PROJECT MODULE 5

Predicting Diabetic Signals in a Select Pima Indian Population

Data Scientist In Training to Data Scientists In Training

MaryJo Zaborowski

DS 04222019
28 June 2019

THE STAGGERING COSTS OF DIABETES

More than
30 MILLION
Americans
have diabetes

Health care costs for Americans with diabetes are **2.3X** greater than those without diabetes

Diagnosed diabetes costs America
$327 BILLION
per year

What if we could predict
– then prevent –
the onset of Diabetes?

THE STUDY

The Pima Indian population of Arizona

Using a data set from the National Institute of Diabetes and Digestive and Kidney Diseases, predict whether or not a subject will get a Diabetes diagnosis, based on measurements included in the data resource:

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

The data set contains up to eight clinical observations for 768 women, ages 21-81

Steps Performed:

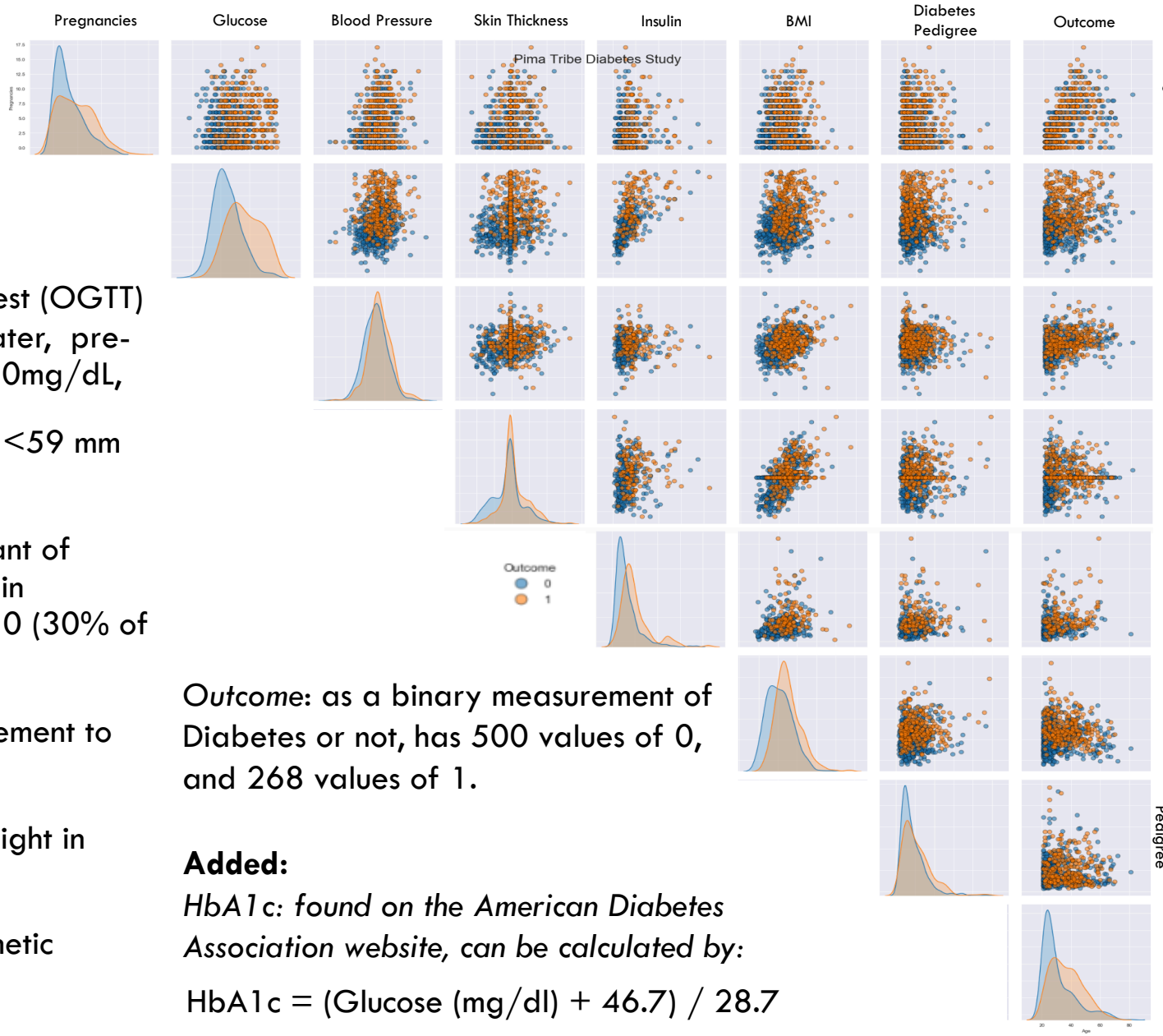
- Data Pipeline
- Data Cleaning
- Feature Engineering
- Feature Selection
- Model Selection

