Gradient Boosting Binary Classification Model

Calculating Risk Prediction for Type 2 Diabetes

Martell Tardy, M.S.
Data Science Intensive Capstone Project
September 13, 2021 Cohort

Springboard

The Problem

How do you identify the leading factors of type 2 diabetes?

The Approach

To explore the relation between the behavioral traits and lifestyle and the probability of being diabetic using machine learning techniques to build a predictive model for the identification of risk factors for type 2 diabetes.

- identify portion of population that is already diagnosed
- identify factors that could potentially be linked to diagnosis
- build a model that predicts which factors have the highest impact
- report the leading factors of type 2 diabetes in the sample population

The Data

Date acquired: August 2, 2021

For the period: 2015

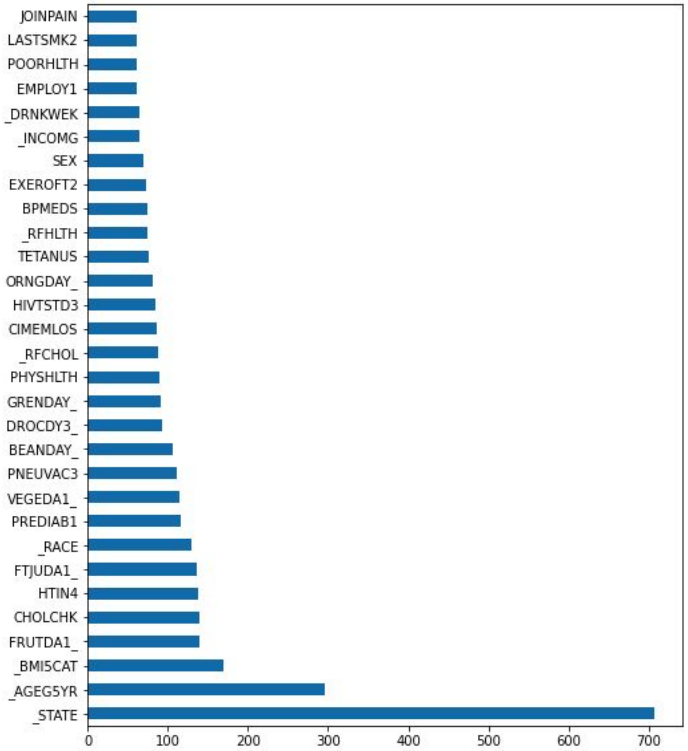
Number of records: 441,456

Number of fields: 330

Feature Selection

Process:

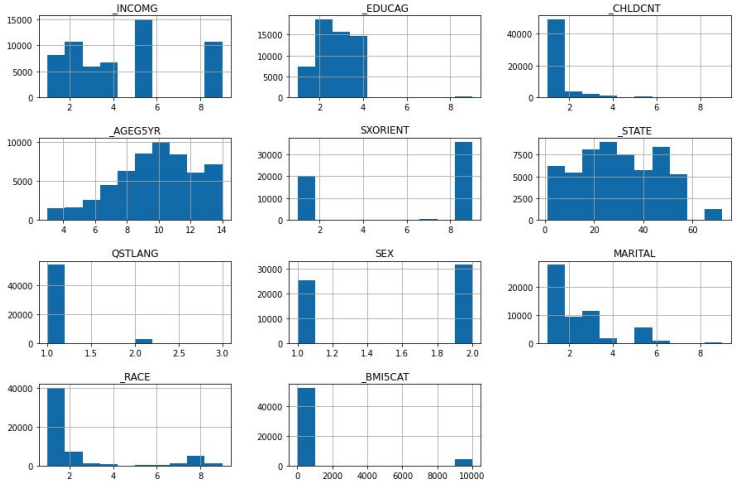
1. Manual feature reduction resulting in 144 fields
2. LightGBM classification for feature importance
3. Target encoding classification
4. Final feature selection of 25 fields (includes target)



Target Variable

Description:

- 56,748 participants (about 13% of the total surveyed population)
- Majority have an annual household income of $50,000 or more
- Majority have high school as the highest level of education
- Majority have no children
- Majority are between the ages of 65 to 69
- Majority are from the state of Kansas
- Majority are female
- Majority are married
- Majority identify racially as “White only, non-Hispanic”
- Majority BMI categorized as obese



Exploratory Data Analysis

10 Research Questions

Research question 1:

In the diabetic sample population `TYPE2`, is there a difference between males and females and the incidence of diabetes type 2?

Solution:

Yes, women are more likely to receive a diagnosis compared to men in the sample population.

Benchmark Model

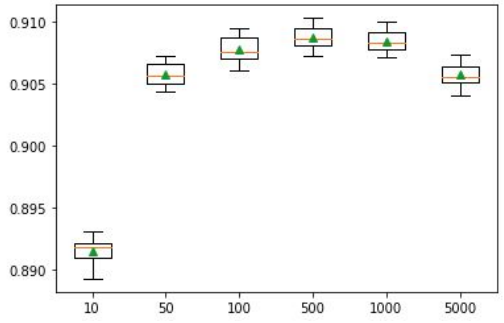
```
[ ] #define the model
    model = lgb.LGBMClassifier()

    #evaluate the model
    cv = RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=1)
    n_scores = cross_val_score(model, X, y, scoring='accuracy', cv=cv, n_jobs=-1)

    #report performance
    print('Accuracy: %.3f (%.3f)' % (mean(n_scores), std(n_scores)))

    Accuracy: 0.908 (0.001)
```

