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 an 80/20 train-test split. Model selection was then performed 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 were used to determine an optimal probability cut-off. 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 in this analysis contains 253,680 survey responses with 21 health-related features. These features represent participants' answers to behavioral and medical questions, along with derived indicators of overall health status. Most binary variables follow a consistent coding scheme in which 1 indicates "yes" and 0 indicates "no." Similarly, the variable Sex is encoded with 0 representing male and 1 representing female.

Several originally numeric variables were pre-categorized into ordered groups. For instance, Age is represented as a 13-level categorical variable, where 1 corresponds to ages 18–24, 9 corresponds to ages 60–64, and 13 represents age 80 or older. Similar ordered categories also appear in variables related to income, physical health, and mental health.

The target variable, Diabetes_012, is an ordered factor capturing three diabetes-related outcomes: 0 for non-diabetes, 1 for prediabetes, and 2 for diabetes. Most individuals classified as diabetic fall into the Type 2 diabetes category, which is consistent with broader population trends.

2.2 Data Preparation

Minimal data cleaning was performed to ensure proper variable types. Only 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.

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.

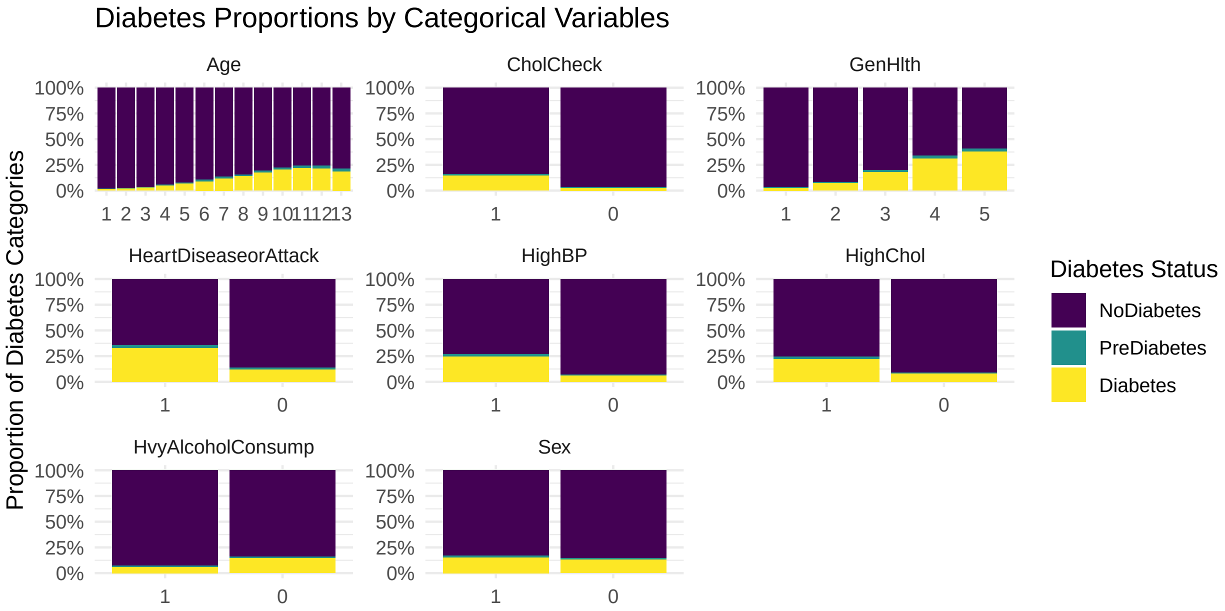

Figure 1: The distribution of categorical data collected from the 2015 BRFSS survey along with respective diabetes class

```
##   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 Feature Selection

The cleaned data set includes 21 feature variables. Many of these variables are highly correlated (eg. fruit consumption vs vegetable consumption), which can lead to multicollinearity. Hence, we first performed Ordinal Lasso Regression. We chose Lasso over Ridge regression because the coefficients of many of the feature variables were already extremely low/close to 0 in our baseline model: hence, Lasso would allow the coefficients to actually go to 0 - allowing us to make clearer choices on which features to select.

We also noted which coefficients had a sign change in order to see what direct coefficients of certain features were heading towards. Using the selected features, we were still left wih 15 candidate features. To further reduce our feature space, we performed stepwise selection criteria using BIC. We used BIC as our criteria

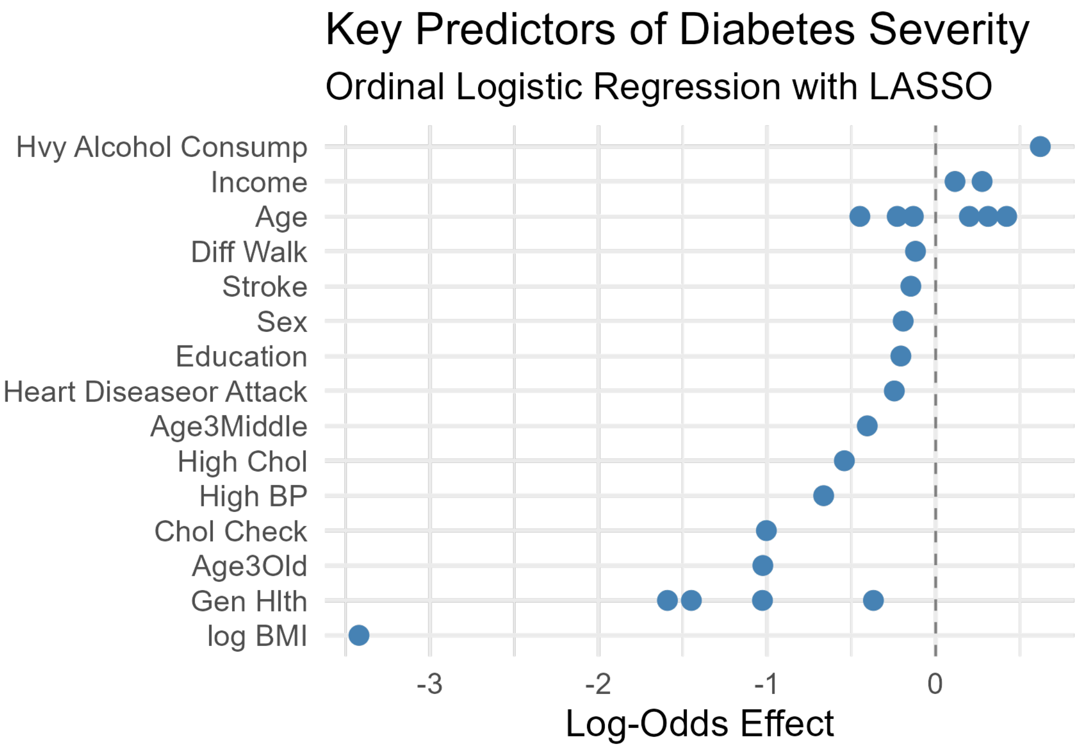

Figure 2: Key predictors of diabetes severity estimated using an ordinal logistic regression model with LASSO regularization.

because it penalizes larger feature models a lot more than other criteria (eg. AIC), hence resulting in a smaller feature set. The final feature set includes:

BMI, Age, HighBP, HighChol, Smoker, Stroke, HeartDiseaseorAttack, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, DiffWalk, Sex, Income

In addition to the statistical evidence for using these variables, these variables make medical sense as to why they would factor into a person having diabetes.