Results Achieved:

- Prediction 59-87%*
  (F1-score) (validation-test)

*excluding all Insulin features

THE DATA



Pregnancies: 0 -17

Glucose: A two-hour, 75-gram oral glucose tolerance test (OGTT) The average values for diabetics is 200mg/dL or greater, pre-diabetics is 140–199 mg/dL), and non-diabetics (<140mg/dL,

Blood Pressure: This is diastolic blood pressure. Low BP <59 mm Hg. Normal is 60-80. High blood pressure is >81.

Skin Thickness: Subcutaneous fat is the major determinant of insulin sensitivity and has a strong association with insulin resistance. Normal is ~23mm. We have 227 values of 0 (30% of the data set.)

Insulin: This measurement is likely a companion measurement to the Glucose test.

BMI: The formula for BMI is 703 times weight in lbs/height in inches squared. The data set has 11 values of 0.

DiabetesPedigreeFunction: This is a measurement of genetic frequency, the calculation of which is unclear.

Outcome: as a binary measurement of Diabetes or not, has 500 values of 0, and 268 values of 1.

Added:

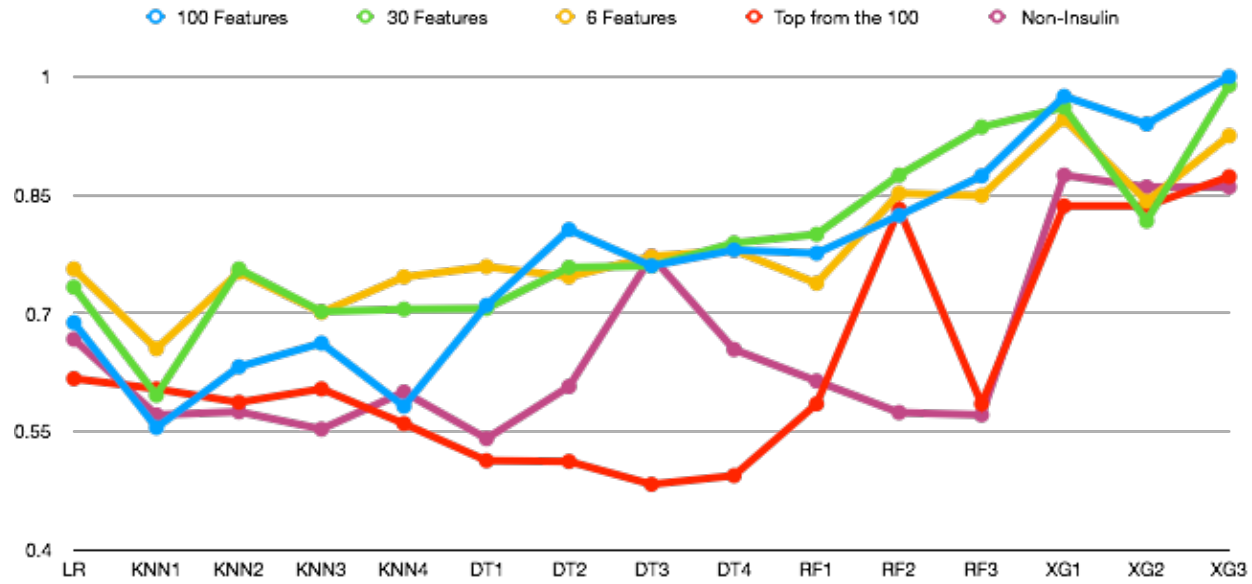
HbA1c: found on the American Diabetes Association website, can be calculated by:

HbA1c = (Glucose (mg/dl) + 46.7) / 28.7

THE MODELS

Can a diagnosis of diabetes be predicted from the given data set?

