Evaluating Risk Factors for Type 2 Diabetes

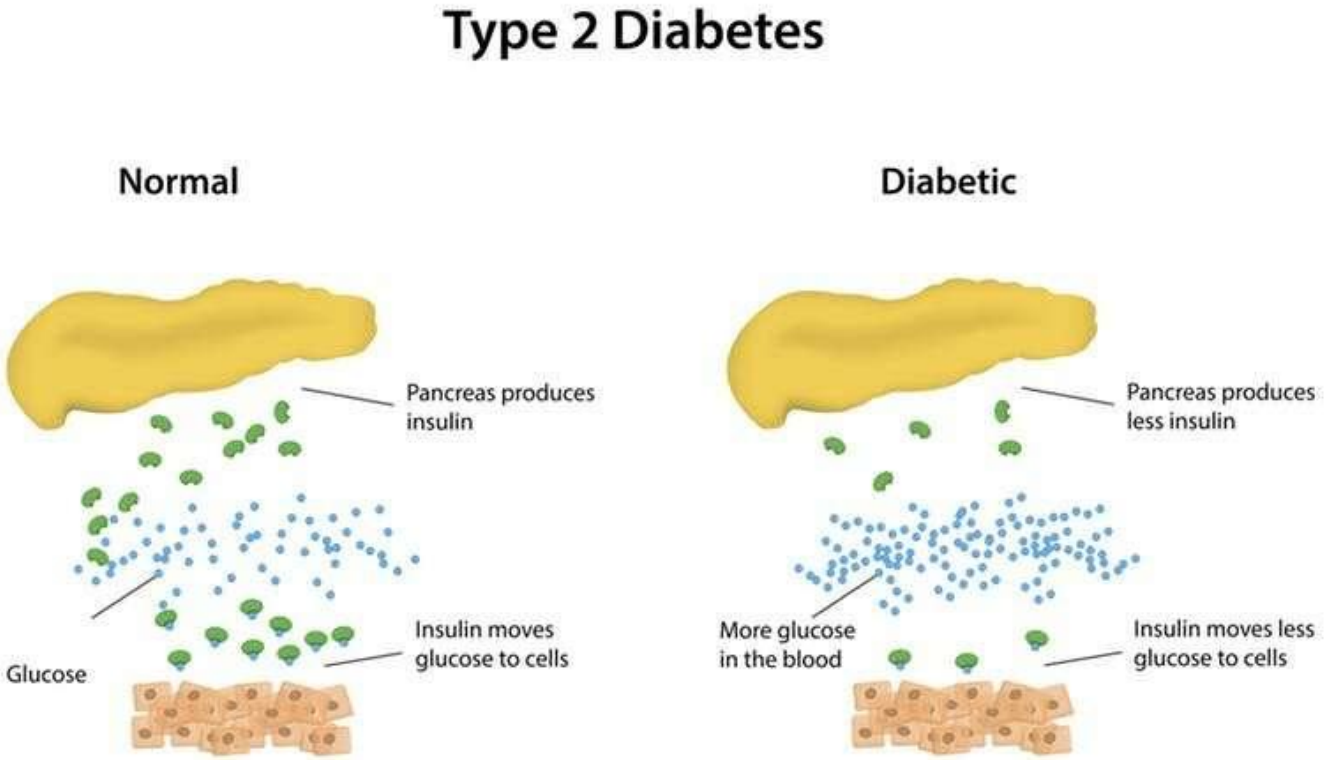
Kevin Sharp

Springboard

Data Science Capstone Project 1

What are the costs?

- 7th leading cause of death
- Higher risk of heart disease, kidney disease, and more
- $327 billion in medical costs
- $90 billion in reduced productivity



What can we do?

Type 2 diabetes is *preventable*.

If we can identify its risk factors, healthcare professionals will be able to use this information to identify at-risk patients

Early intervention will allow the disease to be stopped before it begins

A Classification Problem

Data Source: 2014 Behavioral Risk Factor Surveillance System

- Annual survey data collected by the CDC
- 279 features (survey items)
- 464,664 observations (survey respondents)

The survey data will be used to predict the likelihood that a given respondent has received a diabetes diagnosis.

Data Information

The data are encoded, with digits corresponding to survey response classes, as shown to the right.