2.5 Transformations

Based on the features selected in the previous section, we decided to focus on discovering transformations that would improve the predictive power of the model. After investigating the distributions of each variable and various interaction terms, 2 transformations and an interaction term were ultimately added to the final model.

Firstly, we transformed the BMI feature by first logging and standardizing the data. Originally, the BMI data is right skewed. However, after the aforementioned transformations, the data became a lot more normalized which led to a significant increase in the model's predictive power.

Furthermore, age was also transformed. In our baseline model, we noticed that each level of age had a weak p-value associated with it. However, we collapsed age from 13 ordinal categories to just 3. The p-values associated with each value made the coefficients statistically significant and helped increase the model's overall predictive accuracy.

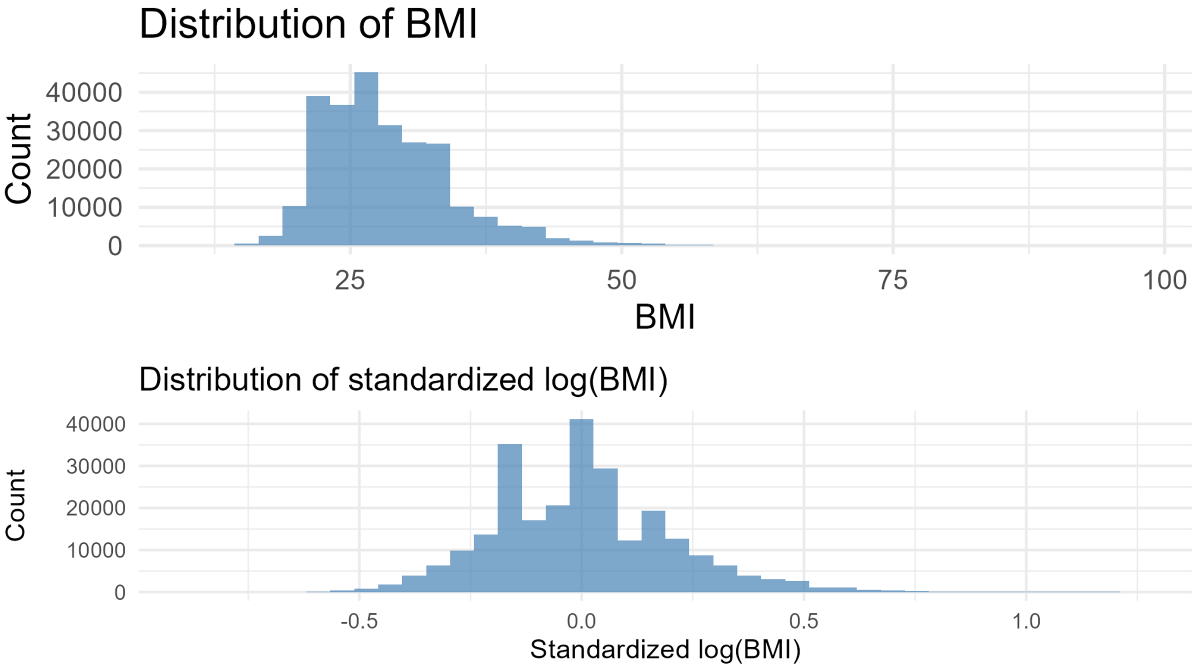

Figure 3: Distribution of Standardized log(BMI)

2.6 Interaction Effects

Within the chosen set of features, there could be a case made for many interaction terms to exist. For example, smoking directly causes higher blood pressure and cholesterol. Hence, these three variables should have interactions with each other. However, after looking at all subsets of models with different interaction terms, there was only one statistically significant interaction term which improved the model's prediction accuracy: *High Cholesterol * High Blood Pressure*.

From the case's perspective, this medically makes sense: higher cholesterol can lead to a build up of plaque in the arteries which directly leads to higher blood pressure.

2.7 Modeling Approach

To account for the ordinal nature of the response, `Diabetes_012`, we used cumulative logistic regression, also known as a cumulative logit model. This approach models the ordered progression from non-diabetes to prediabetes to diabetes by estimating cumulative probabilities rather than modeling each class independently. Specifically, the model defines cumulative probabilities $\pi_j(x) = P(Y \leq j|x)$, which represent the probability that an individual's diabetes status falls at or below category j given covariates x. Category-specific probabilities are then obtained as differences between successive cumulative probabilities. Unlike baseline-category multinomial logistic regression, cumulative logit models do not require the selection of a reference outcome class. Instead, the model estimates category-specific intercepts, or cutpoints, that serve as baselines along an underlying ordinal disease severity scale, while assuming that predictor effects remain constant across thresholds.

Our final model incorporates transformations of key continuous predictors to better capture nonlinear relationships, as well as an interaction effect between high blood pressure and high cholesterol. The final cumulative logistic regression model is given by: polr(Diabetes_012 ~ log_centered(BMI) + Age_3 +

HighBP, + HighChol + Smoker + Stroke + HeartDiseaseorAttack + HvyAlcoholConsump + AnyHealthcare + NoDocbcCost + GenHlth + DiffWalk + Sex + Income + HighBP*HighChol)

