Predicting Diabetes

Elisabeth Johnson

Metis 2022

Classification Module

How prevalent is Diabetes?

Around 1 in 10 Americans (~37M) have diabetes and of this population, around 90%-95% of them have Type 2 Diabetes.

Although a common disease, it can be prevented or delayed with the proper lifestyle changes. This analysis serves to help those at risk identify the possibility for Diabetes so that they may seek appropriate medical treatment.

What Kinds of Diabetes are there?

Type 1

- The failure of insulin production due to the body's immune system malfunctioning.
- Typically occurs in adolescence
- Very rare condition (less thank 200K cases per year)

Type 2

- The failure of insulin production due to the body's inability to produce an adequate supply of insulin.
- Typically occurs in those over 45 years old
- Very common condition (more than 1.4M per year)

How do we get started predicting Diabetes?

Data Cleaning, resolving class imbalance and feature selection.

Feature normalization and noise reduction.

Model Building and testing.

Predictors

#	Column
0	Diabetes_012
1	HighBP
2	HighChol
3	CholCheck
4	BMI
5	Smoker
6	Stroke
7	HeartDiseaseorAttack
8	PhysActivity
9	Fruits
10	Veggies
11	HvyAlcoholConsump
12	AnyHealthcare
13	NoDocbcCost
14	GenHlth
15	MentHlth
16	PhysHlth
17	DiffWalk
18	Sex
19	Age
20	Education
21	Income

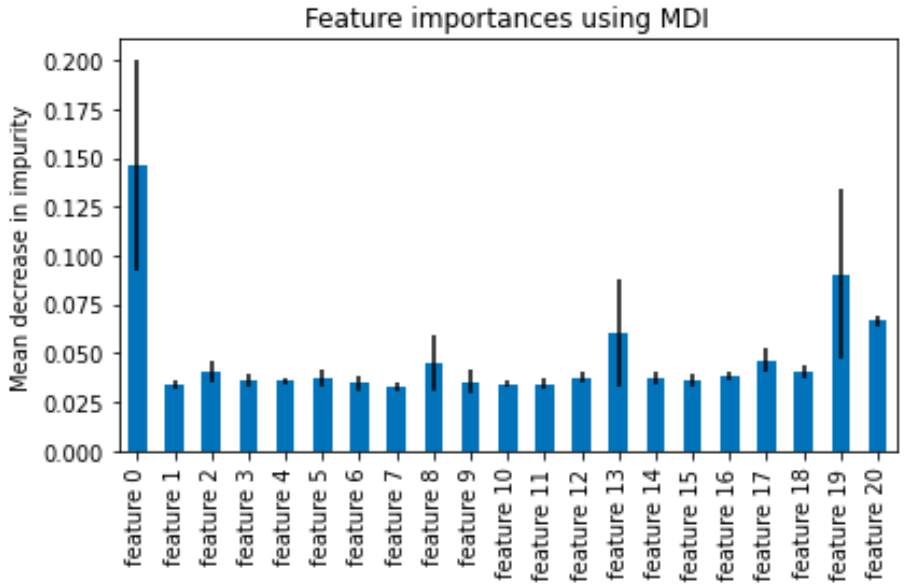
Target

Diabetes presence

Feature Selection

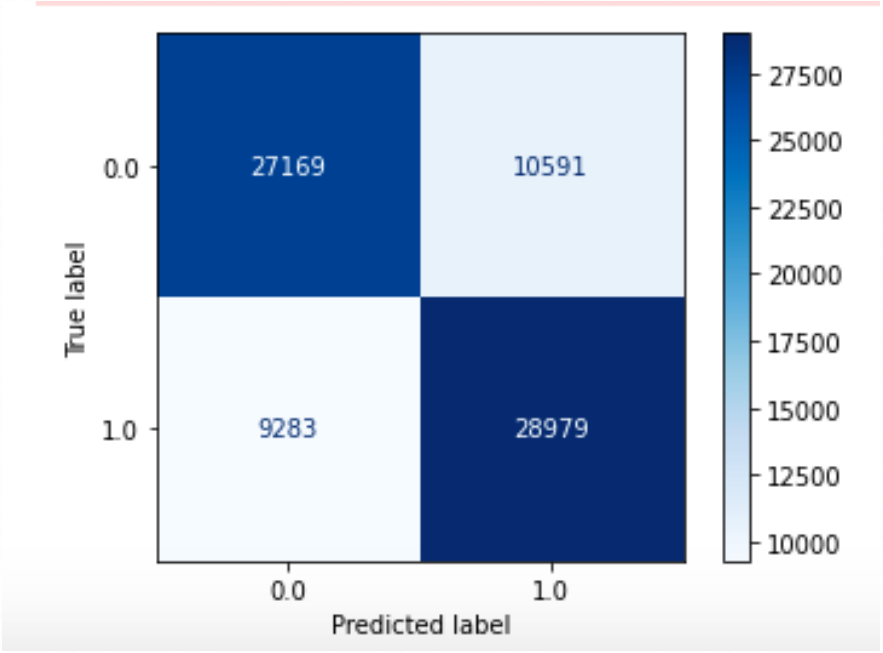
A lot of features are correlated with one another so we chose to use PCA decomposition which serves to reduce the dimensionality of features that are not as important.

Blood pressure and, surprisingly, education and income are three of the most important variables when it comes to Diabetes prediction.



