Predicting 2015 American Diabetes Cases via Health Indicators

Krish Bansal, Pierre Quereuil, & Michael Puglise

December 15, 2025

Contents

1 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?

While our data is 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. Strikingly, 1 in 5 diabetics and 8 in 10 prediabetics are not aware of their risk today. The only way to truly diagnose prediabetes or diabetes are through testing glucose levels, so our research focuses instead on helping identify high-risk populations, guiding doctor-patient conversations, and better informing the public about early-intervention strategies rather than replacing current diagnosis tools.

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 cumulative logistic regression models using k-fold cross-validation. Model selection is based on threshold-independent metrics such as BIC, 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.

2 Methodology

(Make sure to label figures numerically with descriptive captions)

2.1 Data Description

The dataset used for this analysis contains 253,680 survey responses with 21 features. These features include health-related responses to survey questions and derived variables summarizing participant health status. Our target variable, `Diabetes_012`, represents three ordered categories: non-diabetes (0), prediabetes (1), and diabetes (2). Type 2 diabetes accounts for the majority of cases in this dataset.

2.2 Data Preparation

Minimal data cleaning was performed to ensure proper variable types. The variables `BMI`, `MentHlth`, and `PhysHlth` were treated as numeric, while all other variables were converted to factors. The target variable `Diabetes_012` was encoded as an ordered factor to reflect the severity of disease.

The data were split into training and testing sets using an 80-20 split. Exploratory data analysis on the training set revealed that the classes were heavily imbalanced, with prediabetes being underrepresented (see Figures 1 and 2 below). To address this, downsampling of the majority classes was applied to improve model performance while keeping computation manageable.

2.3 Exploratory Data Analysis

We examined the distributions of categorical variables and visualized the proportion of each `Diabetes_012` category across levels of the features (see Figure 1). This helped identify variables with strong associations with diabetes status. Additionally, (Figure 2) helped us identify the need for data balancing as there were far more non-diabetes and diabetes case observations compared to prediabetes ones.

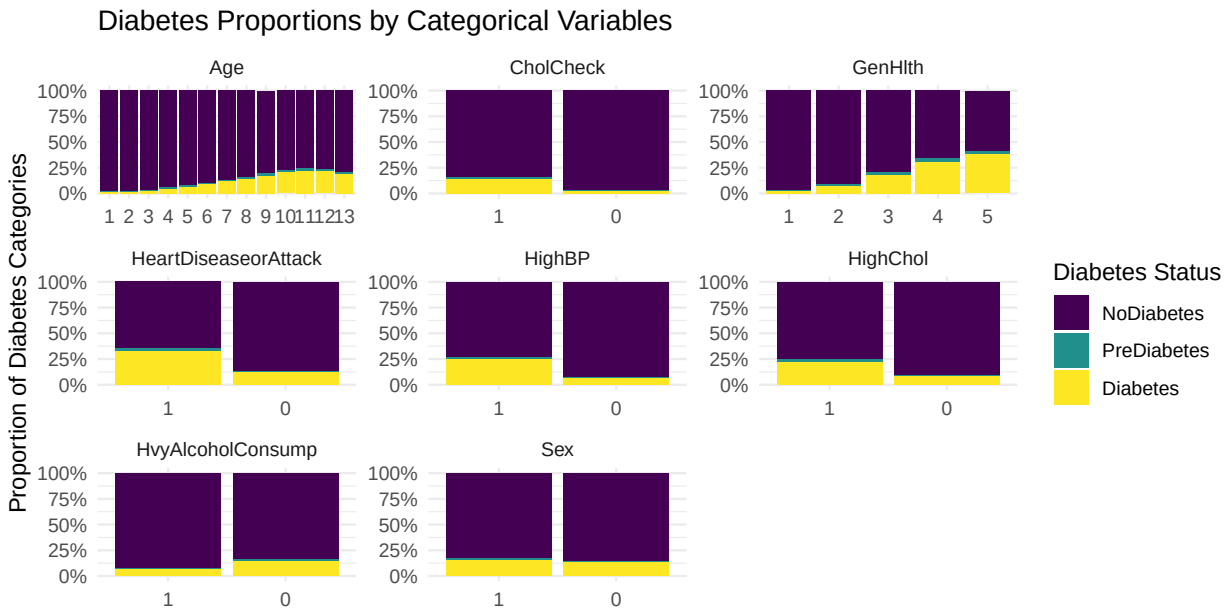

Figure 1: A scatter plot of random data.

Continuous variables such as `BMI`, `MentHlth`, and `PhysHlth` were visualized using histograms to assess their distributions and potential need for transformations. Interaction effects and transformations of variables were also explored to capture non-linear relationships that may improve model performance.

```
##   Diabetes_012      n
## 1   NoDiabetes 213703
## 2  PreDiabetes   4631
## 3     Diabetes  35346
```

After building intuition for the underlying feature distributions, our group experimented with various model types, transformations, and predictors to get best predictive accuracy.

2.4 Transformations

[Krish's work]

2.5 Interaction Effects

[Krish + Pierre]

2.6 Modeling Approach

To account for the ordinal nature of the outcome, we used cumulative logit classification, also known as proportional odds logistic regression. This approach models the ordered progression from non-diabetes to prediabetes to diabetes. Multiple model specifications were tested, including different combinations of predictors, interaction terms, and variable transformations.

replace with final model

$Diabetes \sim HighBP + HighChol + CholCheck + BMI + HeartDisease + HeavyAlc + GenHlth + Sex + Age + HighBP : GenHl$

Other Methods: Black Box Models (remove if we decide not to use)

While our main focus was to develop a model on a subset of predictors that allowed for interpretation of the driving factors of diabetes risk, we also considered several "black box" models that focused on predictive power rather than inference. - Random Forests: - Ordinal Forests: - Boosted Tree: - Support Vector Machine (SVM):

3 Results

what types of metrics used?

results of multinomal logistic regression model

results of SVMs, decision trees. pros and cons.

3.0.1 Assumptions and Uncertainty

3.0.2 Interpretation

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

4 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.