Predicting Diabetes Without Lab Tests

Kristin Cooper | May 23, 2021

Background & Stakeholders

In 2015, diabetes was the seventh leading cause of death in the United States. More than 30 million Americans are living with diabetes, and another 86 million are living with prediabetes.

10.5%
Percent of the US population with diabetes

$327B
Total cost of diagnosed diabetes in the US in 2017

2.3x
Increase in average medical expenditures for people with diagnosed diabetes

Many stakeholders in healthcare and public sectors - such as **public health agencies, hospitals and healthcare providers, insurance companies, employer benefits coordinators, and health & wellness companies** - are investing in programs to reduce healthcare costs and improve quality and length of life for Americans living with or at risk for diabetes.

Data

Using the CDC's National Health and Nutrition Examination Survey (NHANES) data from 2017-2018, a classification model has been developed to predict diabetes or prediabetes using only demographic and easily-accessible body measurement data.

This model is intended to be used by healthcare and public sector organizations to target outreach, advertisement, and investment in preventative care, social determinants of health, health & wellness, and health literacy programs.

Sample Size:
After merging data across source files, 5,951 labeled samples were split into:
- 4,760 elements to train models
- 1,191 elements to validate model performance

Predictors:
37 total features comprised of demographic, physical activity, nicotine usage, and health insurance survey data as well as body measures and pulse were modeled.

Target Class Calculation:
The target class - Diabetic/ Prediabetic (1) or Normal (0)- was calculated based on measured A1C levels using the Mayo Clinic's guidance*.

*The Mayo Clinic suggests A1C levels >6.5 indicate diabetes, >5.7 and <6.5 indicate prediabetes, and <5.7 indicates normal blood sugar health.

Modeling Process

Multiple algorithms with various parameters were iteratively trained on the sample data in order to develop the best possible model.

Methods conducted:

- Logistic Regression with Cross-Validation
- K-Nearest Neighbors
- Decision Trees, Bagged Trees, and Random Forest
- Boosting Ensemble Methods including AdaBoost, GradientBoost, & XGBoost
- Support Vector Machine
- Grid Search

Model complexity was also iterated, with one set of models leveraging all 37 predictors available and another set focusing on the top 12 features identified in the initial models.

How To Evaluate The Models

The primary metric used to evaluate this particular model is the **Recall Score**, which represents *out of all the true diabetic/prediabetic people, how often did the model predict the correct diagnosis?* Consequences for predicting a diabetic/prediabetic person is healthy are significant.

As a Patient...
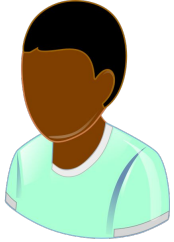
- I miss out on a potentially life-changing program that may have given me the resources and knowledge I need to reverse prediabetes or other risk factors, or manage my diagnosed diabetes

- My quality of life is lower

- I spend more on healthcare in my lifetime

- My life expectancy may decrease

As a Stakeholder...

- My program did not reach the population it would most benefit

- I did not spend my marketing budget as effectively as I could have

- We did not reach our target metrics in:
  - Hospital utilization
  - Readmission rates
  - Insurance claims
  - Employee productivity
  - Population health

Conversely, the **Precision Score** representing *out of all the predicted positive diagnoses, how many were actually diabetic/prediabetc* is less consequential. While some marketing dollars may go towards already healthy people, greater access to health programs never hurts!

Model Summary

Overall, the best models are consistently performing better than random chance and simple stratification models.

The maximum **recall scores** across all models is consistently around .75-.8, meaning **75-80% of true diabetic/prediabetic diagnoses are correctly predicted by the model**. The best performing models resulted in between **8.5-9.5% false negatives**, or diabetics who were incorrectly predicted to be healthy.

The highest accuracy scores are consistently between .71 and .76, indicating that t**he model predicts the correct diagnosis about 71-76% of the time**.

