this context. We used f2-score as our scoring function because detecting diabetes is the priority: a false negative could be much worse than a false positive.

In the test dataset: the decision tree model correctly detected 8283 of 10604 positive cases (recall rate is about 78%). This result does come at a fairly significant cost in terms of false positives (precision rate is about 30%) with 19650 false positives. Depending on the actual cost of false positive this may need significant improvement to be a viable screening model.

Introduction

Methods

Modeling Approach

Results

EDA

Data Summary

Visualizations

Modeling

Classification Analysis

Result Visualizations

Discussion

References