To keep the classification model simple, the cleaned data set contains only strict “Yes” and “No” responses to the question of whether a respondent has been diagnosed with diabetes, which is our target variable.

25 of the features are selected for further analysis, and only respondents who answered all of these are retained.

Ever told) you have diabetes

Section: 6.12 Chronic Health Conditions **Type:** Num

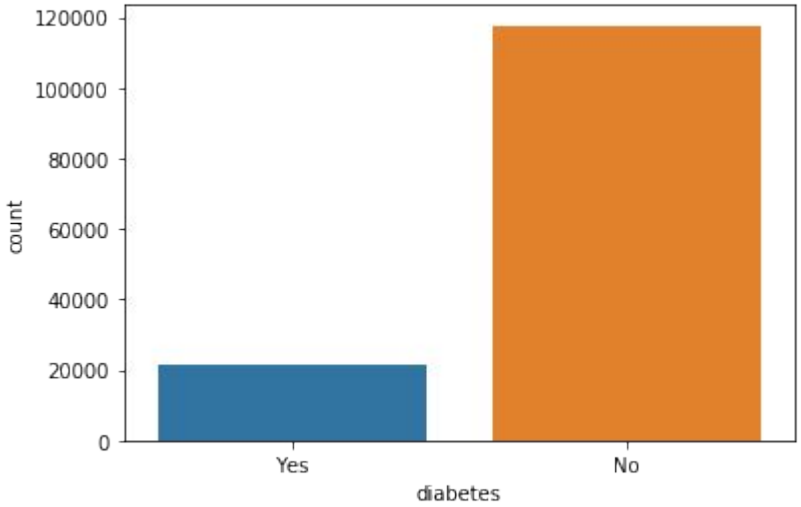
Column: 105 **SAS Variable Name:** DIABETE3

Prologue:

Description: (Ever told) you have diabetes (If "Yes" and respondent is female, ask "Was this only when you were pregnant?". If Respondent says pre-diabetes or borderline diabetes, use response code 4.)

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	61,118	13.15	10.53
2	Yes, but female told only during pregnancy—Go to Section 07.7.1 LASTDEN3	4,207	0.91	1.01
3	No—Go to Section 07.7.1 LASTDEN3	390,827	84.11	86.78
4	No, pre-diabetes or borderline diabetes—Go to Section 07.7.1 LASTDEN3	7,668	1.65	1.48
7	Don't know/Not Sure—Go to Section 07.7.1 LASTDEN3	551	0.12	0.14
9	Refused—Go to Section 07.7.1 LASTDEN3	291	0.06	0.05
BLANK	Not asked or Missing	2		

Exploratory Data Analysis - Part 1

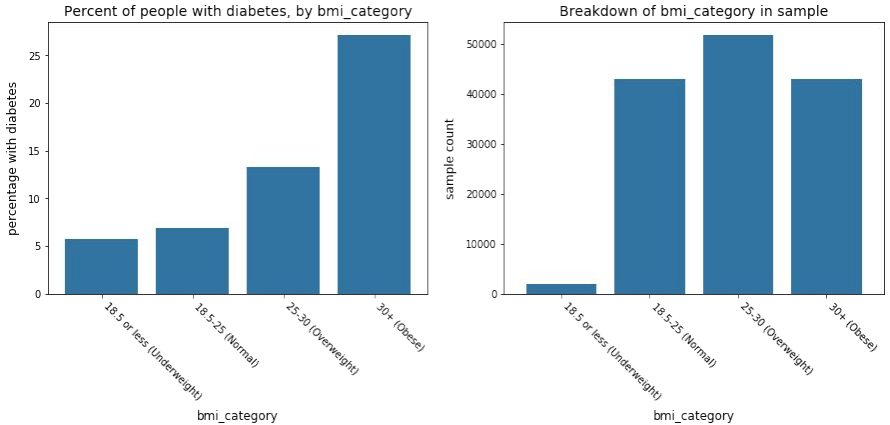


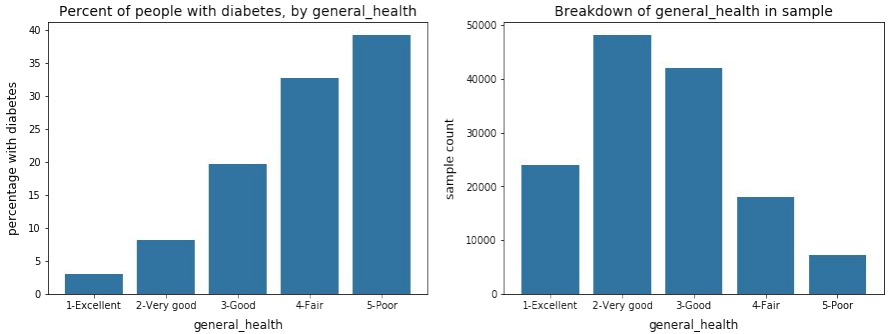
After cleaning the data, about 138,000 observations are retained.

