Diabetes Prediction & Risk Factor Analysis

Understanding Differences in Prediction of Prediabetes and Diabetes

Taylor Erickson
Applied Computer Science
University of Colorado Boulder
Boulder Colorado US
taer7274@colorado.edu

ABSTRACT

Prediabetes and Type II diabetes are both associated with many deleterious health outcomes. However, investment into diagnosis and treatment historically has been primarily focused on Type II diabetes. In contrast to this original stance, recent research has identified prediabetes as a distinct and meaningful stage of hyperglycemia. The goal of this paper was to investigate risk factors for prediction of both prediabetes and diabetes as separate health outcomes. Logistic regression modeling techniques identified similarity in the strongest risk factors for diabetes and prediabetes. These factors were BMI, stroke, poor physical health, heart attack, and high blood pressure. However, there were many more strong predictors for diabetes than there were for prediabetes, including many demographic attributes like marital status and sex. Ultimately, the strongest predictors for prediabetes and diabetes are similar, but there is increased strength for additional predictors when considering the diabetes class.

1 Introduction

Type 2 diabetes is a common chronic illness, affecting roughly 1 in 10 American adults. It is caused by a prolonged inability to maintain healthy blood sugar levels. Diabetes is associated with many adverse outcomes, including heart disease, blindness, and kidney disease [1]. While there is no known cure for Type 2 diabetes, it is possible for patients to experience reduction of symptom severity or even reversal of the disease through treatment [2]. Thus, early detection and regular screening of high-risk patients is imperative.

One of the most widely studied risk factors for type 2 diabetes is the presence of prediabetes. Considered an intermediate stage of hyperglycemia, prediabetes is diagnosed when blood sugar levels are higher than normal but are not high enough to be considered diabetes. 25% of patients with prediabetes will develop diabetes within 3-5 years, and up to 70% will develop diabetes over the span of their lifetime. In addition to an increased risk of diabetes, prediabetes is also associated with myriad adverse health outcomes, including increased risk of cardiovascular disease, peripheral neuropathy, and stroke. While a very common condition, affecting roughly 1 in 3 American adults, prediabetes often goes undiagnosed, with 90% unaware they have the disease [3, 4]. Widespread screening of high-risk individuals has been recommended for detection of diabetes [5, 6]. However, universal screening for prediabetes is currently not recommended [7]. This is in part due to inconsistent findings on potential risk factors, making it difficult to identify who to screen and how to justify the cost [8, 9]. However, emerging evidence suggests that prediabetes is a distinct disease with its own set of risk factors, diagnostic criteria and disease progression [10, 11, 12, 13]. Unfortunately, there is still a gap in understanding of risk factors and appropriate screening criteria for prediabetes as a distinct disease.

The goal of this project is to leverage logistic regression techniques to identify the strongest predictors of prediabetes and diabetes as separate outcomes and compare them to current screening standards. This will help identify whether screening approaches could be expanded to improve the likelihood of diagnosis and early detection in individuals who are at high risk for prediabetes.

2 Related Work

There are a set of widely accepted risk factors for diabetes. They are commonly grouped as behavioral, cardiometabolic, and non-modifiable attributes. The most significant behavioral risk factors include physical activity, diet, smoking behavior, and alcohol consumption. The most significant cardiometabolic risk factors include obesity, hypertension, and dyslipidemia. Finally, the most significant attribute risk factors include gender, race/ethnicity, age, and genetic factors. [16, 17, 18, 19] When reviewing the available literature, these risk factors are well researched and thoroughly proven to be associated with diabetes. However, there is much less clarity on prediabetes risk factors. The first issue is that many studies combine prediabetes and diabetes into a single outcome, making it difficult to identify the differences between prediabetes and diabetes risk factors [24]. The second issue is that there are differing definitions of prediabetes, which causes inconsistency in modeled risk factor strengths [20]. The third issue is that studies investigating prediabetes often consider prediabetes as a risk factor to progression of diabetes, rather than an outcome of interest [21, 22, 23].

However, if we consider studies that have investigated prediabetes risk factors specifically, we find there are some similar risk factors between diabetes and prediabetes. Examples of overlapping risk factors include gender, age, hypertension, dyslipidemia, obesity, and diet [10, 25, 29, 31, 32, 36]. However, there is not consistency between studies on which attributes are strongly associated with prediabetes. Some studies find that factors like gender, age, or diet are not strongly associated with prediabetes, and others find that factors like marital status, income, and education level are significantly associated [25, 27, 29, 31, 32, 34, 37]. This inconsistency is likely due to unbalanced class sizes, low incidence of prediabetes patients, inconsistency in the definition of prediabetes, and overconfidence in the similarity of the pathogenesis of prediabetes and diabetes [8, 11, 20, 28].

Generally, the variance in risk factors for prediabetes is related to a greater difficulty identifying statistically significant risk factors for prediabetes and a reduced strength of prediction for risk factors that overlap with diabetes [27, 30, 32].

For example, a study by Okwechime et al. (2015) found that being overweight, obese, hypertensive, hypercholesterolemic, and arthritic were significantly associated with prediabetes. However, they also found that these predictors were more strongly associated with diabetes. Additionally, many predictors that were not strongly associated with prediabetes were strongly associated with diabetes, including age, income level, and level of physical activity [27]. This inconsistency demonstrates the need for more thorough analysis of prediabetes risk factors and how they are distinct from or related to diabetes risk factors.

3 Data Sets

This project will leverage multiple data sets, comprised of results from the annual CDC Behavioral Risk Factor Surveillance System (BRFSS) survey. The annual survey has been conducted since 1984 and collects telephone survey responses from over 400,000 Americans on a range of health conditions, health behaviors, and use of health services [14]. The data sets used in this project are a subset of the publicly available records, comprised of features that are specifically related to diabetes, prediabetes, and related risk factors. Data for 2011, 2013, and 2015 were cleaned and aggregated. The dataset cleaning and feature selected is similar to previous work on the dataset, available via Kaggle. [15,41]. The datasets used in the modeling are also available via Kaggle. [42]

Data cleaning and feature transformation created an aggregated dataset of 528,794 responses and 17 features of interest. The attributes were chosen because they were questions asked during the BRFSS survey each year. Questions that were not asked consistently across survey years were excluded from the cleaned dataset. This was to increase the sample size for the prediabetes class, which is a very small proportion of the overall population.