The analysis was performed with five different feature sets.



Initial project focused on 100 features, including calculated value for HbA1c and imputed values for NaN Insulin.

Model performed very well, raising two questions:

1) is 100 features appropriate (as too many)?

2) is XG Boost over-fitting?

I decided to run the analysis again selecting fewer features –targeting less than 10 features. I measured and graphed the classification error and log loss in my XG Boost models, to look for over-fitting.

The XG Boost exercise proved over-fitting, which I partially-remediated with the early-stopping parameter.

The run with <10 features was successful, delivering results consistent with the 100 feature set. But I noticed the large weight of the imputed Insulin features, rasing another question: Did the imputing method create bias? I decided to re-run the analysis:

1) Use the top-performing features from the 100 feature set – which included no Insulin features

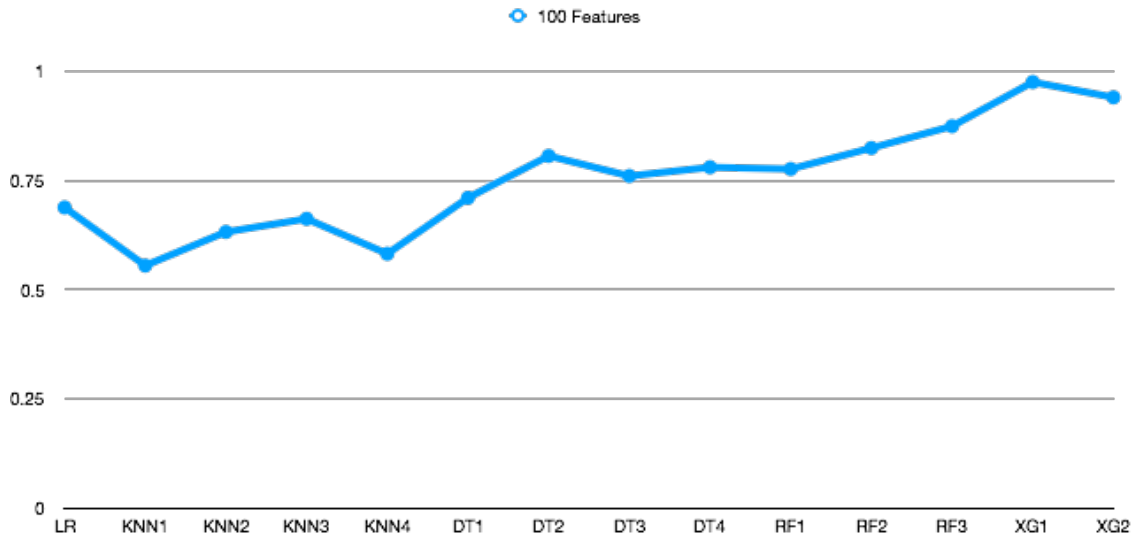
2) Eliminate the use of the Insulin feature and re-select <10 features

THE MODELS

First Approach – 100 Features

Best Performing Features

- A1cNorm AgeLow
- A1cNorm BMINorm
- A1cNorm BPLow
- A1cNorm PregNorm
- A1cPre BMINorm
- BMI DiabetesPedigreeFunction log
- BPHigh AgeHigh
- Glucose BMIHigh
- SkinHigh