Top 5 Performing Models:

| Model	Label	Train Recall	Test Recall	Train AUC	Test AUC	Train Accuracy	Test Accuracy	False Negatives	False Negatives Normalized
SVC(kernel='linear', probability=True, random_state=610)	linear SVC limited feature set scaled	0.786700	0.777500	0.811500	0.818300	0.749000	0.745600	105.000000	0.088161
LogisticRegressionCV(cv=3, random_state=610)	default params, scaled data	0.786700	0.766900	0.823600	0.826900	0.757300	0.757300	110.000000	0.092359
KNeighborsClassifier(n_neighbors=23)	k=23, scaled data	0.826200	0.762700	0.839800	0.778700	0.756900	0.711200	112.000000	0.094039
LogisticRegressionCV(cv=3, random_state=610)	default params, unscaled data	0.764200	0.754200	0.802800	0.805600	0.742000	0.735500	116.000000	0.097397
LogisticRegressionCV(cv=3, random_state=610)	limited feature set; unscaled	0.777400	0.752100	0.812200	0.819700	0.748800	0.743900	117.000000	0.098237

Most Important Features

Coefficients generated by the models tell us how the model uses each feature to make predictions.

The **importance** of the feature to the model's predictions is measured by the absolute value of the coefficient. Coefficients near zero are not used much by the model.

Positive coefficients indicate the feature has a positive relationship with diagnosis. As the feature value increases, the probability of a diabetes diagnosis also increases.

Negative coefficients indicate the features has a negative relationship with diagnosis. As the feature value increases, the probability of a diabetes diagnosis decreases.

Support Vector Machine Model Coefficients (sorted by absolute value)
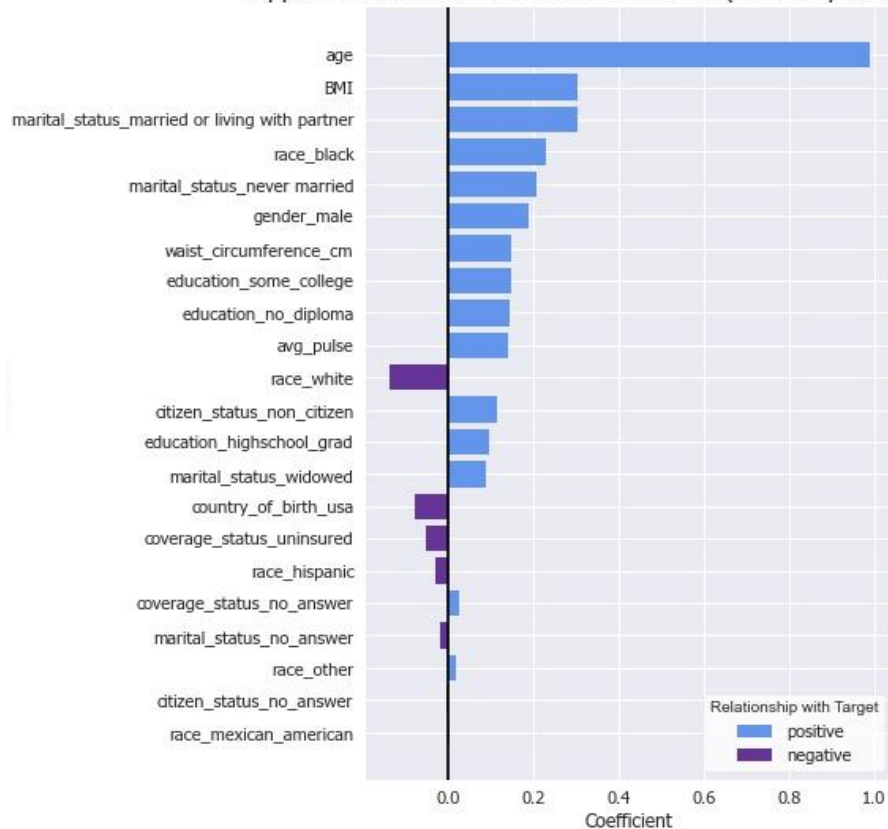
Recommendations

The model as-is can be used to target outreach, advertisement, and investment in preventative care, social determinants of health, health & wellness, and health literacy programs.

➔ **Target older or aging populations.**
Per the CDC, Type 2 diabetes most often develops in people over age 45. The model observed a similar pattern.