4 Main Techniques Applied

First, data cleaning and preparation was conducted to create the aggregated dataset. Raw responses from the

2011, 2013, and 2015 surveys were extracted from Kaggle. These files were cleaned to remove missing and null values, only retaining rows that contained valid responses for all 17 attributes of interest. These were processed in batches due to the size of the raw datasets.

Once the raw datasets were cleaned, the features of interest were transformed according to the original question format from the BRFSS codebooks [44]. Additionally, features like number of poor mental health days and BMI were transformed to binary or categorical variables to account for the large number of outlier responses.

BMI was transformed to capture CDC classification of BMI: below 18.5 is underweight, 18.5-24.9 is healthy, 25 – 29.9 is overweight, 30 – 34.9 is class 1 obesity, 35 – 39.9 is class 2 obesity, and 40 and over is considered class 3 or severe obesity. [38]

Mental health was also transformed to better capture the distribution of the variable without introducing convergence issues. Because most respondents (69.25%) reported having no poor mental health days, it was important to capture no poor mental health days separately from any poor mental health days. This mirrors existing analysis on BRFSS responses on number of poor mental health days [39]. Responses were categorized into two values: no poor mental health days or some poor mental health days.

Physical health was transformed in a similar way to mental health. Most respondents (63.09%) reported having no poor physical health days, so responses were categorized into two values: no poor physical health days or some poor physical health days. This mirrors existing analysis on BRFSS responses on number of poor physical health days. [39]

Once the dataset cleaning and feature transformation was completed, feature selection was conducted to identify the strongest potential features. Chi-squared analyses were conducted for the selection of categorical variables, and one-way ANOVA tests, as well as f-tests and t-tests were conducted for continuous numerical predictors. Features with statistically insignificant correlation to the outcome of interest were dropped. In addition, collinearity analysis was conducted through correlation inspection and VIF score analysis. Collinear features were dropped from analysis.

Once the initial feature selection was completed, initial logistic regression models were built using methods from the sklearn and statsmodels libraries. Recursive feature elimination was conducted to reduce the number of dimensions, and oversampling and under-sampling methods like SMOTE were used to account for the imbalanced class sizes and boost performance.

Once the models were fitted, performance of the models was assessed using precision, recall, and accuracy scores, with a particular focus on the precision score given the imbalance in the size of the classes. Additionally, ROC/AUC scores were assessed to understand model performance, with a particular focus on the micro-average AUC score to account for the class size imbalance.

The project was completed in a series of interactive Jupyter notebook file, to improve readability and understanding of the modeling process. Visualizations and written explanations are displayed next to the model generation and statistical analysis code blocks.

5 Key Results

Multinomial Logistic Regression Model

The first model created was a multinomial logistic regression model, to predict the diabetes and prediabetes classes relative to the no diabetes class as the reference group. Once this model was fitted and created, logistic regression models were fitted for diabetes and prediabetes as distinct outcomes, to assess the consistency and accuracy of the multinomial logistic regression model.

Feature Selection

The first dataset explored was the three class 2015 BRFSS dataset, with responses labeled as prediabetes vs. diabetes vs. no diabetes. Initial feature selection focused on identifying strong predictors and reducing collinearity between selected variables.

The three numerical variables were skewed, with many outliers. Because outliers can cause issues with convergence, the variables were transformed using

domain knowledge into discrete categorical predictors. BMI, physical health, and mental health were transformed into discrete categorical variables. Chi-square test results indicated that the transformed variables BMI X^2 (2, N = 253,680) = 256.13, $p < 0.0$, physical health X^2 (2, N = 253,680) = 4671.23, $p = 0.0$, and mental health X^2 (2, N = 253,680) = 256.13, $p < 0.0$ had a high correlation to our outcome of interest and were selected as part of the initial feature set.

The remainder of the variables were categorical attributes, so chi-square tests were conducted on each predictor. All predictors were statistically significantly related to the difference in means between our classes, suggesting that they are strong predictors and should be selected as initial features.

Once the initial features were selected, it was necessary to investigate and reduce collinearity between features. Multinomial logistic regression requires the assumption that features are independent, which means that any strongly collinear features must be dropped. A correlation heatmap using the Pearson correlation coefficient was generated to identify collinear features. The strongest correlations were between mental health and physical health (r = -0.48), education and income (r = 0.45), and income and marital status (r = 0.44). However, a correlation coefficient of 0.8 or greater is usually used to identify collinear features. [40] Therefore, no additional predictors were dropped from the analysis.

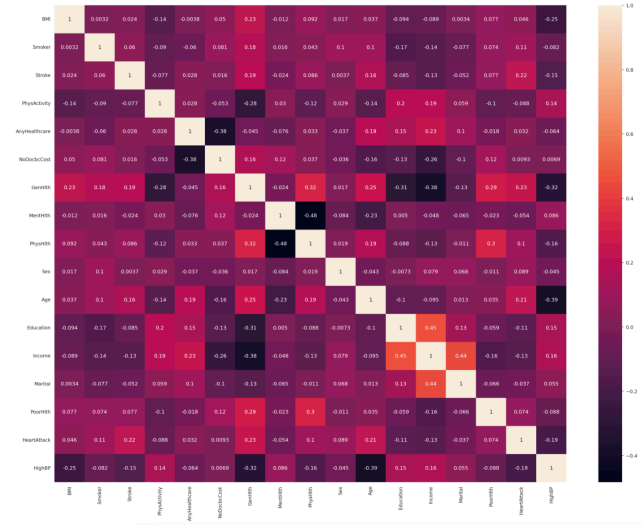

Figure 1: Pearson Correlation Coefficient Heatmap for Three Class Dataset

To further identify collinear features, a Variance Inflation Factor (VIF) calculation was used. A VIF score of greater than 5 indicates moderate to strong collinearity between variables. [40] Variable selection using VIF score was done recursively, dropping the variable with the highest VIF score and recalculating until all variables had a VIF score of less than 5. Education (*VIF* = 21.66), General Health (*VIF* = 10.39), General Health (*VIF* = 9.94), Income (*VIF* = 7.26), and Age (*VIF* = 5.79) had high VIF scores and were dropped from analysis. The remaining attributes had VIF scores less than 5 and were retained in the analysis.