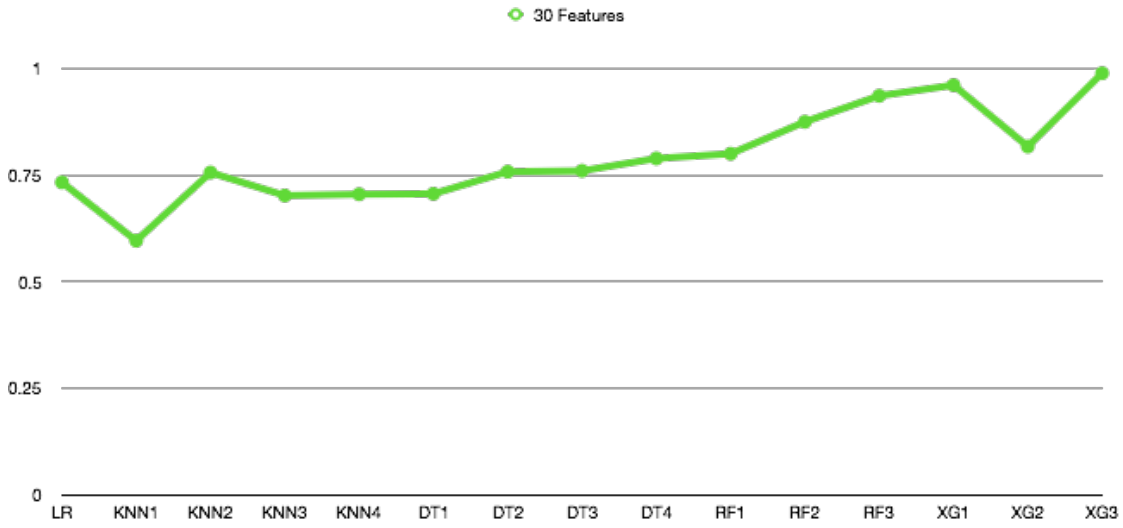
	Best F1
Logistic Regression default parameters	68.8%
KNN K: 14, p: 2 (Manhattan Distance)	66.2%
Decision Tree min_samples_leaf=20	80.6%
Random Forest criterion: entropy, max_depth: 6, min_samples_leaf: 9	87.4%
XG Boost default parameters	97.5%

THE MODELS

Second Approach – 30 Features

Best Performing Features

- Glucose
- Insulin_log
- Glucose BloodPressure
- Glucose BMI
- Glucose BMIHigh
- BloodPressure Insulin_log
- BMI Insulin_log
- GluNorm AgeLow



Best F1

- Logistic Regression 73.3%
  default parameters
- KNN 66.1%
  K: 4, p: 1 (Euclidian Distance)
- Decision Tree 78.0%
  min_using gini, max_depth=2, and min_samples_leaf = 50
- Random Forest 93.6%
  criterion: entropy, max_depth: 6, min_samples_leaf=9
- XG Boost 97.5%
  learning_rate=0.1, max_depth=9, min_child_weight=1, n_estimators=100, and subsample=0.5