*using only default parameters

Hypertuning Parameters

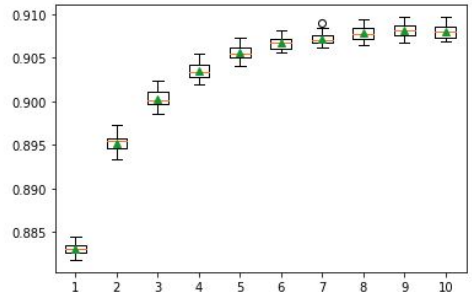
Number of Trees (n_estimators) = **500**

```
>10 0.892 (0.001)
>50 0.906 (0.001)
>100 0.908 (0.001)
>500 0.909 (0.001)
>1000 0.908 (0.001)
>5000 0.906 (0.001)
```



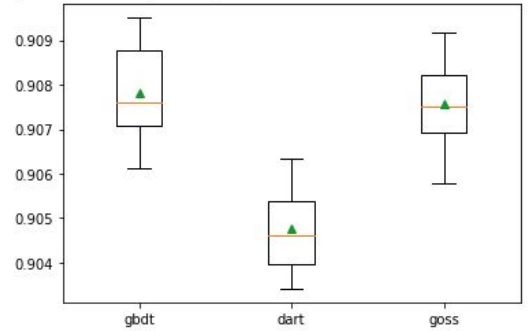

```
>1 0.883 (0.001)
>2 0.895 (0.001)
>3 0.900 (0.001)
>4 0.904 (0.001)
>5 0.906 (0.001)
>6 0.907 (0.001)
>7 0.907 (0.001)
>8 0.908 (0.001)
>9 0.908 (0.001)
>10 0.908 (0.001)
```

Tree Depth (max_depth) = **9**



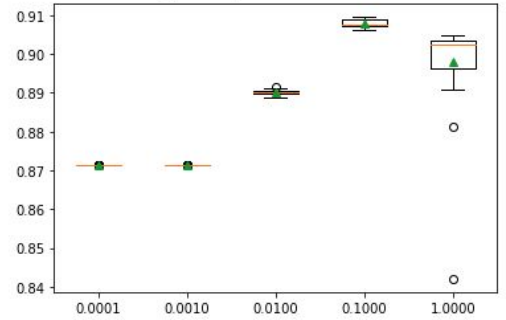

```
>gbdt 0.908 (0.001)
>dart 0.905 (0.001)
>goss 0.908 (0.001)
```

Boosting Type = **gbdt**




```
>0.0001 0.871 (0.000)
>0.0010 0.871 (0.000)
>0.0100 0.890 (0.001)
>0.1000 0.908 (0.001)
>1.0000 0.898 (0.012)
```

Learning Rate = **0.1**



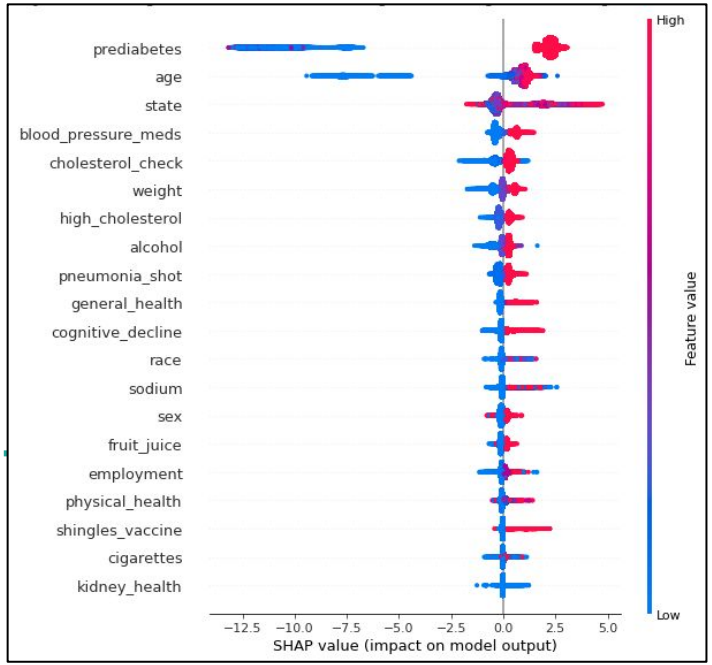
The Model

The goal of our binary classification problem is to create a machine learning model that makes a prediction in situations where the thing to predict can take one of just two possible values. For this study, we want to predict whether a person is diabetic(1) or not diabetic(2) based on the 24 predictor variables selected.

	precision	recall	f1-score	support
0	0.92	0.98	0.95	115306
1	0.77	0.40	0.53	17131
accuracy			0.91	132437
macro avg	0.85	0.69	0.74	132437
weighted avg	0.90	0.91	0.89	132437

Increase of 0.8% from the benchmark model with an accuracy of **91.5%**

Applying SHAP



5 Leading Predictors:

Prediabetes, Age, State, Blood Pressure Medication and Cholesterol Check

- Specifically, we can conclude that if a participant has prediabetes or has had prediabetes before this is the leading indicator for a type 2 diabetes diagnosis.
- As the age of a participant increases so does the likelihood of that participant having a type 2 diabetes diagnosis.
- If a participant takes blood pressure medication they are more likely to have a type 2 diabetes diagnosis.
- Lastly, participants who got their cholesterol checked "within the last year" and those who answered "don't know" are more likely to be diagnosed with type 2 diabetes.