features	vif_factor
BMI	4.188308
PhysHlth	3.844187
PhysActivity	2.786064
MentHlth	2.590753
PoorHlth	2.050418
HighBP	1.986534
Smoker	1.951992
Marital	1.928597
Sex	1.638646
NoDocbcCost	1.261773
HeartAttack	1.186699
Stroke	1.131868

Figure 2: Final VIF Score Output for Initial Selected Features for Three Class Dataset

The feature selection process identified 12 attributes out of the original 17 features for initial modeling and analysis. There were 3 medical behavior attributes: Physical Activity, No Doctor Because of Cost, Smoker. There were 6 health outcome attributes: BMI, Mental Health, Physical Health, Poor Health, Heart Attack, High Blood Pressure. Finally, there were 2 demographic attributes: Sex, Marital Status.

Given the large number of attributes, the goal was to first identify the strongest predictors to reduce dimensionality, then tune to produce a well-fitted model.

Initial Multinomial Logistic Regression Model

To begin identifying the strongest predictors, a simple multinomial logistic regression model with all available features was generated using the statsmodels library. The initial model demonstrated a relatively poor fit to the data (pseudo-$R^2 = 0.15$) and had room to be further improved.

MNLogit Regression Results

Dep. Variable:	Diabetes_012	No. Observations:	396595
Model:	MNLogit	Df Residuals:	396569
Method:	MLE	Df Model:	24
Date:	Thu, 10 Aug 2023	Pseudo R-squ.:	0.1479
Time:	01:48:08	Log-Likelihood:	-1.7214e+05
converged:	True	LL-Null:	-2.0203e+05
Covariance Type:	nonrobust	LLR p-value:	0.000

Figure 3 Statsmodels Multinomial Logistic Regression Model – Descriptive Statistics

For the prediabetes class, the Sex attribute had a statistically insignificant coefficient value ($\beta = -0.01$, $p = 0.59$). All other predictors, including the constant, had a statistically significant coefficient value, and thus were retained in the analysis. The Sex attribute was dropped from further analysis.

Diabetes_012=1	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.4531	0.055	-81.549	0.000	-4.560	-4.346
BMI	0.3106	0.010	32.554	0.000	0.292	0.329
Smoker	0.1069	0.024	4.432	0.000	0.060	0.154
Stroke	0.3516	0.046	7.720	0.000	0.262	0.441
PhysActivity	-0.1243	0.025	-4.931	0.000	-0.174	-0.075
NoDocbcCost	0.1642	0.030	5.523	0.000	0.106	0.222
MentHlth	0.1469	0.028	5.327	0.000	0.093	0.201
PhysHlth	0.2801	0.033	8.384	0.000	0.215	0.346
Sex	-0.0136	0.025	-0.550	0.582	-0.062	0.035
Marital	-0.1578	0.024	-6.551	0.000	-0.205	-0.111
PoorHlth	0.1267	0.025	5.015	0.000	0.077	0.176
HeartAttack	0.3578	0.041	8.803	0.000	0.278	0.437
HighBP	-0.8081	0.025	-31.960	0.000	-0.858	-0.759
Diabetes_012=2	**coef**	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-2.7106	0.023	-117.339	0.000	-2.756	-2.665
BMI	0.4412	0.004	110.502	0.000	0.433	0.449
Smoker	0.0778	0.010	7.757	0.000	0.058	0.097
Stroke	0.5279	0.018	29.803	0.000	0.493	0.563
PhysActivity	-0.2790	0.010	-27.281	0.000	-0.299	-0.259
NoDocbcCost	-0.1024	0.013	-7.665	0.000	-0.129	-0.076
MentHlth	-0.0242	0.011	-2.186	0.029	-0.046	-0.002
PhysHlth	0.5854	0.015	38.600	0.000	0.556	0.615
Sex	0.0584	0.010	5.692	0.000	0.038	0.078
Marital	-0.1164	0.010	-11.660	0.000	-0.136	-0.097
PoorHlth	0.1277	0.010	12.323	0.000	0.107	0.148
HeartAttack	0.7865	0.015	52.280	0.000	0.757	0.816
HighBP	-1.3476	0.011	-122.255	0.000	-1.369	-1.326

Figure 4 Statsmodels Multinomial Logistic Regression Model – Feature Coefficients

Using this set of features, a multinomial logistic regression model was fitted using the sklearn multinomial logistic regression function. This was done to leverage some additional parameters to better account for the imbalanced class sizes in our dataset. A newton-cg solver was used, with additional weighting on rare classes through the balanced-classes parameter. The multinomial logistic regression model was fitted using the 11 features identified as having statistically significant coefficient values in the statsmodels multinomial logistic regression model.

The performance of this model was improved over the first iteration, but still saw mediocre performance. There was a training and accuracy score of 63% but had a precision score of 83%. The higher precision score is good, given our focus on accurately predicting in our minority classes, prediabetes and diabetes.

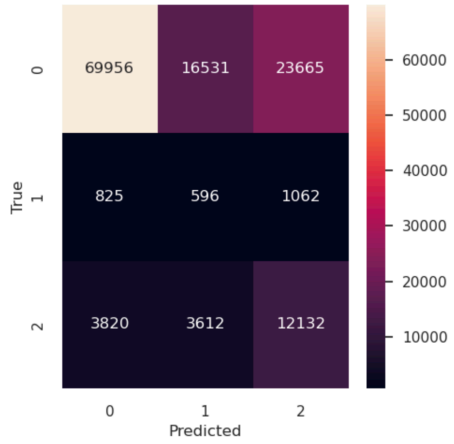

Figure 5 – Confusion Matrix for Multinomial Logistic Regression Model (All Features)

When investigating the ROC scores for each of our classes, we saw good performance for the diabetes (AUC = 0.78) and healthy (AUC = 0.79) classes, and worse performance for the prediabetes class (AUC = 0.64). This is likely due to the very small sample size of the prediabetes class, and the high dimensionality of the dataset.