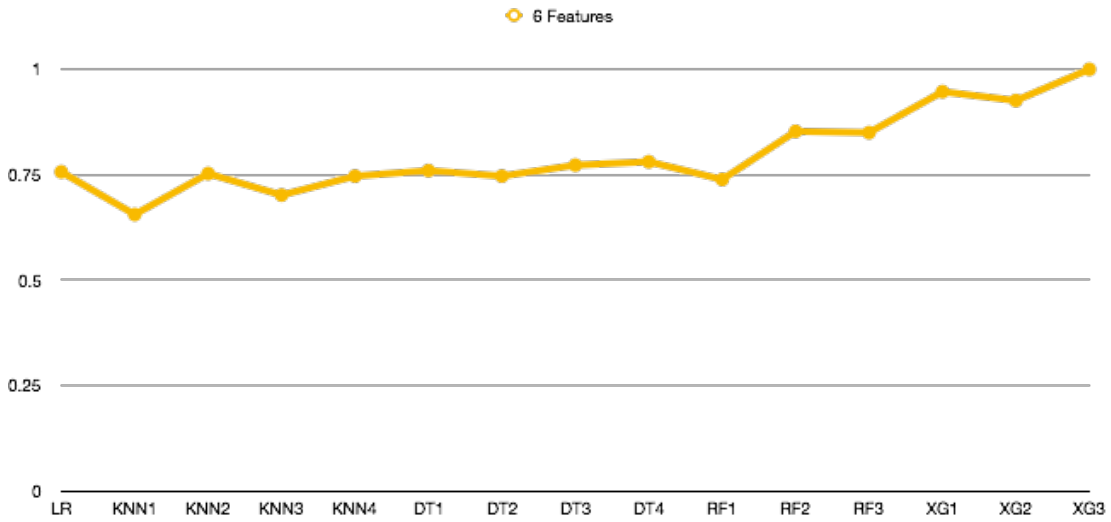
THE MODELS

Third Approach – 6 Features

Features Selected

- A1cDia
- Insulin log
- Glucose BMI
- Glucose BMI_High
- BMI Insulin log
- Glu_Norm Preg_Norm

Noting the significant weight of the Insulin features, and knowing that 336 values were imputed, I suspect inappropriate bias was introduced into the data set, thus rendering this outcome invalid, as well as prior ones.



Best F1

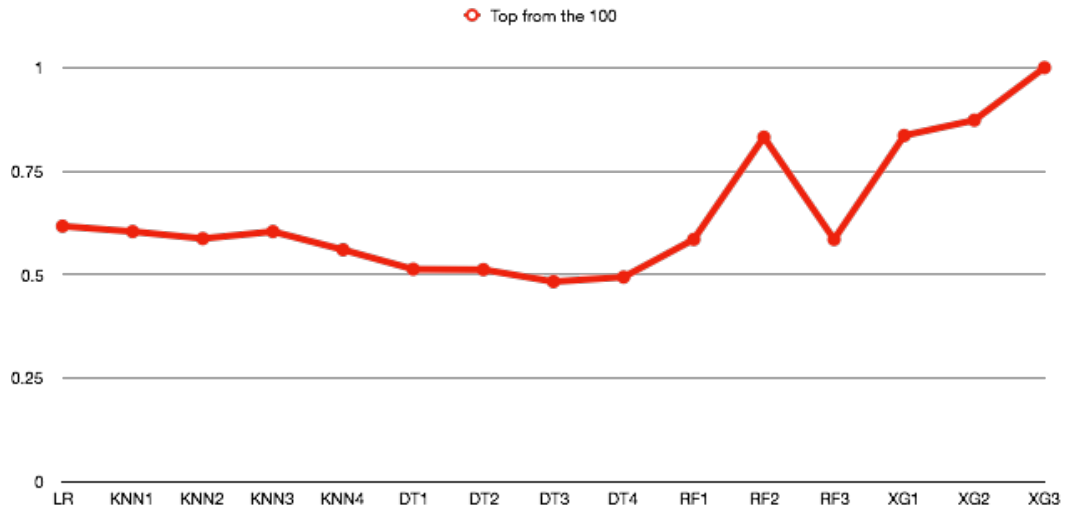
- Logistic Regression 75.6%
  default parameters
- KNN 75.2%
  K: 14, p: 1 (Euclidian Distance)
- Decision Tree 78.3%
  min_using gini, max_depth=2, and min_samples_leaf = 50
- Random Forest 85.2%
  criterion: entropy, max_depth: 6, min_samples_leaf=5
- XG Boost 99.8%
  learning_rate=0.1, max_depth=6, min_child_weight=1, n_estimators=250, and subsample=0.9

THE MODELS

Fourth Approach – Features from 1st Analysis

Features Used

- A1cNorm AgeLow
- A1cNorm BMINorm
- A1cNorm BPLow
- A1cNorm PregNorm
- BMI DiabetesPedigreeFunction_log
- BPHigh AgeHigh
- Glucose BMIHigh
- SkinHigh



Best F1

- Logistic Regression 61.7%
  default parameters
- KNN 60.4%
  K: 12, p: 2 (Manhattan Distance)
- Decision Tree 51.3%
  default parameters
- Random Forest 83.1%
  gini, max_depth=8, min_samples_leaf=2
- XG Boost 87.3%
  learning_rate=0.1, max_depth=6, min_child_weight=10, n_estimators=250, and subsample=0.7