➔ **Target individuals with rising BMI, waist circumference, and average heart rates.**
Observing a rise in these 3 body measurements may be a good indicator of a person's rising risk of diabetes.

➔ **Consult with subject matter experts before using sensitive information such as race to target populations.**
While this should not be considered a causal relationship, the model observed that black individuals are more likely to have diabetes and white individuals are less likely to have diabetes. There are many underlying factors not captured in this analysis that may be more direct predictors

This model should <u>not</u> be used in place of physician advice and care plans.

Model Limitations & Future Enhancements

Please note the following caveats and suggested future enhancements for this model.

Model Limitations:

- Coefficients represent observed relationships of a relatively small sample and should not be considered causal. There are countless factors that influence a person's health such as family history, living environment, and many more.

- This model is not intended to predict diabetes in children under 18 or pregnant women.

- This model was trained on a sample comprised only of individuals in the USA. Some data, notably insurance coverage status, is quite specific to this sample.

Future Enhancements:

- Add samples to training set to improve performance.

- Conduct additional analysis with subject matter experts to consider the underlying reasons for some observations such as the correlation between race and diagnosis.

- Create a multivariate classifier to differentiate between prediabetes and diabetes.

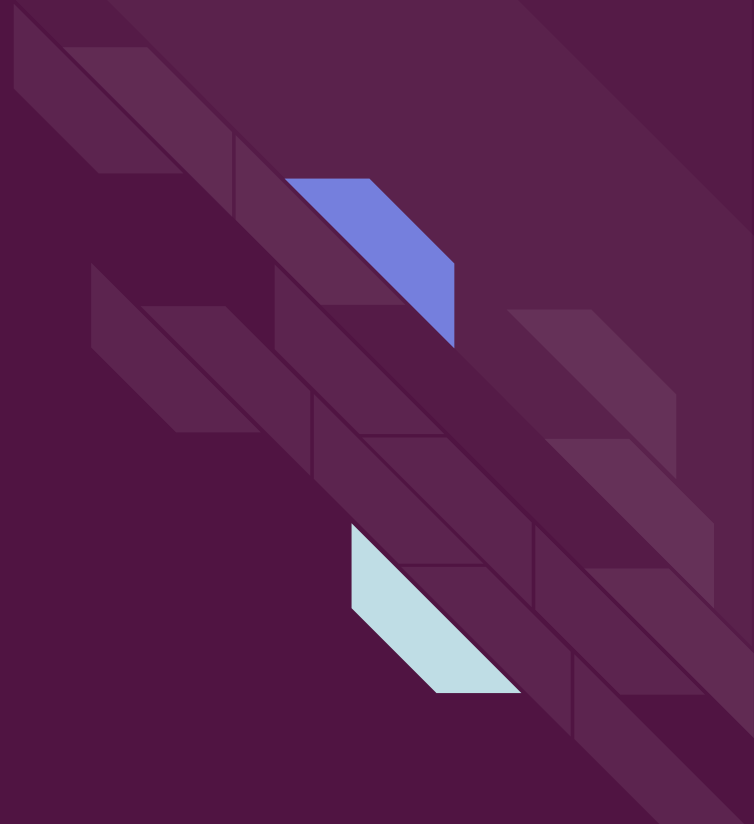- Incorporate time-series using data from prior year NHANES.

Thank You!

Full Report: https://github.com/kcoop610/phase-3-project
Email: kcoop610@gmail.com
LinkedIn: https://www.linkedin.com/in/kristincooper16

Appendix

Best-Performing Model - Report, Confusion Matrix, & ROC curve

CHECK ACCURACY, PRECISION, RECALL, & F1 SCORE - seek to maximize recall and accuracy

	precision	recall	f1-score	support
0	0.81	0.71	0.76	719
1	0.63	0.75	0.69	472
accuracy			0.73	1191
macro avg	0.72	0.73	0.72	1191
weighted avg	0.74	0.73	0.73	1191

****

CHECK Test ROC CURVE - seek to maximize area under the curve

AUC: 0.8092321609580161