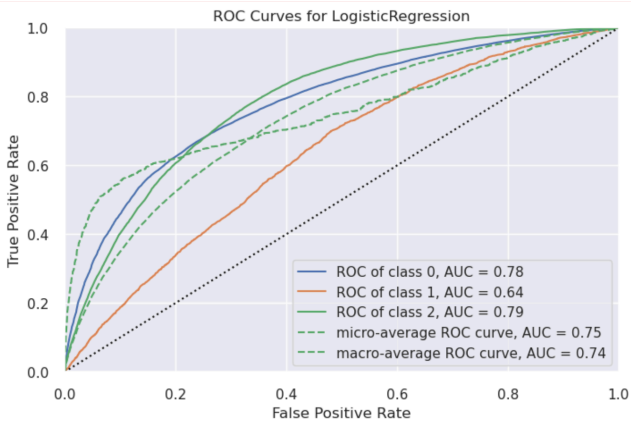

Figure 6 – ROC/AUC Curves for Multinomial Logistic Regression Model (Initial Feature Set)

Recursive Feature Elimination

In order to reduce the dimensionality, recursive feature elimination (RFE) was used to select the top 5 strongest predictors of the prediabetes and diabetes classes. The RFE algorithm selected BMI, Stroke, Physical Health, Heart Attack, and High Blood Pressure as the best predictors. A multinomial logistic regression model was fitted using the sklearn library.

The training and test accuracy was slightly improved to 65%, and the recall score slightly improved to 65%, with a precision score of 83%.

However, the ROC curves and AUC scores indicated that there was not significant improvement in the model fit. There was no improvement in the prediction of the healthy class (AUC = 0.78) or the diabetes class (AUC= 0.79), and a slight decrease in performance on the prediabetes class (AUC = 0.62).

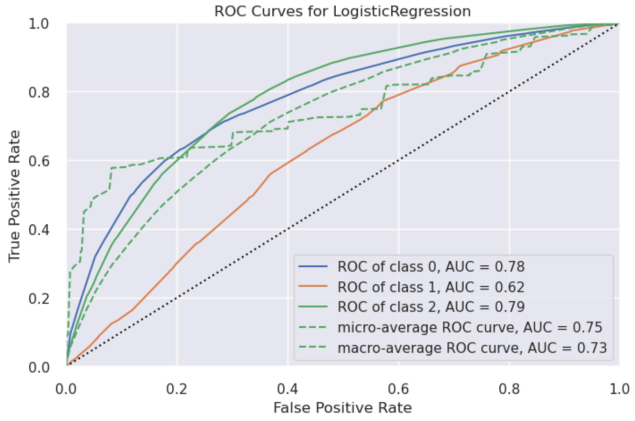

Figure 7 ROC/AUC Curves for Multinomial Logistic Regression Model (RFE Feature Set)

SMOTE

Because SMOTE and other under-sampling and oversampling techniques work best with low dimensionality datasets, we had to eliminate potentially redundant features prior to attempting to further account for the imbalanced class sizes. The use of SMOTE oversampling technique and the random under sampler method from the imblearn library created a much stronger model. There was a significant improvement in the accuracy, with a training accuracy score of 80%, test accuracy of 84%, and recall score of 84%. There was a small decrease in precision, with a precision score of 79%.

However, the performance of the model was much improved with the over-sampling and under-sampling techniques applied.

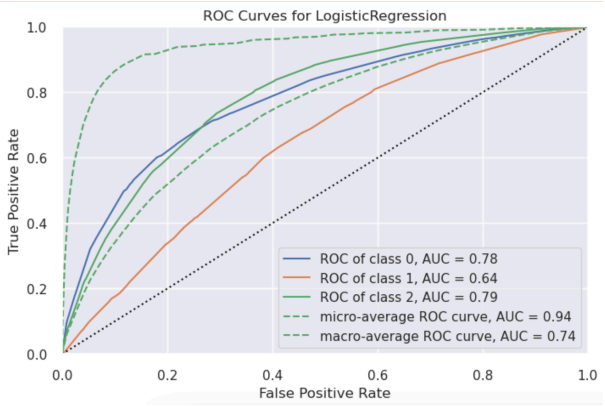

Figure 8 ROC/AUC Curves for Multinomial Logistic Regression Model (SMOTE/Random Under Sampler feature set)

The ROC score was improved for the prediabetes class (ROC=0.64), albeit still low. However, the micro-average AUC score was significantly improved (AUC=0.94). The micro-average AUC score is preferred when there is an imbalance between class sizes, which is true for our dataset.

Interpretation of the Model

The best performing model was fitted using BMI, Stroke, Physical Health, Heart Attack, and High Blood Pressure using the SMOTE and Random Under Sampler transformed training dataset fitted using features selected using recursive feature elimination (RFE).

For the diabetes class, BMI ($\beta = 0.20$, $p = 0.000$), Stroke ($\beta = 0.28$, $p = 0.000$), poor Physical Health ($\beta = 0.37$, $p = 0.000$), Heart Attack ($\beta = 0.45$, $p = 0.000$), and High Blood Pressure ($\beta = 0.66$, $p = 0.000$) were associated with an increase in risk of diabetes. In other words, there is a 22% increase in the risk of diabetes with each increase in BMI category, 32% increase in the risk of diabetes with history of Stroke, 45% increase in the risk of diabetes with poor physical health, 57% increase in the risk of diabetes with history of heart attack, and 93% increase in the risk of diabetes with high blood pressure.

For the prediabetes class, BMI ($\beta = 0.07$, $p = 0.000$), Stroke ($\beta = 0.08$, $p=0.000$), and High Blood Pressure

(β = 0.09,p=0.000) were associated with an increase in the risk of prediabetes. In other worse, there is a 7% increase in the risk of prediabetes with each increase in BMI category, 8% increase in the risk of prediabetes with history of stroke, 9% increase in the risk of prediabetes with high blood pressure. However, unlike diabetes, poor physical health (β = -0.04, p = 0.000) was associated with a 4% reduction in the risk of prediabetes, and history of heart attack (β = -0.01, p = 0.000) was associated with a 1% reduction in the risk of prediabetes.