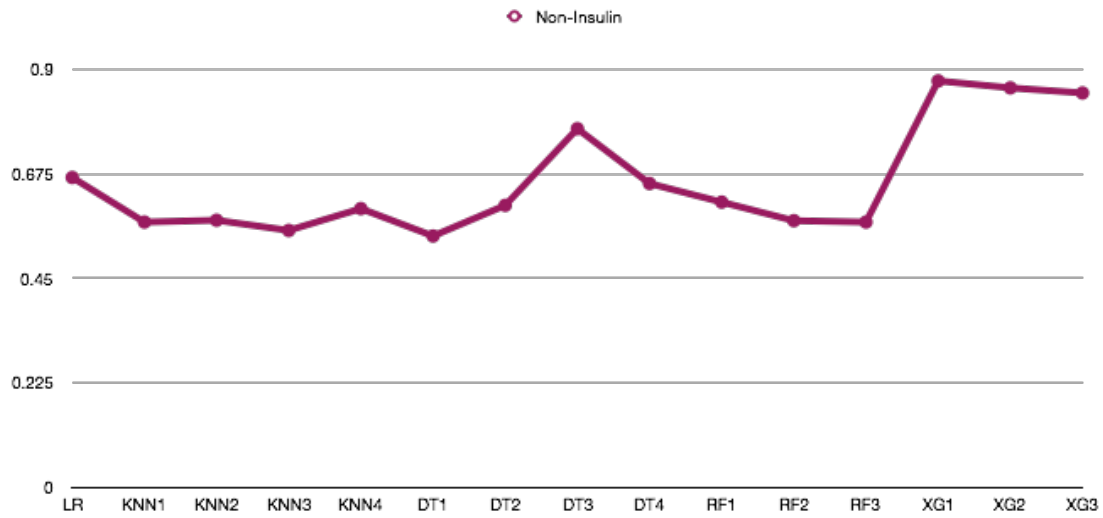
Selected Feature Set

THE MODELS

Fifth Approach – From Scratch With
All Insulin Features Excluded

Features Selected

- Glucose
- A1cNorm
- A1cDia
- Glucose BMI
- Glucose BMIHigh
- A1cNorm DiabetesPedigreeFunction_log
- GluNorm PregNorm



Best F1

- Logistic Regression 66.7%
  default parameters
- KNN 60.0%
  leaf_size=15, n_neighbors=19, p=1, weights=distance
- Decision Tree 77.2%
  max_depth=5
- Random Forest 61.4%
  default parameters
- XG Boost 87.5%
  default parameters

THE ALGORITHMS

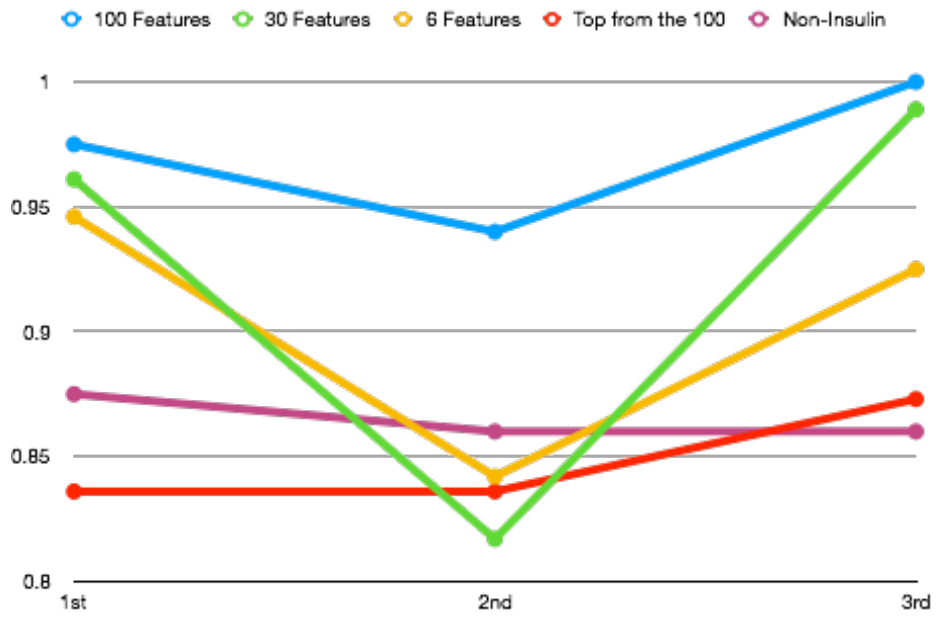
XG Boost – A library implementing the gradient boosting decision tree algorithm

Each analysis cycled through five model approaches, including XG Boost.