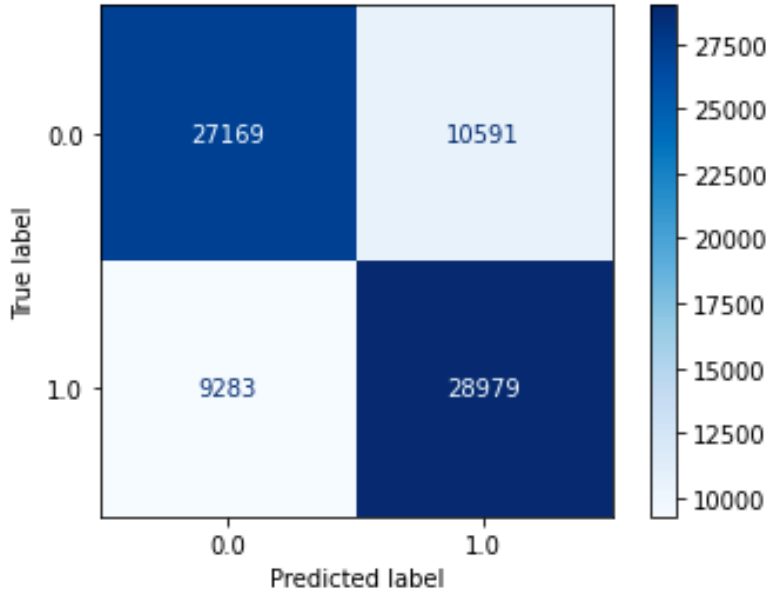
Logistic Regression Model

	precision	recall	f1-score	support
0.0	0.75	0.72	0.73	37760
1.0	0.73	0.76	0.74	38262
accuracy			0.74	76022
macro avg	0.74	0.74	0.74	76022
weighted avg	0.74	0.74	0.74	76022



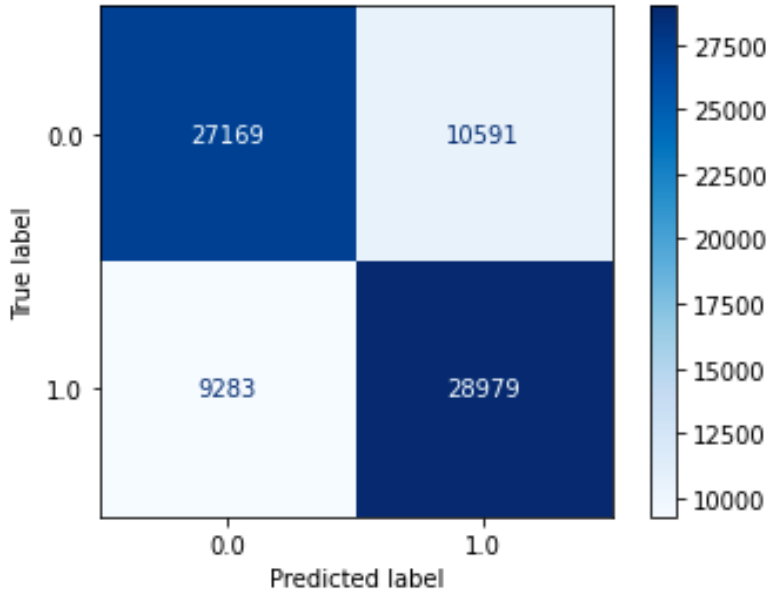
Random Forest Model

	precision	recall	f1-score	support
0.0	0.99	0.90	0.94	37760
1.0	0.91	0.99	0.95	38262
accuracy			0.95	76022
macro avg	0.95	0.95	0.95	76022
weighted avg	0.95	0.95	0.95	76022



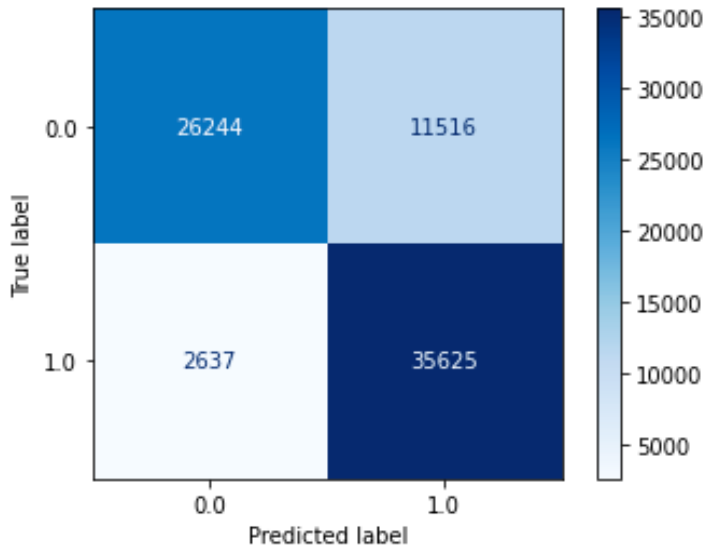
Gradient Booster Model

	precision	recall	f1-score	support
0.0	0.76	0.70	0.73	37760
1.0	0.73	0.78	0.75	38262
accuracy			0.74	76022
macro avg	0.74	0.74	0.74	76022
weighted avg	0.74	0.74	0.74	76022



KNN Model

	precision	recall	f1-score	support
0.0	0.91	0.70	0.79	37760
1.0	0.76	0.93	0.83	38262
accuracy			0.81	76022
macro avg	0.83	0.81	0.81	76022
weighted avg	0.83	0.81	0.81	76022



Conclusion

Overall our top performing model is Random Forest! This could be because RF has a more sophisticated way for dealing with overfitting. Because our model has over 15 predictors, our particular model may be prone to overfitting.

Resources Used

kaggle

CDC

WebMD

Thank You!