This tells us that while BMI, stroke, and high blood pressure are associated with an increased risk of prediabetes and diabetes, poor physical health and history of heart attack were not consistently strong predictors of increased risk for both prediabetes and diabetes.

To test these outcomes, logistic regression models were fitted to predict diabetes vs. healthy, and prediabetes vs. healthy, to determine whether the best fit models would contain the same predictors for both classes separately.

Prediabetes Logistic Regression Model

The same feature selection process was conducted for the prediabetes logistic regression model as was performed for the multinomial logistic regression model. Chi-squared analyses indicated that all initial features were strongly correlated with our outcome. Additionally, VIF score analysis indicated that Education (*VIF* = 21.72), General Health (*VIF* = 10.33), Any Healthcare (*VIF* = 9.94), Income (*VIF* = 7.29), and Age (*VIF* = 5.76) had high collinearity and were dropped from the analysis.

The best fit logistic regression model for the prediction of the prediabetes class vs. healthy class was created using the sklearn logistic regression model, with parameter balanced class weight and the imblearn SMOTE and random under sampler techniques for oversampling and under-sampling.

The training accuracy was 91%, the test accuracy was 98%, the recall was 98%, and the precision was 96%. This is a very good fit to our dataset and a strong model.

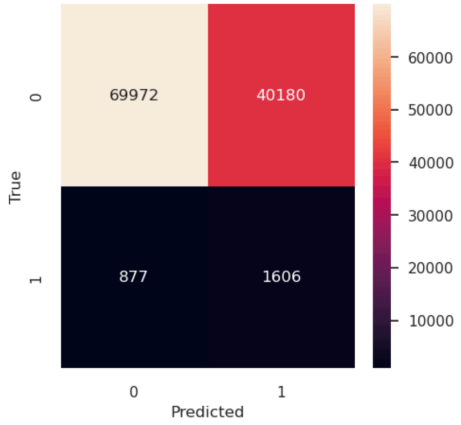

Figure 9 Confusion Matrix for Prediabetes Logistic Regression Model

This performance can also be seen in the micro-average AUC score of 0.99. While the AUC score for each class is lower, at 0.69, the strength of the micro-average AUC score is preferred in model performance assessment due to the imbalanced class sizes.

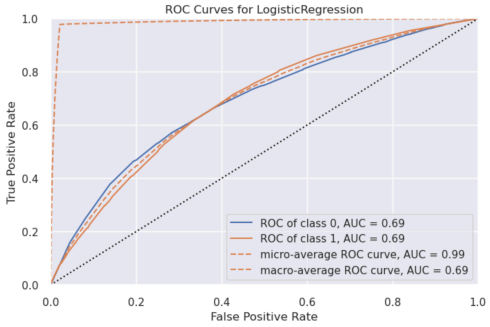

Figure 10 ROC/AUC Curves for Prediabetes Logistic Regression Model

The coefficients of the model were BMI (β = 0.33, p = 0.000) , Stroke (β = 0.38, p = 0.000), poor Physical Health (β = 0.25, p = 0.000), Heart Attack (β = 0.39, p = 0.000), and High Blood Pressure (β = 0.84, p = 0.000). In other words, for every increase in BMI category, there is a 39% increase in the risk of prediabetes. History of stroke increases risk of prediabetes by 46%, history of heart attack increases risk of prediabetes by 48%, poor health increases risk of prediabetes by 28%, and high blood pressure increases risk of prediabetes by 132%.

Diabetes Logistic Regression Model

The same feature selection process was conducted for the diabetes logistic regression model as was

performed for the multinomial logistic regression model. Chi-squared analyses indicated that Smoker (p = 0.40) and Marital Status (p = 0.11) were not significantly associated with our outcome and were dropped from analysis. Additionally, VIF score analysis indicated that Education (*VIF* = 19.56), General Health (*VIF* = 15.70), Any Healthcare (*VIF* = 13.12), Age (*VIF* = 8.71), and Physical Health (*VIF* =

The best fit logistic regression model for the prediction of diabetes class vs. healthy class was created using the sklearn logistic regression model method, with parameter balanced class weight and SMOTE techniques for oversampling and under-sampling.

Unlike with the prediabetes logistic regression model, a model for diabetes prediction was able to be fitted using many of the original features. The included features were BMI, smoker, stroke, physical activity, no doctor because of cost, mental health, physical health, sex, marital status, poor health, heart attack, and high blood pressure.

Logit Regression Results

Dep. Variable:	Diabetes_012	No. Observations:	389148
Model:	Logit	Df Residuals:	389135
Method:	MLE	Df Model:	12
Date:	Thu, 10 Aug 2023	Pseudo R-squ.:	0.1747
Time:	21:52:01	Log-Likelihood:	-1.3621e+05
converged:	True	LL-Null:	-1.6505e+05
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.0513	0.023	-173.755	0.000	-4.097	-4.006
BMI	0.4425	0.004	110.574	0.000	0.435	0.450
Smoker	0.0776	0.010	7.726	0.000	0.058	0.097
Stroke	0.5240	0.018	29.553	0.000	0.489	0.559
PhysActivity	-0.2786	0.010	-27.161	0.000	-0.299	-0.259
NoDocbcCost	-0.1111	0.013	-8.288	0.000	-0.137	-0.085
MentHlth	-0.0274	0.011	-2.469	0.014	-0.049	-0.006
PhysHlth	0.5738	0.015	37.838	0.000	0.544	0.604
Sex	0.0646	0.010	6.292	0.000	0.044	0.085
Marital	-0.1072	0.010	-10.714	0.000	-0.127	-0.088
PoorHlth	0.1187	0.010	11.418	0.000	0.098	0.139
HeartAttack	0.7937	0.015	52.845	0.000	0.764	0.823
HighBP	1.3433	0.011	121.797	0.000	1.322	1.365

Figure 11 Statsmodels Diabetes Logistic Regression Model (Full Feature Set)

Even with this high dimensionality, the model achieved 71% training and test accuracy, with 71% recall score and 84% precision score.

The performance of the model was similar for both the no diabetes and diabetes classes, with an AUC score of 0.79 for both classes, and a micro-average AUC score of 0.79.