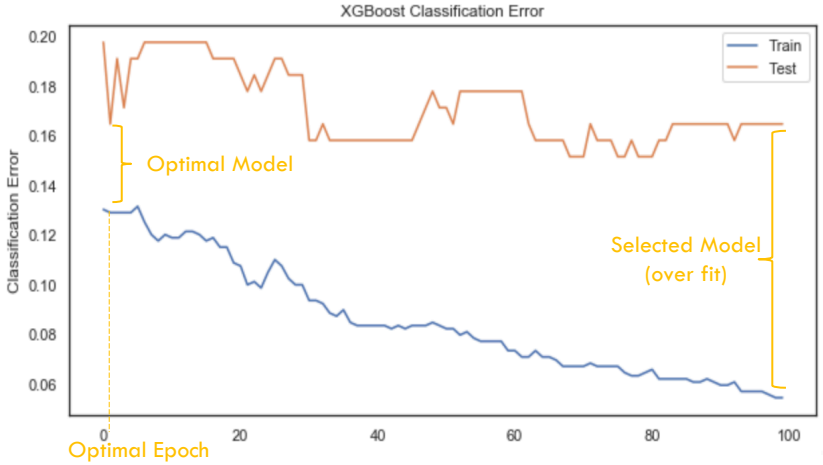
Each cycle included use of default parameters as the first approach.

The first approach in XG Boost was captured, where most values drove to >99% accuracy. It appeared the algorithm was over-fitting.

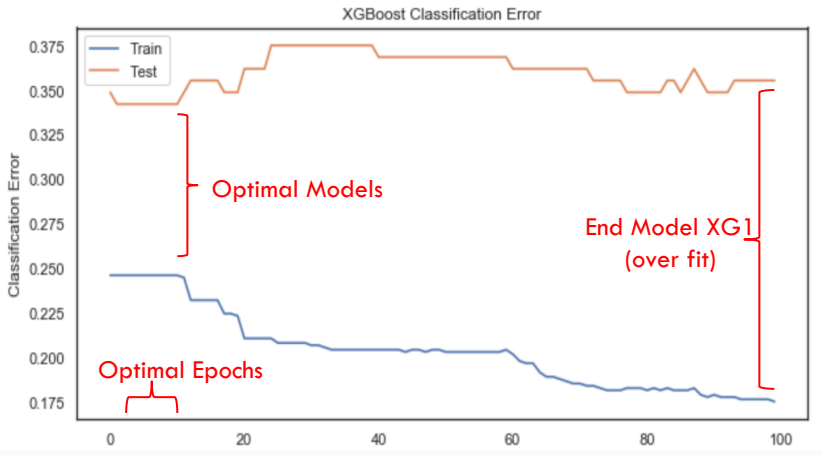
Based on the visual inspection of where best model performance occurred, the early_stopping parameter was applied.