21,587 (15%) of those report having a diagnosis of diabetes.

Since this is our target variable, our predictor classes are imbalanced, which will influence our modeling approach later.

Exploratory Data Analysis - Part 2

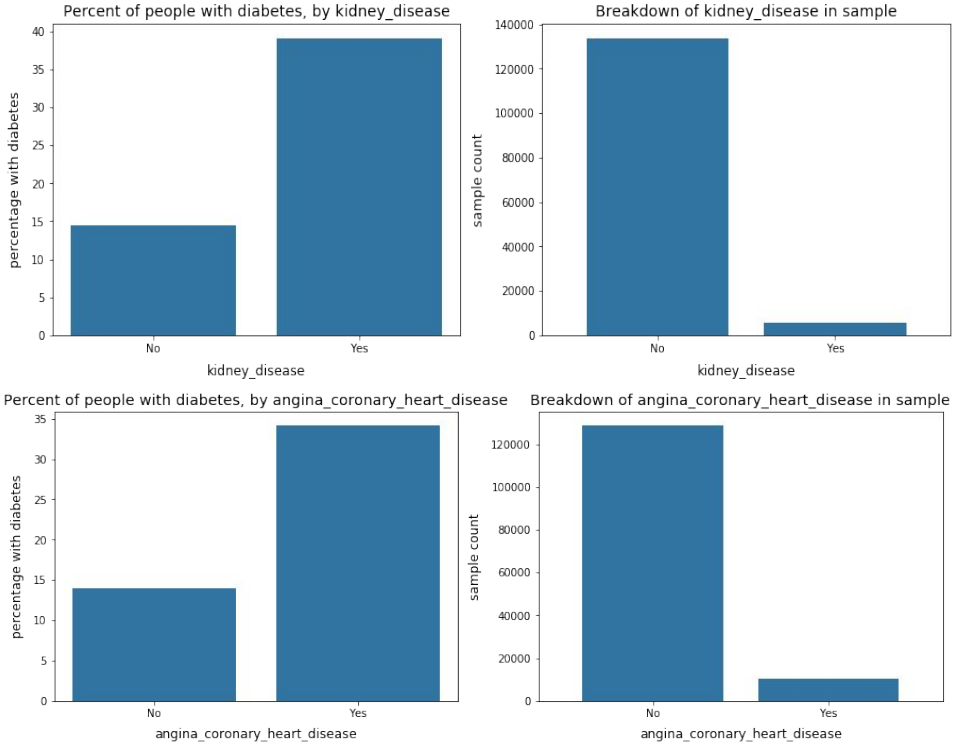




Two features that immediately appear to have a strong association to diabetes are General Health and BMI Category.

As general health worsens and BMI increases, the rate of diabetes increases.

Exploratory Data Analysis - Part 3



Recall that about 15% of this sample has a diabetes diagnosis.

Of those with kidney disease, nearly 40% also have diabetes.