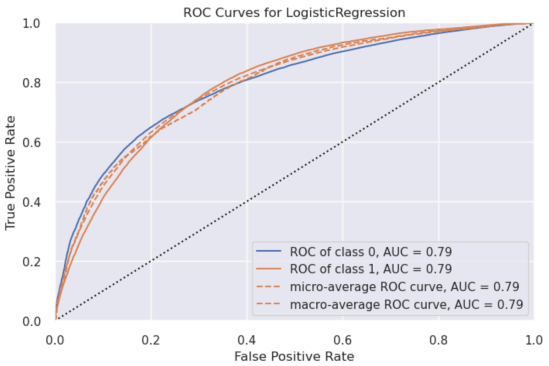

Figure 12 ROC/AUC Curves for Diabetes Logistic Regression Model (Full Feature Set)

Unlike the prediabetes class, the model performance was not significantly improved when reducing the number of features or performing under-sampling and over-sampling techniques to better balance the class sizes.

A model fitted with the top 5 strongest predictors selected using RFE had very similar performance to the initial logistic regression model. The training and test accuracy was 72%, the recall score was 72%, and the precision was 84%.

There was no improvement in the AUC scores for either the diabetes or no diabetes classes, with only a small improvement in the micro-average AUC score, with a value of 80%.

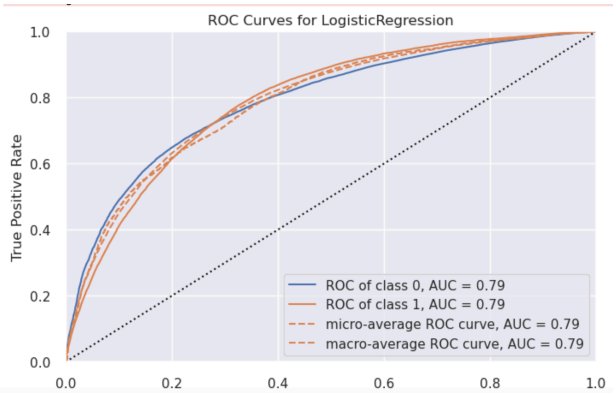

Figure 13 ROC/AUC Curves for Diabetes Logistic Regression Model

The coefficients were BMI (β = 0.47, p = 0.000) Stroke (β = 0.62, p = 0.000), Physical Health (β = 0.70, p = 0.000), Heart Attack (β = 0.93, p = 0.000), and High Blood Pressure (β = 1.41, p = 0.000).

In other words, for every increase in BMI category, the risk of diabetes increased 60%. Diabetes risk increases

86% with history of stroke, 101% with poor physical health, 153% with history of heart attack, and 310% with high blood pressure.

This tells us that there are more predictors that are strongly associated with diabetes outcome than the prediabetes class.

6 Applications

Overall, the predictors consistently identified as the strongest risk factors for diabetes and prediabetes were BMI, Stroke, Poor Physical Health, Heart Attack, and High Blood Pressure. The risk factor associated with the greatest increase in risk of both diabetes and prediabetes is high blood pressure.

When comparing these risk factors to the established risk factors, there is close alignment to existing guidelines and research [10, 25, 29, 31, 32, 36]. However, the directionality of these relationships is not accounted for in the model, and it is important to note that the identification of these risk factors is not for the risk of developing the disease, but rather the probability of experiencing that outcome. For example, there is a well-established relationship between stroke and diabetes. However, it is noted that stroke is usually considered to be an outcome of diabetes. [43] This illustrates a limitation of the modeling performed: we cannot assume the directionality of these results, only that there is an increased risk of the disease given the selected attributes. This limitation should be noted when interpreting these results.

An additional challenge with this research is the high dimensionality of the dataset. When fitting the models, there were many features significantly associated with diabetes, while there were not as many significantly associated with prediabetes. While there is alignment between models on the strongest predictors of diabetes and prediabetes, it is possible that noisiness of the dataset of poor encoding of the survey results is responsible for the relative strength of some predictors. Finally, a challenge with these models is the large imbalance in class sizes, particularly with the prediabetes class. Because the prediabetes class size is so small, there are limits on the statistical power of the model fitted to the data, and the results must be interpreted carefully.

Despite these limitations, these findings mirror the results found by Okwechime et al (2015) and other research conducted that compares the predictors of prediabetes with diabetes [27, 30, 32]. These results also align with existing research and screening guidelines, which suggest that there are similar risk factors for prediabetes and diabetes. This aligns with the assumption that there is a similar pathogenesis of prediabetes and diabetes.

However, the most interesting findings of this research is the findings of the logistic regression model fitted on the diabetes and healthy dataset, relative to the prediction of the prediabetes class. There were many more features in the cleaned dataset that were strong predictors of diabetes. The strongest predictors of prediabetes tended to be the medical attributes, with demographic attributes dropped due to covariance and collinearity early on. This suggests that there may not be a clear distinction in variance and distribution of demographic attributes for the prediabetes and no diabetes classes, while there is a significant difference between the diabetes and no diabetes classes. What would be interesting to explore further is the possibility that there are important risk factors for the progression of prediabetes to diabetes that may not be the same risk factors for prediabetes, and that these may include more of the demographic attributes that are significantly associated with diabetes. Because not all those with prediabetes will necessarily develop diabetes later, this difference in risk factor may be related to a separate underlying mechanism that is related to the progression from prediabetes to diabetes. There is evidence to suggest that there are separate risk factors that predict the progression from prediabetes to diabetes, and is worth additional investigation [21,22,23].

To answer our original research questions, we can conclude based on our model findings that there is not a meaningful difference between the strongest risk factors of prediabetes and diabetes. However, there are many more attributes that are strong risk factors for diabetes than prediabetes, which can be attributed to limitations of the dataset or potential underlying mechanisms that require further research and investigation.