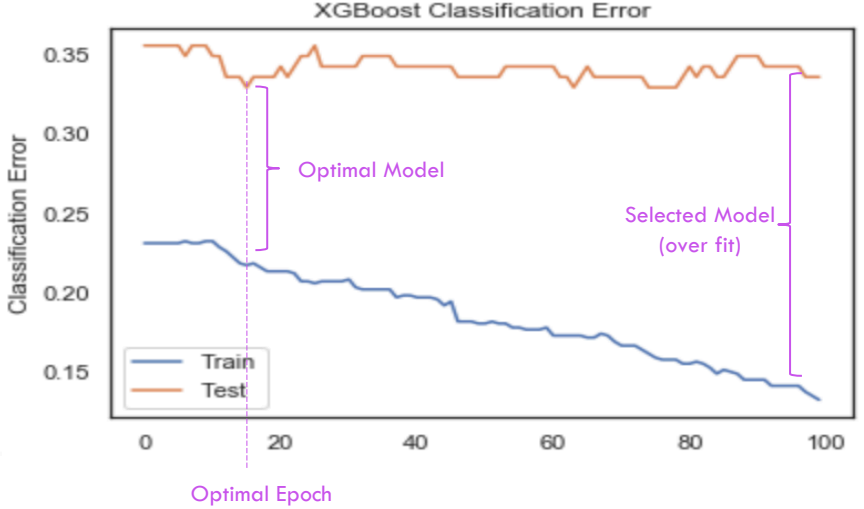
6 Features



Top from the 100



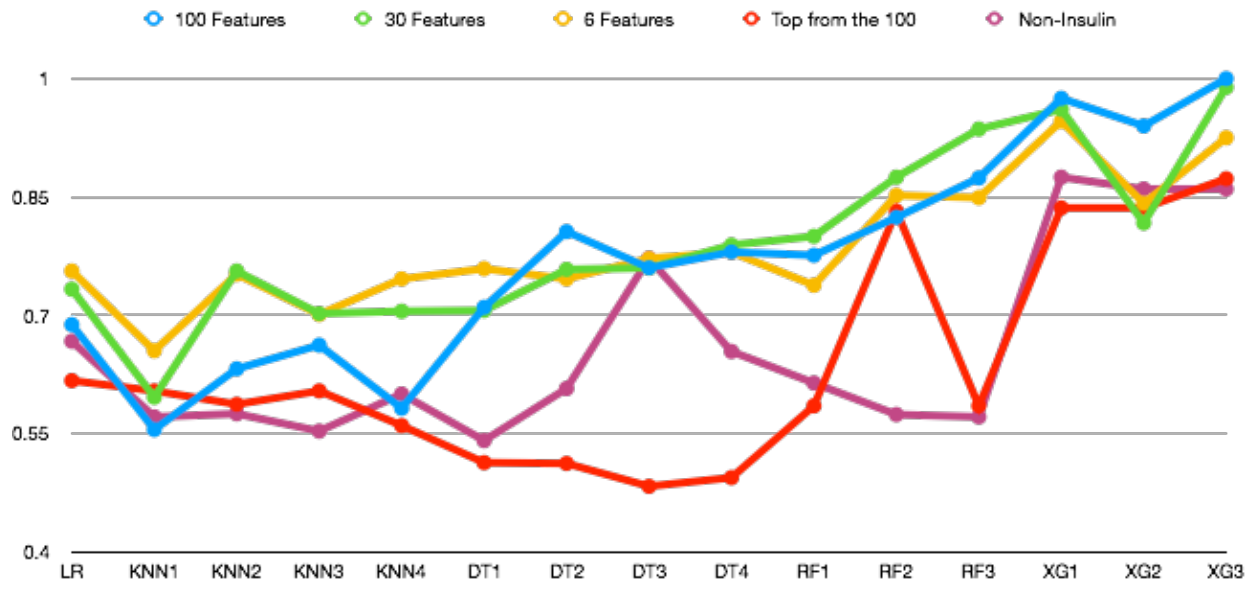
Non-Insulin



THE INSIGHTS

Can a diagnosis of diabetes be predicted from the given data set?

Based on the selected feature set (Non-Insulin Features)



Using the XG Boost model, the diabetes diagnosis can be predicted at an F-score accuracy of 62-87% (validation – test)

Given the size of the data set, it is most appropriate to use a small feature set – 7 features selected.

The successful model performances is highly associated to the Glucose feature, including the HbA1c calculation.

Interaction features appear to be most powerful for this data set and prediction use case. Successful features were created that I personally would not have thought to combine.

XG-boost tends to over-fit, yet can be controlled with the early stopping parameter.

Model development and tuning appears to be an art. I am not yet an artist.

Conclusion: While this work shows promise, there are needed refinements. Not ready for public use.