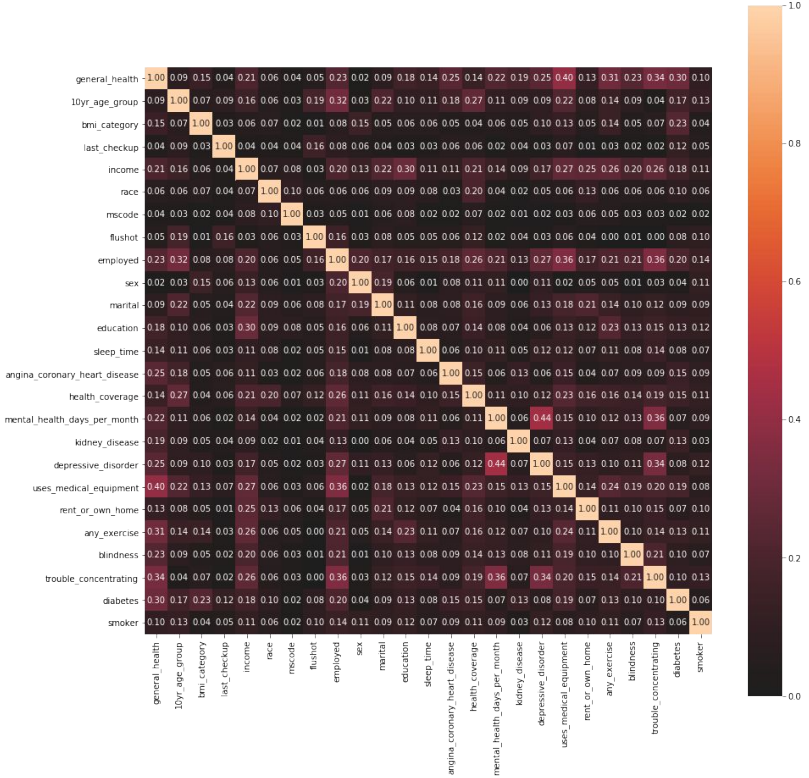
Of those with angina or other coronary heart disease, nearly 35% also have diabetes.

Exploratory Data Analysis - Part 4

To the right is a heatmap of *Cramer's V* statistics for all pairs of features. *Cramer's V* uses the *chi-squared* statistic for a given pair of features to compute their strength of association.

Although some are low, all associations between diabetes and selected features were found to be statistically significant ($a = 0.01$).

From this plot, there does not appear to be strong collinearity between predictive features.

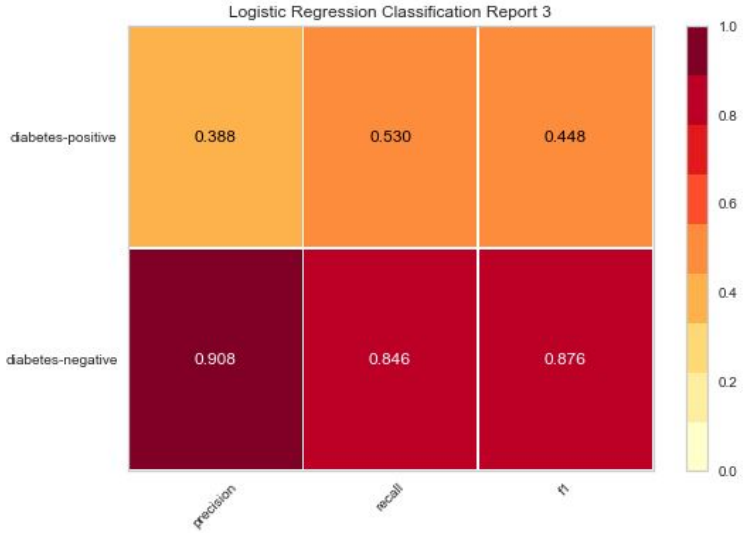


Machine Learning Overview

- Supervised Learning - all data are labelled according to the target variable.
- Since all predictive features are categorical and we are predicting a binary classification, Logistic Regression is well-suited to this problem.
- Imbalanced Data - only 15% have diabetes
- Tools: scikit-learn, yellowbrick

Modeling Steps

- Train-test split (25%-75%)
- Cross-validation (CV) for hyperparameter tuning
  - 5-fold CV with grid-search method
  - Include class weight parameters to account for imbalanced class proportions
  - Scoring metric: weighted f1 score
- Train the classifier using the selected optimal parameters



