Predicting 2015 American Diabetes Cases via Health Indicators

Krish Bansal, Pierre Quereuil, & Michael Puglise
December 15, 2025

Introduction

Diabetes remains one of the most prevalent chronic diseases in the United States, affecting over 38 million Americans, or roughly one in ten people. Type 2 diabetes accounts for the majority of cases, and tens of millions more have prediabetes, a condition characterized by elevated blood sugar and an increased risk of developing diabetes. While several contributing factors of diabetes and prediabetes are well established, including obesity, high blood pressure, and poor diet, the complex interactions among these and other health indicators are still an active area of research. Understanding these relationships is critical to identifying high-risk populations and informing public health interventions.

Motivated by this context, the present case study focuses on analyzing data from the 2015 CDC Behavioral Risk Factor Surveillance System (BRFSS). The BRFSS is a large-scale annual telephone survey of approximately 400,000 Americans that collects detailed information on health-related behaviors, chronic health conditions, and use of preventive services. This dataset provides a rich source of information for examining potential predictors of diabetes and prediabetes at the population level.

The primary goal of this project is to develop predictive models that estimate an individual's risk of diabetes using a subset of survey responses. Specifically, we aim to answer the following research questions:

1. Can survey responses from the BRFSS accurately predict whether an individual has diabetes?
2. Which health indicators and behaviors are the most predictive of diabetes risk?
3. Can a reduced set of features provide reliable predictions, enabling the creation of a short-form screening tool?

Although the data are from 2015, understanding the relationships between different health indicators and the risk of prediabetes and diabetes can help inform public health officials, healthcare providers, and researchers. The hope is that this research can help identify high-risk populations, guide doctor-patient conversations, and better inform the public about early-intervention strategies.

To achieve these goals, we combine exploratory data analysis (EDA), feature selection, and predictive modeling. The analysis starts with an examination of the distributions of key health indicators and their associations with diabetes status, followed by an assessment of useful interaction effects and variable transformations. We then train and evaluate ordinal logistic regression models using k-fold cross-validation. Model selection is based on threshold-independent metrics such as quadratic weighted kappa, log loss, mean absolute error, and macro AUC. Once the best model is chosen, threshold-dependent metrics such as One-vs-All (OvA) ROC, F1, precision, and recall can be used to determine an optimal probability cutoff. This framework provides insights into the most predictive health indicators and supports both clinical practice and public health applications.

Methodology

(Make sure to label figures numerically with descriptive captions)

Data Description / Distribution

- Used diabetes _ 012 _ health _ indicators.csv dataset

The *diabetes _ 012 _ health _ indicators.csv* dataset was chosen for

of 253,680 survey responses to the CDC's BRFSS2015. The target variable Diabetes_012 has 3 classes. 0 is for no diabetes or only during pregnancy, 1 is for prediabetes, and 2 is for diabetes. There is class imbalance in this dataset. This dataset has 21 feature variables

Goal is to correctly identify cases of diabetes.

Wish to predict 3 cases. ordinal relationship. non-diabetes< prediabetes < diabetes

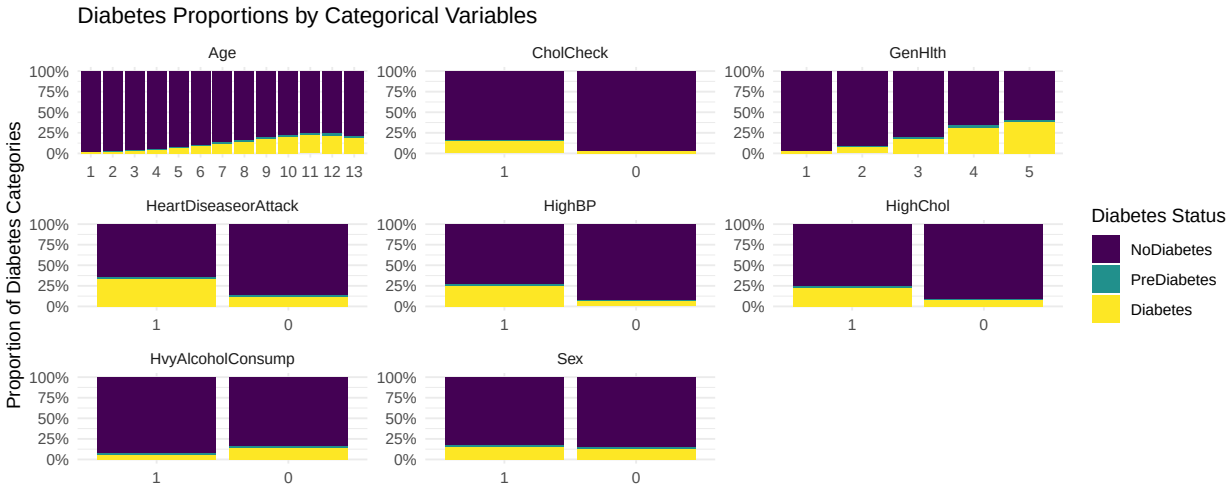
We model these statistics using an ordinal multinomal logistic regression model to capture the ordinal relationship (or severity levels of) non-diabetes (0), prediabetes (1), and diabetes (2).

Note that diabetes in this study is mostly type 2 (I think)

Data cleaning (created Diabetes_012 as a factor).

Train and test set. split

downsampling (tried upsampling but too expensive).



My figure. Most of the thing is prediabetes.

Instead of distributions .

Transformations

Interaction Effects

Results

what types of metrics used?

results of multinomal logistic regression model

results of SVMs, decision trees. pros and cons.

Assumptions and Uncertainty

Interpretation

This time around, need more plots showing degree of interactions and nonlinearities

Conclusion

In this statistical study, we have shown that it is possible to consistently and accurately predict cases of diabetes, prediabetes, and non-diabetes.

Not only do we achieve great performance on the testing set (computed through _____), but we preserve interpretability through _____.

These models already deliver great results, but we believe they could be reinforced and improved upon with future study and additional data. For instance, we only have significantly fewer observations for prediabetes compared to diabetes and non-diabetes, which caused _____. Perhaps study on the areas between our provided data would reveal even more patterns to exploit.

In conclusion, we recommend the use of ordinal multinomal logistic regression models with interaction effects to model varying degrees of diabetes. With continued study and more data, these could be refined further to shed even more insight on diabetes risk factors.