REFERENCES

[1] CDC. 2023. Type 2 Diabetes. (April 2023). Retrieved July 10, 2023 from https://www.cdc.gov/diabetes/basics/type2.html

[2] Sarah J Hallberg, Victoria M Gershuni, Tamara L Hazbun, and Shaminie J Athinarayanan. 2019. Reversing Type 2 Diabetes: A Narrative Review of the Evidence. *Nutrients* 11, 4, (April 2019), 766. DOI: https://doi.org/10.3390/nu11040766

[3] Adam G. Tabák, Christian Herder, Wolfgang Rathmann, Eric J. Brunner, and Mika Kivimäki. Prediabetes: a high-risk state for diabetes development. *Lancet* 379, 9833, (June 2012), 2279-2290. DOI: https://doi.org/10.1016/S0140-6736(12)60283-9

perspectives. *Clinical Diabetes and Endocrinology* 5, 5, (May 2019). DOI: https://doi.org/10.1186/s40842-019-0080-0

[5] US Preventive Services Task Force. Screening for Prediabetes and Type 2 Diabetes: US Preventive Services Task Force Recommendation Statement. *JAMA*. 326, 8, (August 2021), 736–743. DOI: https://10.1001/jama.2021.12531

[6] Kenneth Lam and Sei J. Lee. Prediabetes – A Risk Factor Twice Removed. *JAMA Internal Medicine* 181, 4, (August 2021), 520-521. DOI: https://doi.org/10.1001/jamainternmed.2020.8773

[7] Aditya K. Khetan and Sanjay Rajagopalan. Prediabetes. *Canadian Journal of Cardiology* 34, 5, (May 2018), 615-623. DOI: https://doi.org/10.1016/j.cjca.2017.12.030

[8] Ashkann Zand, Karim Ibrahim and Bhargavi Patham. Prediabetes: Why Should We Care? Methodist Debakey Cardiovascular Journal,14, 4, (October 2018), 289-297. DOI: https://doi.org/10.14797/mdcj-14-4-289

[9] Justin B. Echouffo-Tcheugui and Elizabeth Selvin. Prediabetes and What It Means: The Epidemiological Evidence. *Annual Review of Public Health* 42, 59, (April 2021), 59-77. DOI: https://doi.org/10.1146/annurev-publhealth-090419-102644

[10] Alicia Diaz-Redondo, Carolina Giráldez-García, Lourdes Carrillo et al. Modifiable risk factors associated with prediabetes in men and women: a cross-sectional analysis of the cohort study in primary health care on the evolution of patients with prediabetes (PREDAPS-Study). *BMC Family Practice*, 16, Article 5, (January 15). DOI: https://doi.org/10.1186/s12875-014-0216-3

[11] J.F. Elgart, R. Torrieri, M. Ré et al. Prediabetes is more than a pre-disease: additional evidences supporting the importance of its early diagnosis and appropriate treatment. *Endocrine* 79, (January 2023), 80-85. DOI: https://doi.org/10.1007/s12020-022-03249-8

[12] Jan Brož, Jana Malinovska, Marisa A. Nunes et al. Prevalence of diabetes and prediabetes and its risk factors in adults aged 25-64 in the Czech Republic: A cross-sectional study. *Diabetes Research and Clinical Practice*, 170, Article 108470, (September 2020). DOI: https://doi.org/10.1016/j.diabres.2020.108470

[13] Khaled K. Aldossari, Abdulrahman Alidiab, Jamaan M. Al-Zahrani et al. Prevalence of Prediabetes, Diabetes, and Its Associated Risk Factors among Males in Saudi Arabia: A Population-Based Survey. *Journal of Diabetes Research* 2018, Article 2194604 (April 2018), 12 pages. DOI: https://doi.org/10.1155/2018/2194604

[14] CDC. Behavioral Risk Factor Surveillance System. (May 2023). Retrieved July 10, 2023 from https://www.cdc.gov/brfss/index.html

[15] Alex Teboul. Diabetes Health Indicators Dataset. (2021). Retrieved July 10, 2023 from https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

[16] Sania Siddiqui, Hadzliana Zainal, Sabariah Noor Harun, Siti Maisharah Sheikh Ghadzi, and Saadia Ghafoor. Gender differences in the modifiable risk factors associated with the presence of prediabetes: A systematic review. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 14, 5, (July 2020), 1243-1252. DOI: https://doi.org/10.1016/j.dsx.2020.06.069

[17] CDC. Diabetes Risk Factors. (April 2022). Retrieved July 10, 2023 from https://www.cdc.gov/diabetes/basics/risk-factors.html

[18] Barbara Fletcher, Meg Gulanick, and Cindy Lamendola. Risk Factors for Type 2 Diabetes Mellitus. *The Journal of Cardiovascular Nursing*, 16, 2, (January 2002), 17-23.

[19] Phillipa J Talmud, Aroon D. Hingorani, Jackie A Cooper et al. Utility of genetic and non-genetic risk factors in prediction of type 2 diabetes: Whitehall II prospective cohort study. *BMJ* 340, Article 4838, (January 2010). DOI: https://doi.org/10.1136/bmj.b4838

[20] Crystal Man Ying Lee, Stephen Colagiuri, Mark Woodward et al. Comparing different definitions of prediabetes with subsequent risk of diabetes: an individual participant data meta-analysis involving 76,513 individuals and 8,208 cases of incident diabetes. *BMJ Open Diabetes Research and Care,* 7, 1, (November 2019). DOI: https://doi.org/10.1136/ bmjdrc-2019-000794

[21] Bennasar-Veny M, Fresneda S, López-González A, Busquets-Cortés C, Aguiló A, Yañez AM. Lifestyle and Progression to Type 2 Diabetes in a Cohort of Workers with Prediabetes. *Nutrients,* 12, 5, Article 1538 (May 2020,. DOI: https://doi.org/10.3390/nu12051538

[22] Rooney MR, Rawlings AM, Pankow JS, et al. Risk of Progression to Diabetes Among Older Adults With Prediabetes. *JAMA Intern Med*., 181, 4, (February 2021), 511-519. DOI: https://doi.org/10.1001/jamainternmed.2020.8774

[23] N. Anthony, V. Lenclume, A. Fianu, N.Le Moullec, X. Debussche, P. Gérardin, C. Marimoutou, E. Nobécourt, Association between prediabetes definition and progression to diabetes: The REDIA follow-up study. *Diabetes Epidemiology and Management,* 3, Article 100024 (November 2021). DOI: https://doi.org/10.1001/jamainternmed.2020.8774

[24] Hai Wang, Xin Zheng, Zheng-Hai Bai, Jun-Hua Lv, Jiang-Li Sun, Yu Shi, and Hong-Hong Pei. A retrospective population study to develop a predictive model of prediabetes and incident type 2 diabetes mellitus from a hospital database in Japan between 2004 and 2015. *Medical Science Monitor*, 26, (April 2020). DOI: https://doi.org/10.12659/MSM.920880

[25] Jiahua Wu, Jiaqiang Zhou, Xueyao Yin, Yixin Chen, Xihua Lin, Zhiye Xu, and Hong Li. A Prediction Model for Prediabetes Risk in Middle-Aged and Elderly Populations: A Prospective Cohort Study in China. *International Journal of Endocrinology*, 2021, Article 2520806 (November 2021). DOI: https://doi.org/10.1155/2021/2520806.