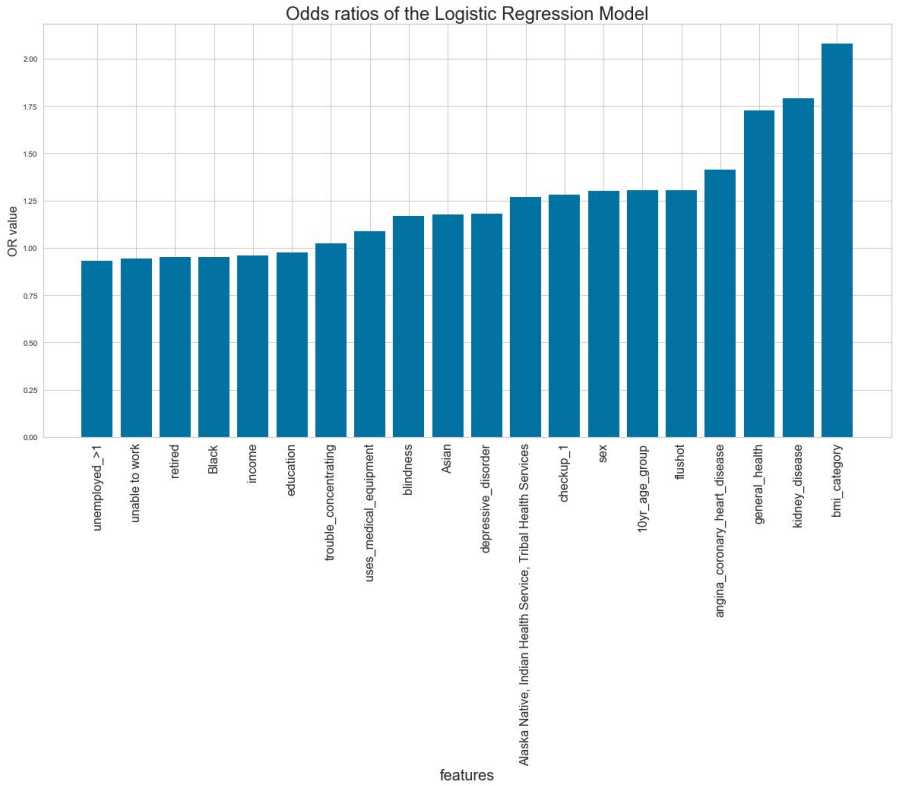
The final model's classification report demonstrates precision, recall, and f1 scores that are on par with other research in this area.

Deeper Analysis - Part 1

Shown to the right is the upper half of the calculated odds ratios for the model's features.

The four strongest predictors appear to be bmi_category, kidney_disease, general_health, and angina_coronary_heart_disease.

Other features with high predictive value are age group, sex (male) and blindness.

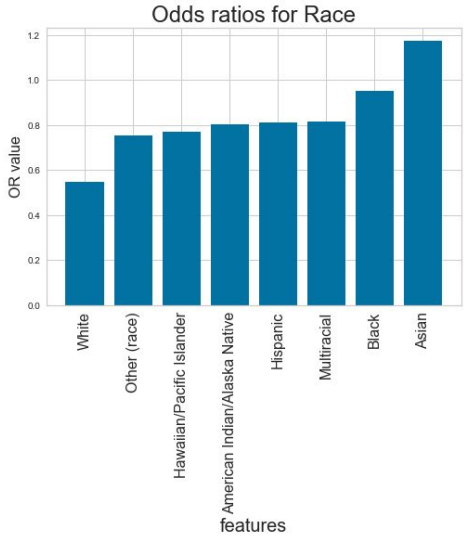


Deeper Analysis - Part 2

Other features result from splitting our original features into their individual categories.

Those who have had a checkup within the last year more frequently have diabetes.

This is likely because those who have received a diagnosis are frequently monitoring their condition.



Asian and black individuals seem more susceptible to diabetes than those of other races, while white individuals seem less susceptible.

Disclaimers and Areas for Improvement

- Due to the cross-sectional nature of the data, the model itself is not able to reveal insights about the causal relationship between features - perhaps a new kind of approach could address this. Statements I make about causality are sourced from the medical community.
- Since the data gathered is sourced from telephone surveys, it may be subject to recall bias. Clinical data and biomarkers could be included in the future to get a more objective picture of features such as “general health”.
- Type 2 diabetes is significantly more common than type 1, so measures taken to approximate the type 2 population, such as limiting the sample to those at least 30 years old, are likely sufficient for estimation purposes. However, survey or clinical data that makes a distinction between the two could lead to improved predictive models.

Recommendations

1. Those in poor health, particularly those who are overweight, should be made aware of their risk of diabetes and encouraged to improve their health before a serious condition develops.
2. Elderly patients, particularly male patients, black patients, and Asian patients should be made aware of their risk of diabetes and the importance of maintaining their health since these risk factors are outside of a person's control.
3. Kidney disease, heart disease, and blindness should be indicated as possible complications that may result from diabetes.