This modeling framework balances predictive performance with interpretability, making it well suited for understanding how health behaviors and indicators jointly influence diabetes risk across different levels of severity.

2.8 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):

[Pierre's work]

3 Results

[Pierre's work]

what types of metrics used?

results of multinomal logistic regression model

confusion matrix

results of SVMs, decision trees. pros and cons.

Talk about how model selection was performed based on threshold-independent metrics such as BIC, quadratic weighted kappa, log loss, mean absolute error, and macro AUC.

After the best model is chosen, analyze threshold-dependent metrics such as One-vs-All (OvA) ROC, F1, precision, and recall that we used to determine an optimal probability cutoff.

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 study, we demonstrate that survey responses from the BRFSS can be used to reasonably predict whether an individual has diabetes. Using cumulative logistic regression, we achieved strong performance on a held-out testing set, evaluated through classification accuracy and class-specific error rates, while maintaining a clear and interpretable modeling framework.

Through LASSO regularization and stepwise model selection using BIC, we identified a reduced set of influential health indicators that balances predictive performance with interpretability. In particular, our final model prioritizes minimizing false negatives (cases in which individuals with diabetes are incorrectly classified as non-diabetic) given the potential health consequences of missed risk identification. While more complex black-box models such as support vector machines, random forests, and boosted trees provided marginally better predictive power, we ultimately favored an ordinal logistic model to preserve transparency and insight into how specific risk factors contribute to diabetes severity.

Despite these encouraging results, some important limitations remain. Diabetes cannot be definitively diagnosed using survey data alone, as accurate classification ultimately requires clinical glucose testing. Consequently, our model is best suited for risk stratification and screening rather than diagnostic decision-making. Another key challenge in this analysis was the relatively small number of prediabetes observations compared to diabetes and non-diabetes cases, which introduced additional uncertainty and limited classification accuracy for this intermediate category.

Future work should focus on collecting more comprehensive prediabetes data to improve class separation and reduce uncertainty across disease stages. Additionally, the reduced set of influential predictors identified in this study could serve as the foundation for a more targeted, short-form screening tool to help flag individuals who may benefit from further clinical evaluation. Overall, we recommend the use of cumulative logistic regression models with interaction effects to the CDC as an effective and interpretable approach for classifying varying degrees of diabetes risk, with significant potential for refinement as additional data become available.