[26] Xue-Hui Meng, Yi-Xiang Huang, Dong-Ping Rao, Qiu Zhang, and Qing Liu. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors, *The Kaohsiung Journal of Medical Sciences*, 29, 2 (October 2012), 93-99. DOI: https://doi.org/10.1016/j.kjms.2012.08.016

[27] Ifechukwude Obiamaka Okwechime , Shamarial Roberson, and Agricola Odoi. Prevalence and Predictors of Pre-Diabetes and Diabetes among Adults 18 Years or Older in Florida: A Multinomial Logistic Modeling Approach. *PLOS ONE,* 10, 12, Article e0145781 (2015). DOI: https://doi.org/10.1371/journal.pone.0145781

[28] Gary S. Collins, Susan Mallett, Omar Omar, and Ly-Mee Yu. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med*icine 9, 1, (December 2011), 1-14.

[29] Mayo Clinic. 2022. Prediabetes. (November 2022). Retrieved July 10, 2023 from https://www.mayoclinic.org/diseases-conditions/prediabetes/symptoms-causes/syc-20355278

[30] Vijay Viswanathan, Satyavani Kumpatla, Vigneswari Aravindalochanan et al. Prevalence of Diabetes and Pre-Diabetes and Associated Risk Factors among Tuberculosis Patients in India. *PLOS ONE,* 7, 7, (2012), Article ID e41367. DOI: https://doi.org/10.1371/journal.pone.0041367

[31] Zhao, Ming, Hongbo Lin, Yanyan Yuan, Fuyan Wang, Yang Xi, Li Ming Wen, Peng Shen, and Shizhong Bu. Prevalence of Pre-Diabetes and Its Associated Risk Factors in Rural Areas of Ningbo, China. *International Journal of Environmental Research and Public Health* 13, 8, Article 808 (August 2016). DOI: https://doi.org/10.3390/ijerph13080808

[32] Hemavathi Dasappa, Farah Naaz Fathima, Rugmani Prabhakar, Sanjay Sarin. Prevalence of diabetes and pre-diabetes and assessments of their risk factors in urban slums of Bangalore. *Journal of Family Medicine and Primary Care*, 4, 3 (July 2015), 399-404. DOI: https://doi.org/10.4103/2249-4863.161336

[33] Xie Z, Nikolayeva O, Luo J, Li D. Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. *Prev Chronic Dis* 16, Article 130, (September 2019). DOI: https://doi.org/10.5888/pcd16.190109

[34] F Hadaegh, A Derakhshan, N Zafari, D Khalili, M Mirbolouk,N Saadat, and F Azizi. Pre-diabetes tsunami: incidence rates and risk factors of pre-diabetes and its different phenotypes over 9 years of follow-up. *Diabet Med*. 34, 1, (January 2017), 69-78. DOI: https://doi.org/10.1111/dme.13034

[35] Rui Wang, Peng Zhang, Zhijun Li et al. The prevalence of pre-diabetes and diabetes and their associated factors in Northeast China: a cross-sectional study. *Scientific Reports*, 9, Article 2513, (February 2019). https://doi.org/10.1038/s41598-019-39221-2

[36] Parisa Amiri, Sara Jalali-Farahani, Mehrdad Karimi et al. Factors associated with pre-diabetes in Tehranian men and women: A structural equations modeling. *PLOS ONE*, 12, 12, Article e0188898, (December 2017). https://doi.org/10.1371/journal.pone.0188898

[37] Anna Zamora-Kapoor, Amber Fyfe-Johnson, Adam Omidpanah, Dedra Buchwald, and Ka'imi Sinclair. Risk factors for pre-diabetes and diabetes in adolescence and their variability by race and ethnicity. *Preventive Medicine*, 115, (August 2018), 47-52. DOI: https://doi.org/10.1016/j.ypmed.2018.08.015

[38] CDC. Defining Adult Overweight & Obesity. (2022). Retrieved July 23, 2023 from https://www.cdc.gov/obesity/basics/adult-defining.html

[39] Lindsey Lanigan and Colin Planalp. Measuring State-level Disparities in Unhealthy Days. (2022). Retrieved July 23, 2023 from https://www.shadac.org/news/measuring-unhealthy-days-SHC

[40] Noora Shrestha. Detecting Multicollinearity in Regression Analysis. *American Journal of Applied Mathematics and* Statistics, 8, 2, (June 2020), 39-42. DOI: https://doi.org/10.12691/ajams-8-2-1

[41] Alex Teboul. Diabetes Health Indicators Dataset Notebook. (2022). Retrieved July 10, 2023 from https://www.kaggle.com/code/alexteboul/diabetes-health-indicators-dataset-notebook/notebook

[42] CDC. Behavioral Risk Factor Surveillance System. (2017). Retrieved July 10, 2023 from https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system?select=2011.csv

[43] Rong Chen, Bruce Ovbiagele, and Wuwei Feng. Diabetes and Stroke: Epidemiology, Pathophysiology, Pharmaceuticals, and Outcomes. *American Journal of Medical Sciences*, 351,4 (April 2016),380-386. DOI: https://doi.org/10.1016%2Fj.amjms.2016.01.01

[44] CDC. Behavioral Risk Factor Surveillance System: 2015 Codebook Report. (2016). Retrieved August 10, 2023 from https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf