Final Report

Predicting Type 2 Diabetes (2019-2022)

Thomas Dinh, Scott Bennett, Mark Hoyt

***Each author contributed equally to the design, coding & development, analysis, and writing of this project**

2 May 2024

Abstract

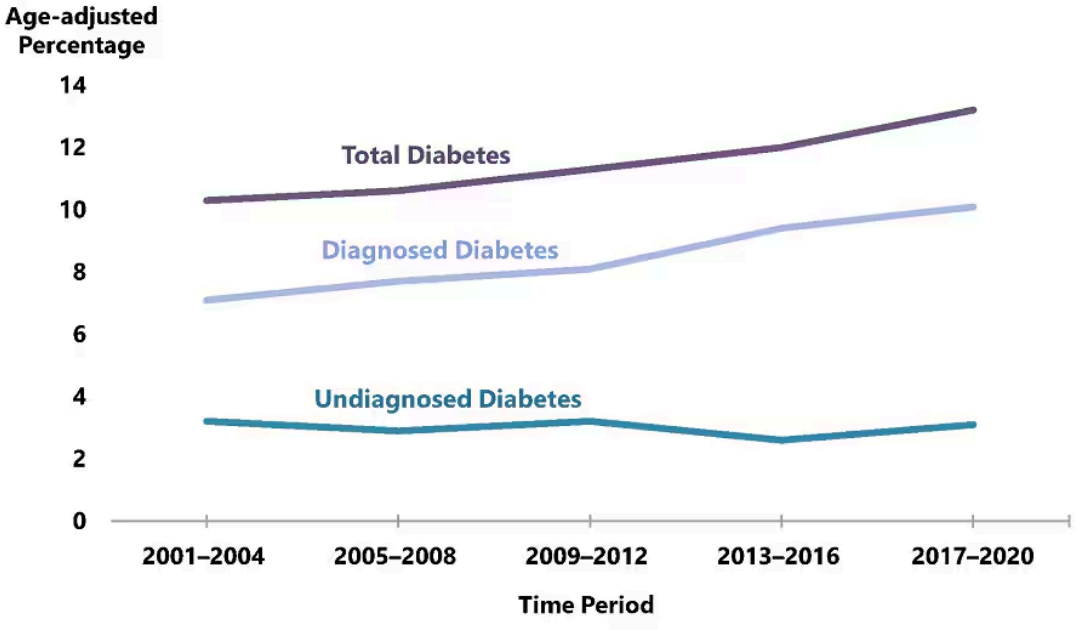
The purpose of this project was to determine if there was a method to predict an individual's risk of having type 2 diabetes based on information available to them. Through this, the intent was to be able to develop a model that could ask a person or their physician some simple questions about their health background, their socioeconomic status, and their lifestyle and determine if that individual was at high risk of having type 2 diabetes. The motivation for this project was due to the high rates of individuals living with undiagnosed diabetes. Through an appropriate tool, they would be able to know if they should go see their doctor. In order to accomplish this project, data was collected through the Center for Disease Control's (CDC) National Health Interview Survey (NHIS). Following the data collection, initial exploratory data analysis was conducted using the data and then predictive models were generated with the data.

Multiple models were developed and tested to predict an individual's risk of having type 2 diabetes based on responses gathered in the NHIS. While these models had high total accuracy, they were poor at correctly predicting if a patient was type 2 diabetic. A final model was trained and tuned, and while it represented a marked improvement over the first two models in accurately predicting if a patient was type 2 diabetic, it still performed at a poor enough level that this project can not recommend deploying the model to predict a patients risk of type 2 diabetes. The following report will discuss in detail how this conclusion was reached by going through the background, question, data selection, data pre-processing and feature engineering, modeling methods, the finalized model, and a conclusion.

Background and Question

According to the American Diabetes Association (ADA), 38.4 million Americans were living with diabetes as of 2021. Breaking that number down a bit more shows that 2 million of

those with diabetes had type 1 diabetes, and 8.7 million were living with undiagnosed diabetes (American Diabetes Association, n.d.). At the time, that accounted for 11.6% of the population living with diabetes, and Figure 1 shows that the total number of Americans living with diabetes



has increased over time.

Figure 1: Percentage of Total Diabetes Share over Time

There is a large percentage of people living with undiagnosed diabetes, with 2021 estimates from the ADA showing that 22.7% of people living with diabetes are undiagnosed. According to the Centers for Disease Control and Prevention (CDC), type 2 diabetes detection can be difficult due to the fact that symptoms can take a while to present, if they are ever present at all (CDC, 2023). Alarmingly, according to the ADA, in 2021, diabetes was the 8th leading cause among Americans (American Diabetes Association, n.d.). Coupled with the high rate of undiagnosed patients living with diabetes, it is apparent that there is a need to ensure that patients can recognize that they should speak to their healthcare provider, receive their necessary

diagnosis, and begin treatment for their diabetes as soon as possible in order to ensure the best care and outcome.

Through this background research, it became apparent that there was an important research question that needed to be answered. In order to help undiagnosed patients know that they should visit their physician, can a model be developed to show the risk of type 2 diabetes based on information available to the patient? This project hypothesized that there would be a connection between a person's responses to specific NHIS questions and whether or not they were non-diabetic or had type 2 diabetes because there are known health and lifestyle indicators associated with an increased risk of type 2 diabetes (*Symptoms & Causes of Diabetes - NIDDK*, n.d.). Given a patient's responses to specific NHIS questions, our team predicted that a model would be able to classify whether or not a patient had type 2 diabetes based on their responses.

Data

The NHIS was selected as a data source because this project aims to develop a model to predict diabetes based on information that would be readily available to the patient. The NHIS is a survey conducted by the CDC via the National Center for Health Statistics. "The main objective of the NHIS is to monitor the health of the United States population through the collection and analysis of data on a broad range of health topics. A major strength of this survey lies in the ability to categorize these health characteristics by many demographic and socioeconomic characteristics" (National Center for Health Statistics, 2023). Since the NHIS is focused on monitoring the health of Americans using health characteristics categorized by various demographic and socioeconomic characteristics, it was the best data set for this project to assess if there is a link between type 2 diabetes and any of the numerous questions asked in the NHIS.

An apparent issue with using survey data was that our survey data seemingly displayed sampling bias. While 11.6% of the population of the United States suffers from diabetes, the data from the surveys contained just under a 4% positive rate for a diabetes diagnosis. Since this project was interested in modeling type 2 diabetes, this presented an issue with class imbalance that will be addressed later on. Additionally, this data was only data on adults from the United States, so any model developed would be limited in its utility to just adult patients. The survey also did not include persons without a fixed household address, members of the military, persons living in long-term care institutions, people in correctional facilities, and US nationals living abroad. Regardless, the survey was gathered using clustered samples, which was sufficient to develop the model. Future iterations of this project should work on weighing the survey data to avoid the biases introduced and faced by this project.

The data was split by year on the NHIS web page in a CSV format. Initially, the project was trying to look at data from a ten-year span, but due to the collection methodology changing between the 2018 and 2019 surveys, only data from 2019 - 2022 was included. Within each year, there were discrepancies in the total number of features. As the survey matures, more questions are added each year, and so to combine all the years into a single dataset, only features that were shared among all years were included. 2019 data contained 534 features, 2020 data contained 617 features, 2021 data contained 622 features, and 2022 data contained 637 features. Using RStudio, these data sets were loaded and combined on common features for a final dataset containing 120,698 observations and 399 features.

Using Excel, the data dictionary provided by the NHIS, and research on diabetes causes from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), the team reviewed each feature and determined whether to keep it (Symptoms & Causes of

Diabetes—*NIDDK*, n.d.). Using information from the NIDDK and the team's intuition and curiosity about the relationship of type 2 diabetes with various socioeconomic features, the 399 features were reduced to 38 features for the initial exploratory data analysis portion of the project. These features are listed and detailed in Appendix B.

This project aimed to predict the risk of type 2 diabetes. While the dataset contained two variables relevant to a patient's diabetes status, 'DIBTYPE_A' and 'PREDIB_A,' a new variable was created to be the response. Within 'DIBTYPE_A,' prediabetic patients were shown as NA, but using 'PREDIB_A' allowed the creation of 'DIA_STATUS' with 0 representing non-diabetic, 1 representing prediabetic, 2 representing type 1 diabetes, and 3 representing type 2 diabetes. This project started out interested in multi-classification but shifted to binary classification after focusing on our interest in predicting type 2 diabetes and an acknowledgment that pre-diabetes is not a long-term diagnosis but rather a temporary situation in which a nondiabetic patient is at great risk of developing type 2 diabetes. For this reason, all type 1 diabetes responses were dropped from the dataset, and our response variable became 0 (non-diabetic) or 1 (type 2 diabetic).

Due to various reasons, extensive data cleaning and feature engineering on the dataset was necessary. The data initially appeared complete with limited NA values, but after initial exploratory data analysis, the data dictionary from the NHIS was consulted, and a big issue was noted. Most of the selected features were numeric variables, but after close inspection, they were seen to contain categorical placeholders. In many instances, values 7, 8, and 9 represented 'refused,' 'not ascertained,' and 'don't know.' These variables were all cleaned by dropping all 7s, 8s, and 9s since those values were in effect NA. Finally, categorical variables that were encoded as numeric were changed to represent their true values better. For example, 'REGION'

1, 2, 3, and 4 became ‘Northeast,’ ‘Midwest,’ ‘South,’ and ‘West’ respectively. The data codebook is linked in the appendix, and the highlighted variables had to have numeric values dropped. Figure 2 demonstrates an issue noticed with the variable age that was discovered during EDA

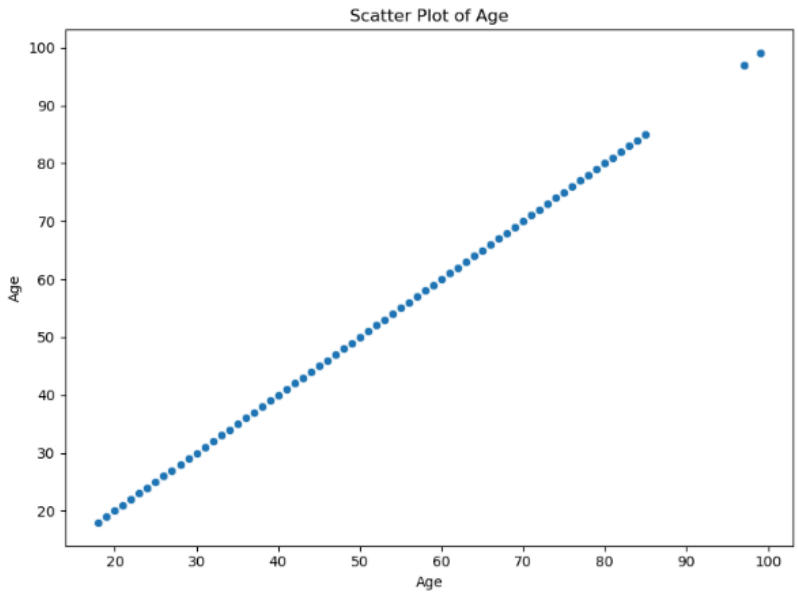


Figure 2: Plot of Age to Demonstrate Hidden Categorical Data

Here, it initially seemed that age had a few outliers at 97 and 99, but after analysis, the group discovered that there were three hidden categories. 85 represented all ages 85 and over, 97 represented refused, and 99 represented not ascertained. This issue was dealt with by breaking ages fully into categories before developing the final model. 18 to 25 were classified as ‘Young Adult,’ 26 to 44 were classified as ‘Middle Adult,’ 45 to 65 were classified as ‘Late Adult,’ and over 65 was classified as ‘Retired Adult.’

After working through the initial data cleaning, exploratory data analysis began with a correlation heatmap shown in Figure 3. This heatmap had a threshold of 0.7 and demonstrated a low correlation between most variables, including the response. The only highly correlated variables were height and weight.

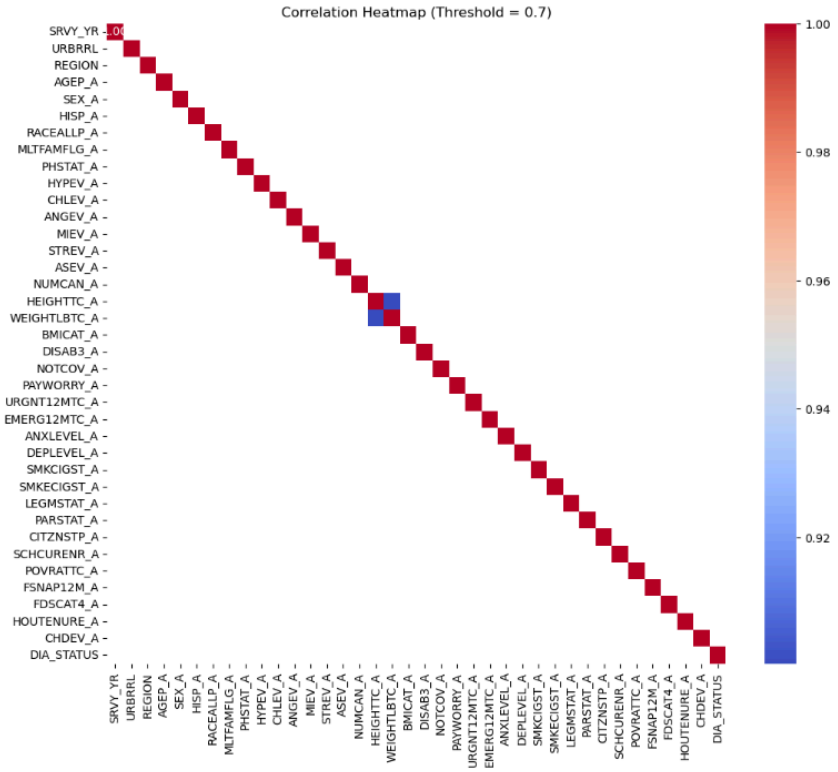


Figure 3: Correlation Heatmap for Initial Selected Variables (Threshold = 0.7)

Figure 4 was a variance inflation factor table that also confirmed the findings of the correlation heatmap. The only variables with high collinearity were height and weight. Due to this collinearity, and since BMI captured the relationship between height and weight in a way that this project was more comfortable with, the decision was made to drop height and weight.

	variable	VIF
0	Intercept	3345113.487
1	SRVY_YR	1.010427096
2	URBRRL	1.107555383
3	REGION	1.029503386
4	AGEP_A	2.233882531
5	SEX_A	1.887358286
6	HISP_A	1.161444607
7	RACEALLP_A	1.092889882
8	MLTFAMFLG_A	1.051341203
9	PHSTAT_A	1.586891831
10	HYPEV_A	1.44497708
11	CHLEV_A	1.315144076
12	ANGEV_A	1.159238582
13	MIEV_A	1.374277347
14	STREV_A	1.08474601
15	ASEV_A	1.044712607
16	NUMCAN_A	1.122723589
17	HEIGHTTC_A	10.64360195
18	WEIGHTLBTC_A	9.816369735
19	BMICAT_A	1.355554075
20	DISAB3_A	1.234712257
21	NOTCOV_A	1.153090659
22	PAYWORRY_A	1.218423614
23	URGNT12MTC_A	1.084168863
24	EMERG12MTC_A	1.165684197
25	ANXLEVEL_A	1.44716423
26	DEPLEVEL_A	1.452999481
27	SMKCIGST_A	1.260823139
28	SMKECIGST_A	1.274597753
29	LEGMSTAT_A	1.059042399
30	PARSTAT_A	1.315583918
31	CITZNSTP_A	1.150789647
32	SCHCURENR_A	1.150940877
33	POVRATTC_A	1.454145837
34	FSNAP12M_A	1.287364195
35	FDSCAT4_A	1.26316871
36	HOUTENURE_A	1.270249025
37	CHDEV_A	1.484471898
38	DIA_STATUS	1.227710103

Figure 4: Variance Inflation Factor Table

Using the ‘variance_inflation_factor()’ function from the ‘statsmodels’ library. The VIF (variance inflation factor) for each predictor variable can be seen. If a VIF value is closer to 1 that means collinearity is not a concern, but if it is significantly greater than 1, it will be a concern. Similar to the corrplot above only Height and Weight have high collinearity.

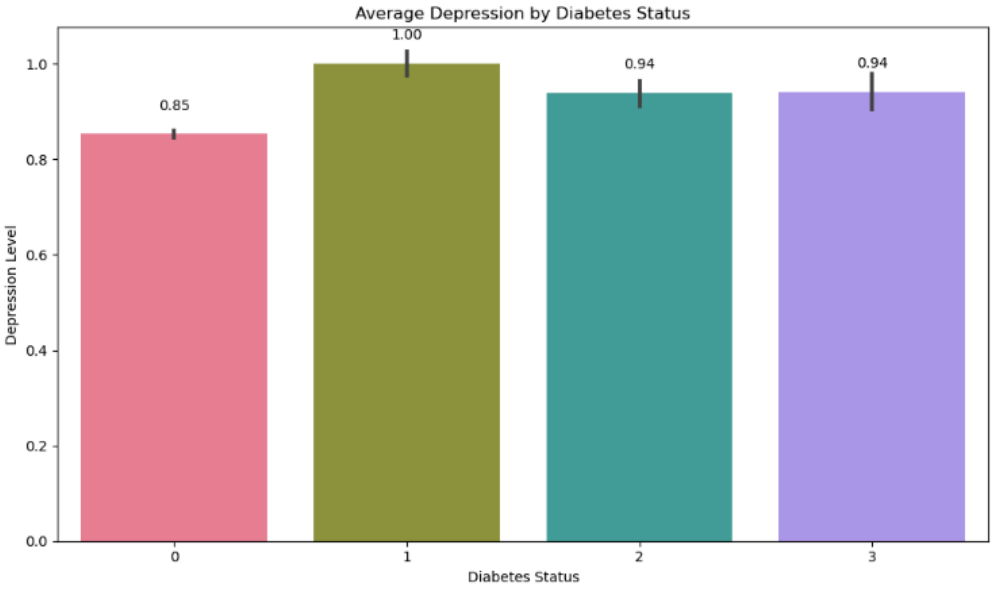


Figure 5: Depression Level by Diabetes Status

Looking at diabetes status vs average depression levels, it is apparent that for pre-diabetic, type 1, and type 2, there was a slightly higher level of depression than non-diabetic, by about 10%. On the other hand, in Figure 6, which looks at anxiety levels by diabetes status, there seems to be no pattern.

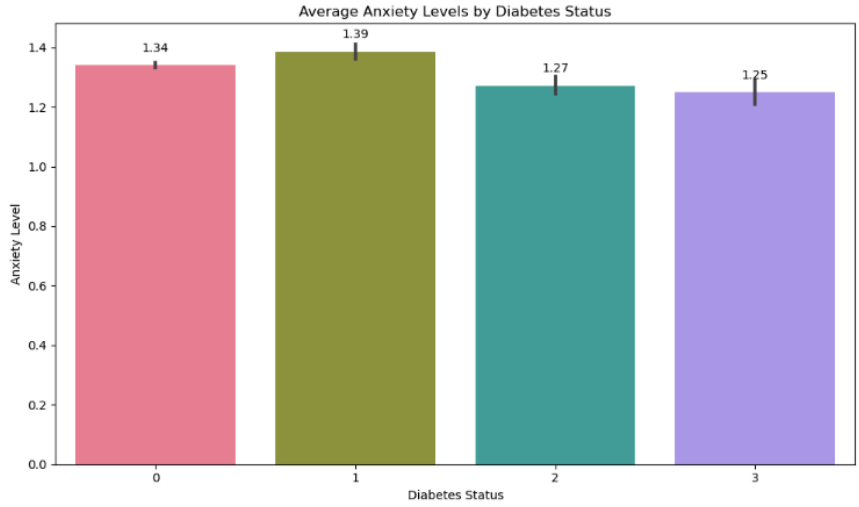


Figure 6: Anxiety Level by Diabetes Status

The histogram in Figure 7 shows four diabetes statuses with the poverty ratio frequency. The one with the most frequency is non-diabetic, and the least is type 1 diabetes. These histograms all show a right-skewed graph, with a small blip up at the tail end. The Boxplot shown in Figure 8 shows a higher average poverty rate in regions 1 and 4, while region 3 has the lowest.

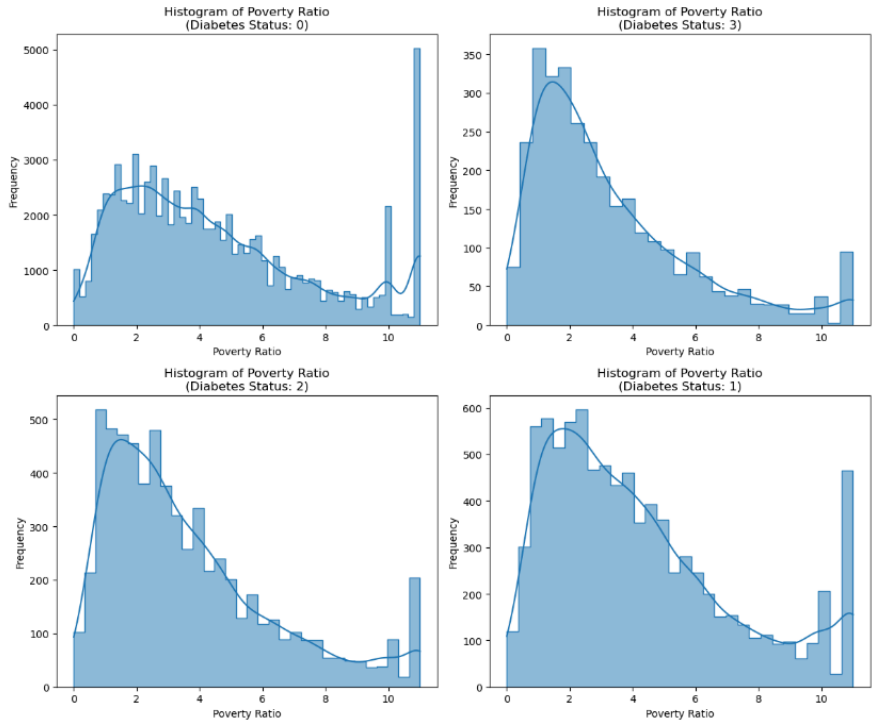


Figure 7: Poverty Level by Diabetes Status

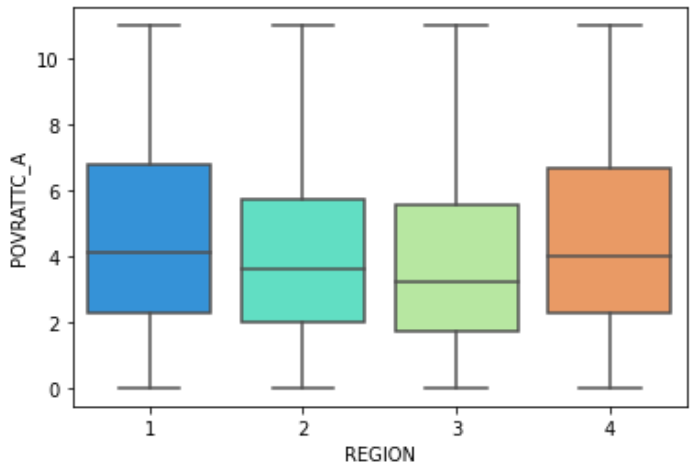


Figure 8: Poverty Rate by Region

Models

The pre-processing and feature engineering plan originally involved cleaning the data by removing NAs or imputing new values to solve any issues with NAs. Then, categorical values would be one-hot encoded, and PCA would be conducted to enable feature selection. A change that occurred in the original plan was brought about due to discoveries during exploratory data analysis. Values that looked like outliers were investigated using the data codebook, and it became apparent that certain numeric values contained categorical placeholders. Before one-hot encoding or conducting PCA, it was important to fix these categorical values within the numeric columns (e.g., AGEP_A).

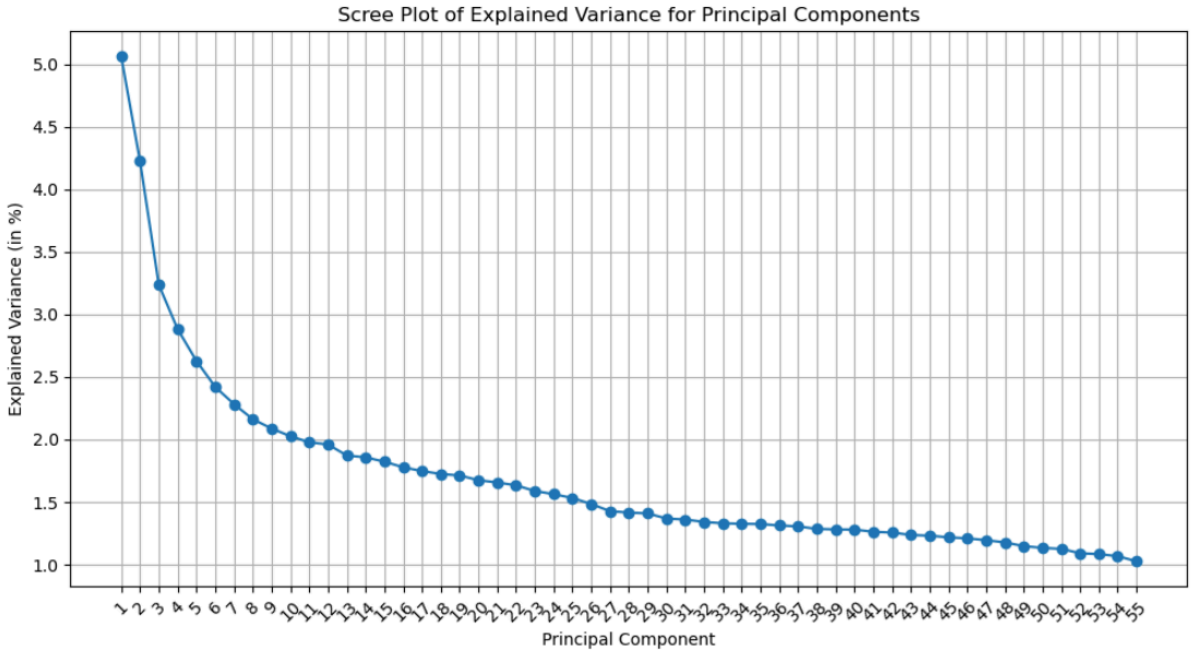
Our initial EDA revealed missing and placeholder values in several variables. For variables with only one missing value, we'll opt to remove the entire row containing the missing data. This approach avoids introducing potentially misleading values through imputation for these isolated cases.

The data format itself presented no issues, so cleaning focuses on ensuring complete cases with minimal missing values. Based on the EDA findings, variables with high redundancy

or low relevance (e.g., weight and height when BMI is available) might be removed entirely. Furthermore, variables like "urgent care visits" and "emergency room visits" with significant overlap will be combined into a single variable indicating healthcare utilization within the past year.

Initially, an attempt was made to model the data using linear discriminant analysis (LDA). However, LDA was restricted to a single component, limiting its ability to capture the data's inherent dimensionality. To address this, Principal Component Analysis (PCA) was conducted, which offers greater flexibility in dimensionality reduction.

Principal Component Analysis (PCA), an unsupervised approach for dimensionality reduction, was executed. This technique helps manage the high number of features (potentially 90 columns) by identifying a smaller set of uncorrelated components that capture most of the data's variance. Before applying PCA, categorical variables were converted to factors and then one-hot encoded to ensure compatibility with the algorithm. Finally, all data was normalized or scaled to ensure all features contributed equally during model training.




```
[0.05062152 0.04225203 0.03236842 0.02882336 0.02628461 0.02420462
 0.02282669 0.02162132 0.02087584 0.02026257 0.01977525 0.01960431
 0.01872041 0.01857343 0.01822678 0.0177782  0.01747553 0.0172443
 0.01712321 0.01674613 0.01656435 0.01633892 0.01588194 0.01563064
 0.01532267 0.01483917 0.01427075 0.01415822 0.01408982 0.01368005
 0.01360199 0.01341077 0.01330087 0.01325984 0.0132401  0.01312642
 0.0130508  0.01285367 0.01281197 0.01279682 0.0126235  0.01255965
 0.0123674  0.01231271 0.01217614 0.01209838 0.01194927 0.01177395
 0.0114771  0.01133473 0.01125445 0.01090414 0.01083571 0.0106742
 0.01026176]
```

Figure 9: PCA Scree Plot and Explained Variance Table

After conducting PCA, it was determined that 55 components were optimal, as this threshold marked a significant decrease in explained variance beyond which diminishing returns were observed. Notably, the first 10 components each contribute over 2% of the variance, while the remaining components collectively explain an additional 45%. Collectively, these 55 components account for 90% of the total variance within the data, suggesting a high degree of successful dimensionality reduction. The scree plot and explained variance ratio array, presented above (Figure 9), visualize and quantify this variance distribution across components.

Leveraging these insights, our team incorporated the derived 55 PCA components as features into each of our three models. This inclusion of the reduced dimensionality representation could enhance model performance by focusing on the most informative aspects of the data.

An initially deployed model was a logistic regression model due to the ease of interpretation and implementation. During the preprocessing and feature engineering step, PCA identified 55 components, shown in Figure 9, that were used to develop the logistic regression model. By transforming the training and testing data using the PCA selection, a logistic regression model was created with an accuracy on the testing data of 0.963. While this appears to be highly effective due to the accuracy, a look at the confusion matrix in Figure 10 shows that that is not the case. As shown by this confusion matrix, most people in the test set are non-diabetic. The logistic regression model tended to overpredict non-diabetic cases, as is shown by the 631 false non-diabetic responses compared to 20 true Type 2 diabetic predictions.

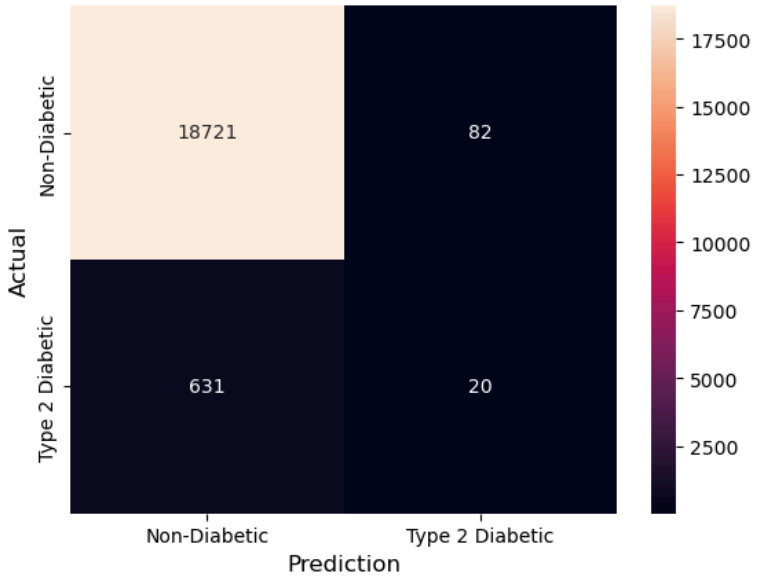


Figure 10: Logistic Regression Confusion Matrix

When proceeding with a logistic regression model, it is important to understand what

assumptions are made to use this model properly. The assumptions for logistic regression include that the dependent variable is binary for binary logistic regression, independence of observations, no multicollinearity between independent variables, linearity of independent variables and log odds, and finally, a large sample size (*Assumptions of Logistic Regression*, n.d.). First, a binary dependent variable was created during the preprocessing step in order to filter down observations to either Type 2 Diabetic or non-diabetic, so the data fulfills the requirement for a binary dependent variable. Through the exploratory data analysis step, it was determined that there was no multicollinearity between the selected features. Additionally, the collection method of the NHIS demonstrates independence of observations and ensures a large enough sample size. Since PCA transforms the original variables, it complicates the ability to test for linearity between the predictors and the logit. For this reason, we proceeded without directly testing for this assumption, but if future logistic regression attempts are made with the features directly, this assumption will have to be tested.

A receiver operating characteristic curve (ROC) shown in Figure 2 was generated to show the model's performance. The graph also shows the area under the ROC curve (AUC) of 0.71. This value demonstrates that our model is not likely to be of clinical usefulness, and a lower value suggests that our model is likely not statistically significant. The ROC, combined with the confusion matrix above, show that the classification of our data presumed that non-diabetic was the positive class and Type 2 Diabetic was the negative class. So, the True Positive Rate (TPR) and the False Positive Rate (FPR) are calculated based on the model’s ability to predict the non-diabetic case correctly. This is still helpful since a model that was better at predicting non-diabetic cases in the binary system would, therefore, be better at predicting the inverse, Type 2 Diabetic cases.

In general, the first pass at a logistic regression model made a model that did not fit our testing data very well. In determining whether or not this model might have been overfitting, looking at the classification report is helpful. This allows a comparison between how the model performed on the training data versus the testing data. If the training data vastly outperformed the testing data, then there could be overfit issues. Figure 12, on the next page, shows how the model performed on the testing data. Regarding the classification of interest, Type 2 Diabetes, it is apparent that the model performed poorly on the test data with a low precision of 0.1961, low recall of 0.0307, and low f1 of 0.0531.

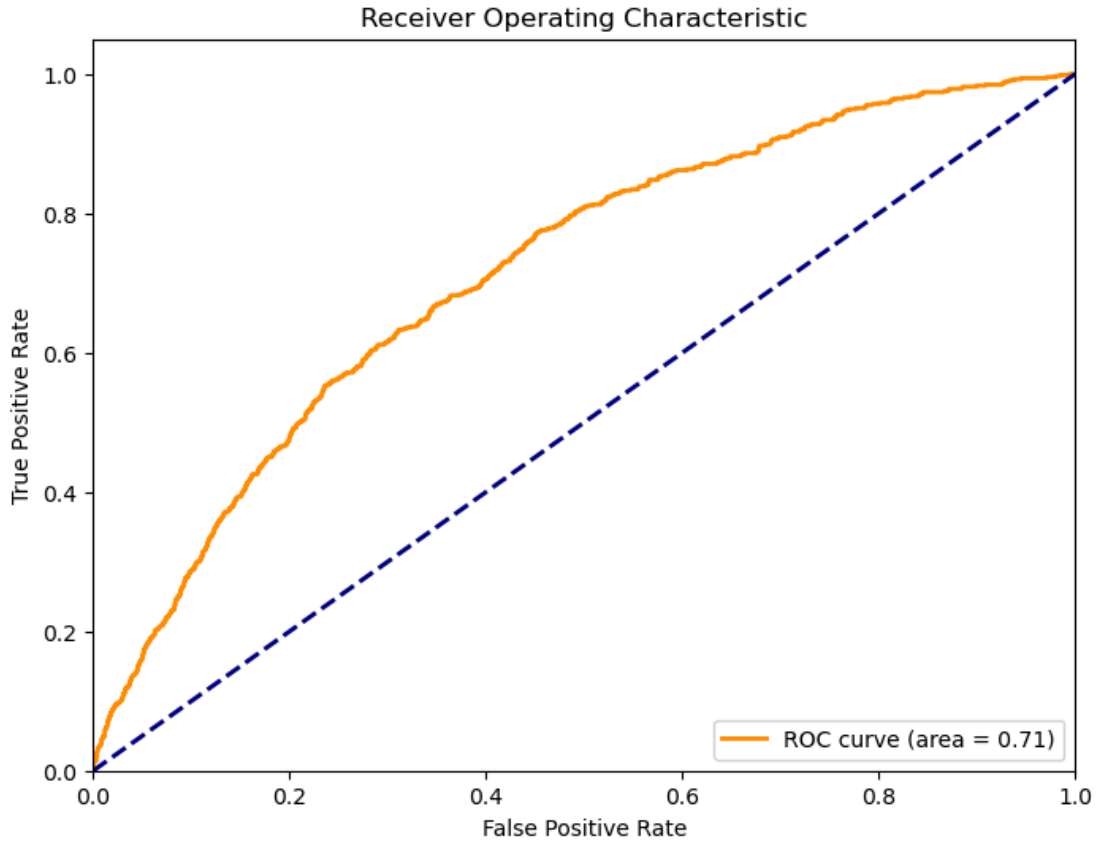


Figure 11: Logistic Regression Model ROC Curve

	precision	recall	f1-score	support
Non-Diabetic	0.9674	0.9956	0.9813	18803
Type 2 Diabetic	0.1961	0.0307	0.0531	651
accuracy			0.9633	19454
macro avg	0.5817	0.5132	0.5172	19454
weighted avg	0.9416	0.9633	0.9503	19454

Figure 12: Classification Report, Logistic Regression Test Data

Figure 13 and Figure 14 both show how the model performed on the training data. Figure 13 is the classification report for the model performing on the training data, and Figure 14 is a confusion matrix for the model's predictions based on the training data. The precision of the model is better than the precision of the testing data since it predicted 17 positives, and of those, 8 are true positives. This is technically a better performance than on the testing data, but the recall and the f1 score are appropriately affected by the tiny percentage of true predictions. On the training data, the model predicts non-diabetic a vast majority of the time, so even though it has high precision, the model appears to underfit.

	precision	recall	f1-score	support
Non-Diabetic	0.9666	0.9999	0.9830	75208
Type 2 Diabetic	0.4706	0.0031	0.0061	2604
accuracy			0.9665	77812
macro avg	0.7186	0.5015	0.4945	77812
weighted avg	0.9500	0.9665	0.9503	77812

Figure 13: Classification Report, Logistic Regression Training Data

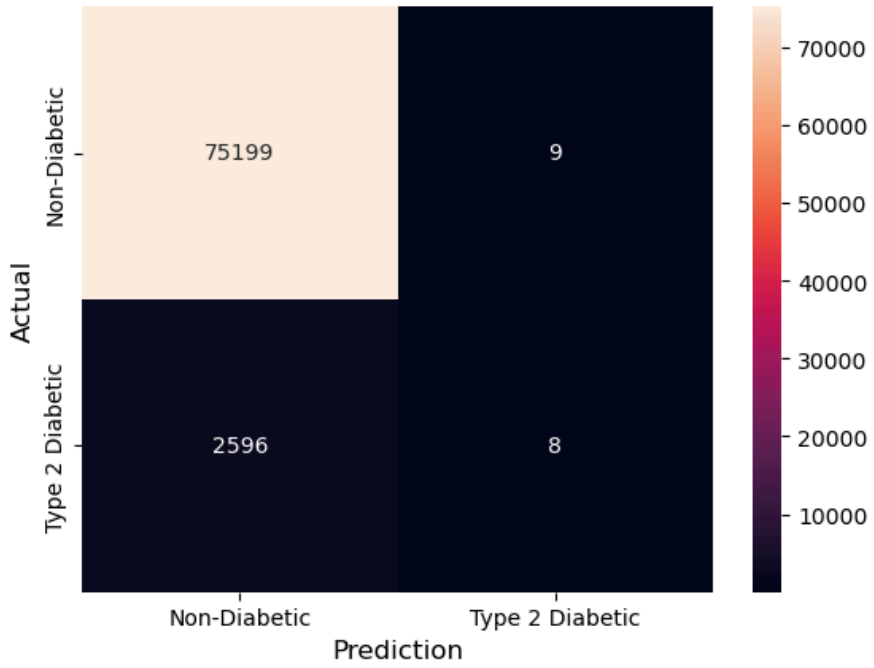


Figure 14: Logistic Regression Confusion Matrix on the Training Data

The other model leveraged in this report was a Random Forest classifier, "A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting" (*Sklearn.ensemble.randomforestclassifier*). Our team chose This model because it can handle datasets with continuous and categorical variables for classification problems. Our dataset consists of all three of these things. Additionally, random forests are complex and prevent overfitting in the data, with many use cases and better performance on new data than other models.

The random forest model utilized the 55 components identified during PCA the previous week. The 'Sklearn.ensemble' was used with 'RandomForestClassifier,' which has 25 n_estimators, an entropy criterion, and a random state of 0. Then, the test results were predicted using a model score, which yielded a score of 0.9662 and a confusion matrix. Based on the confusion matrix results, it is apparent that the model is not as accurate as the model score

shows.

	0	1
0	18796	7
1	650	1

Figure 15: Random Forest Confusion Matrix

A box plot of features sorted from most important to least important was produced.

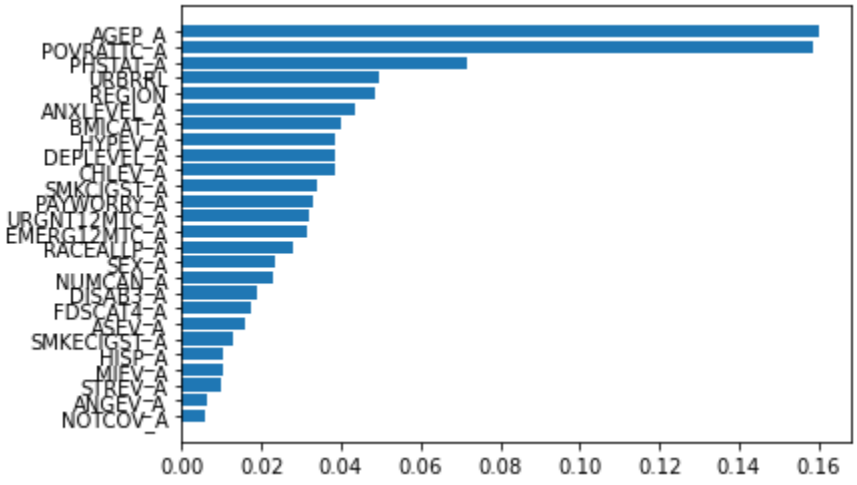


Figure 16: Box Plot Showing Feature Importance in Random Forest

Though this random forest model did not offer accurate predictions, this box plot using the random forest model ranked the importance of each of our variables, which will be useful in testing and refining XGBoost. A receiver operating characteristic curve (ROC) was plotted from the random forest model.

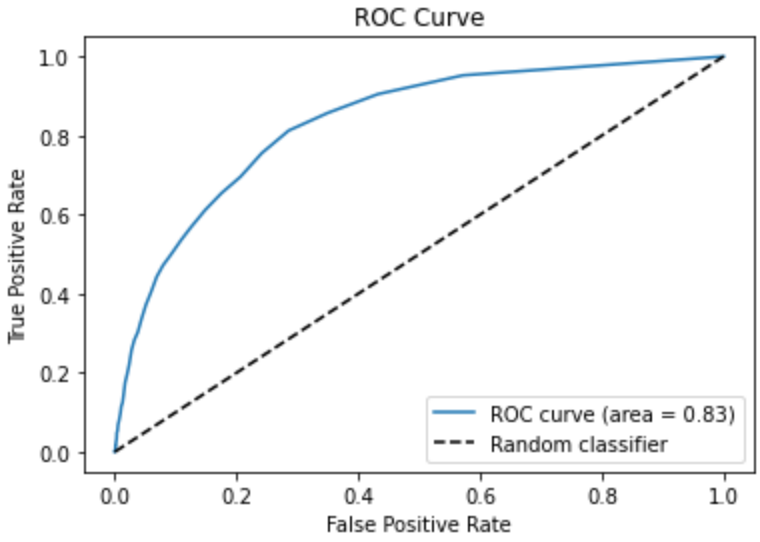


Figure 17: Random Forest ROC Curve

With an area of 0.83, the hope is for the ROC curve to be closer to 1. This will visually have the ROC curve hugging the y-axis and the top of the plot getting as close as possible to a 90-degree angle above the random classifier line. The false positive rate (FPR) and true positive rate (TPR) for different classification thresholds were calculated. A predicted class probabilities histogram was plotted. This shows that, in most cases, the predictive probability is 0, although some rate above 0.4.

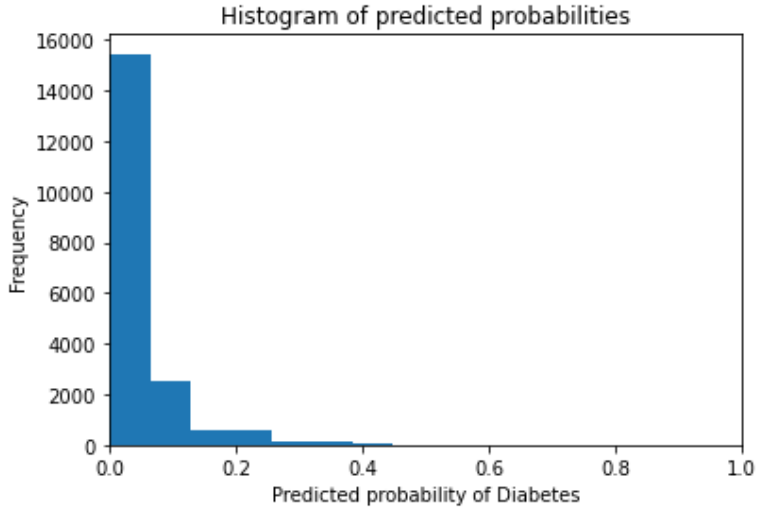


Figure 18: Random Forests Predicted Class Probability Chart

In XGBoost, achieving the best possible performance often involves carefully adjusting hyperparameters, the settings that control how the model learns. The param_grid dictionary is a powerful tool that helps define a search space for these hyperparameters. Also using a PCA with 45 components, instead of 55 from the last model.

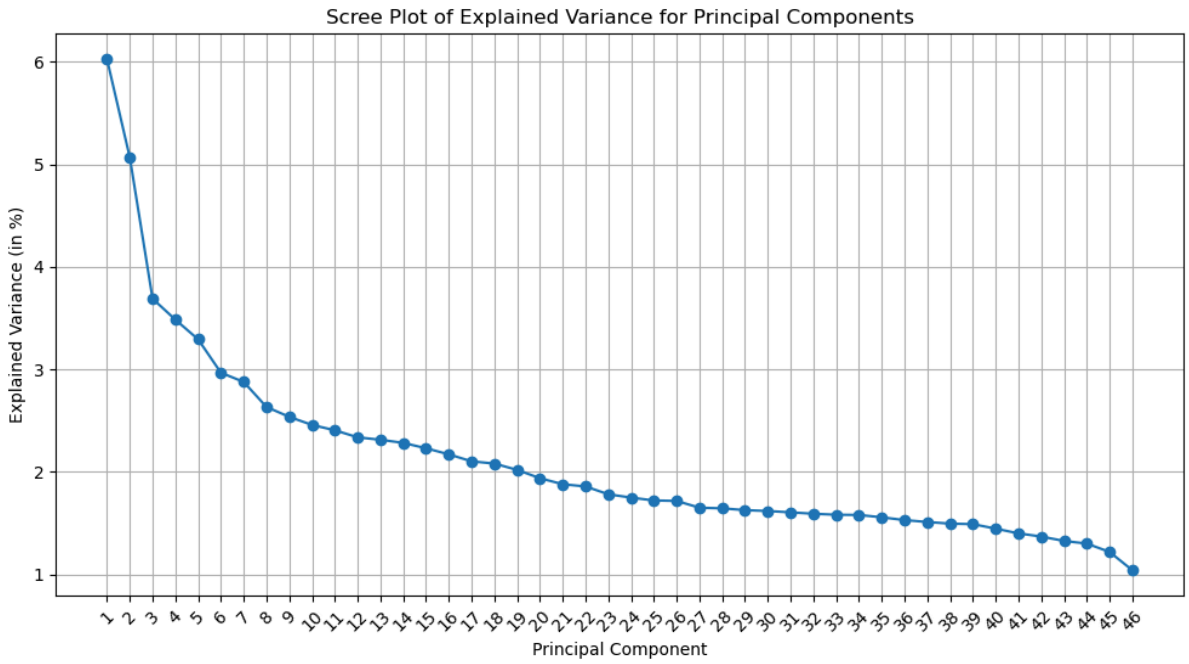


Figure 19: Scree Plot for PCA

The names of the hyperparameters, like max_depth (tree complexity) and learning_rate (learning pace), act as the keys in the dictionary. Each key has a corresponding list of values, which are the different settings to try for that particular hyperparameter. In this example, different tree depths (5, 8, 10, 20, 50, 75) and learning rates (0.85, 0.9, 0.1, 0.2) are being explored.

Max_depth and subsample are crucial for managing model complexity. Higher max_depth allows for more intricate decision trees but can lead to overfitting. Using a subset of data for each tree can introduce randomness and potentially reduce overfitting.

Learning_rate determines how much the model adjusts its weights based on errors. Lower values lead to smaller, more cautious adjustments, which can prevent overshooting the optimal solution but may also slow down learning.

The chosen values (10, 11, 12, 20, 50, 75 for max_depth, 0.85, 0.9, 0.1, 0.2 for learning_rate) provide a range to explore different levels of complexity and learning speed. With param_grid, tools are instructed to train XGBoost models with various combinations of these settings and identify the one that performs the best on the data.

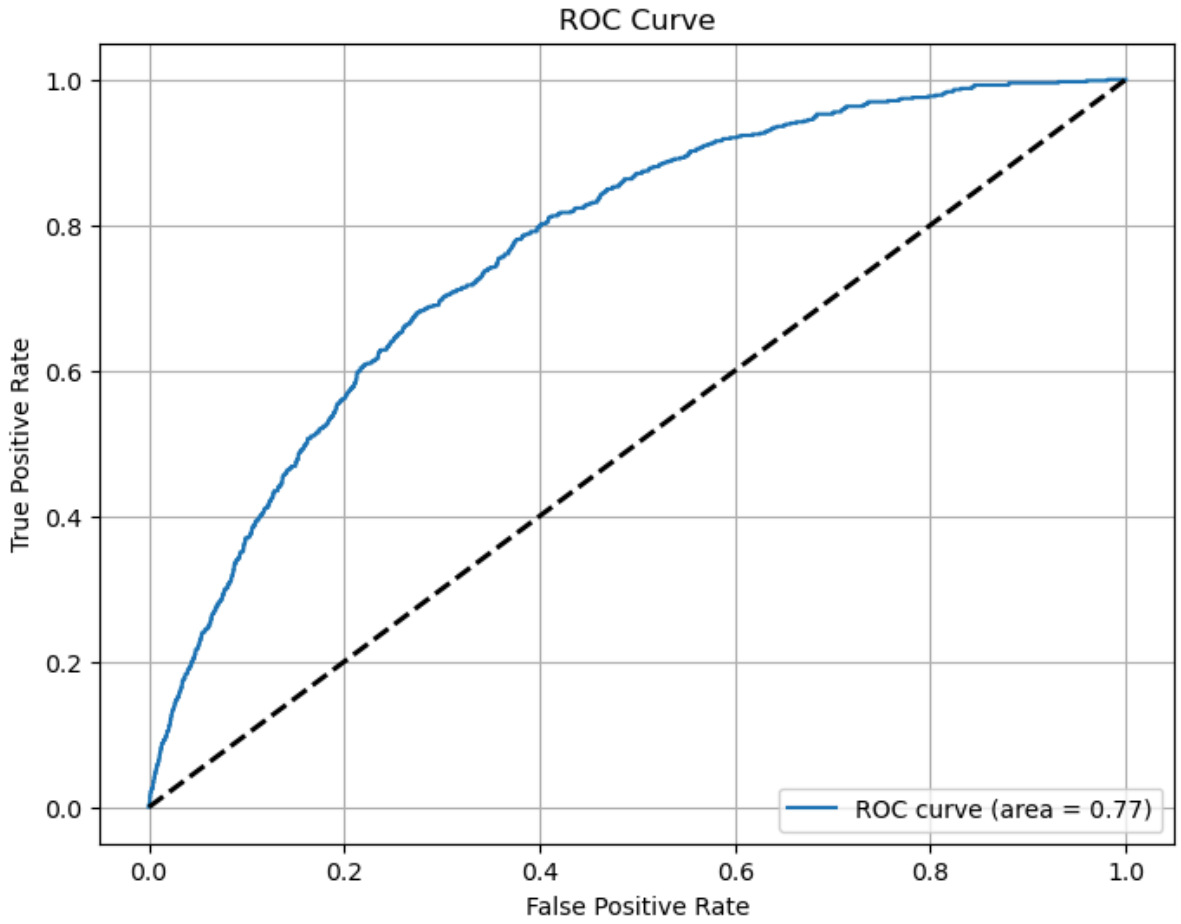


Figure 20: ROC curve for XGboost

While there were initial hiccups getting the model to function smoothly, debugging efforts have been successful. The model currently boasts high accuracy, but this project requires

maximizing true positives, which currently sit at 33. To achieve this, the model's hyperparameters will be fine-tuned. Additionally, exploring a different set of features might prove beneficial in identifying the patterns most relevant to true positives. By combining these adjustments, the model's ability to accurately detect the desired outcomes should be significantly improved.

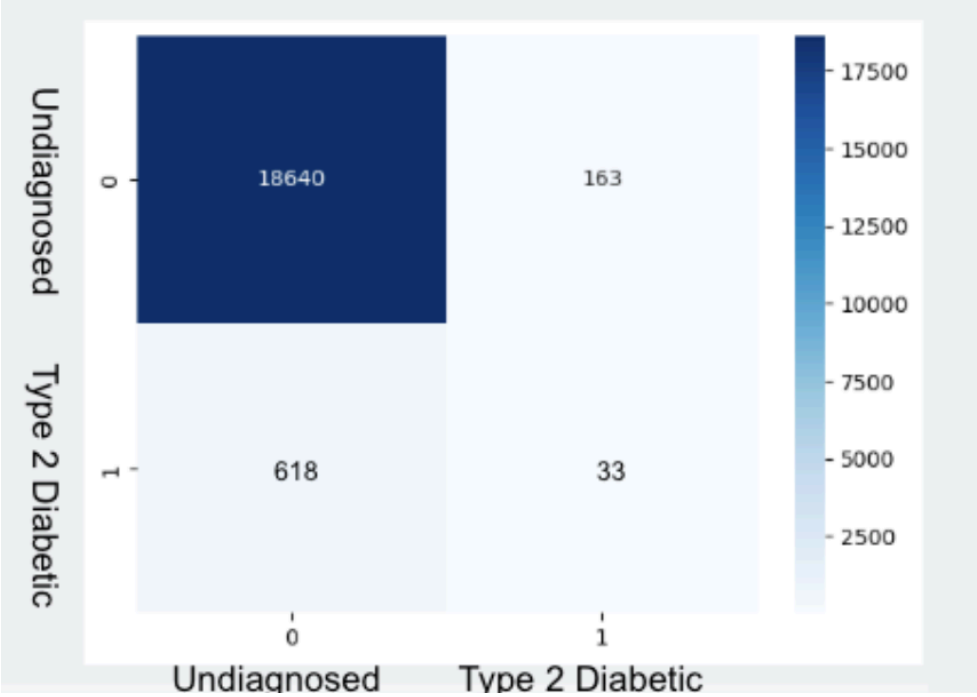


Figure 21: XGBoost confusion matrix

Conclusion

In conclusion, the original goal of providing the stakeholders (patients and doctors) with a product that would give a risk rating for type 2 diabetes based on health background, socioeconomic, and lifestyle variables was unsuccessful. Extensive research on type 2 diabetes risk factors and preprocessing through Principal Component Analysis was conducted to best prepare the dataset for the presented research question. The data set was put through three

supervised learning models: Logistic Regression, Random Forest, and XGBoost. Logistic Regression performed with 47% accuracy, and the Random Forest model with 12% accuracy. XGBoost had marked improvement over Logistic Regression and Random Forests; even still, this was not enough to recommend deploying the model to predict a patient's risk of type 2 diabetes. Though this project was unsuccessful, our team realizes what an important complex issue Diabetes is to the healthcare industry. It will continue working on this idea of at-home healthcare resources using the lessons learned in follow-on products.

Discussion and Next Steps

Our analysis revealed two key areas for improvement in the model's ability to answer the original question. First, the dataset suffers from a significant amount of missing data, likely due to incomplete surveys. This limits the model's ability to learn effectively from the data and hinders its usefulness in providing definitive answers.

While the current model can identify some patterns in the data, these limitations prevent it from definitively answering the proposed question. To address this, we recommend focusing on improving data quality. This could involve addressing missing values in the current dataset or obtaining a new dataset with a lower missing value rate. Additionally, exploring features like waist circumference and blood sugar levels could significantly enhance the model's ability to capture relevant relationships.

Despite these limitations, the current model serves as a valuable foundation for further development. By incorporating a higher-quality dataset with a broader feature set, we can build a more robust model that effectively answers the original question. The insights gained here can also inform future data collection efforts to ensure better data quality for future analysis. It's

important to remember that the findings from this analysis are limited by the current data and feature set, and the current model should not be used for medical predictions.

References

American Diabetes Association. (n.d.). *Statistics About Diabetes*. https://diabetes.org/about-diabetes/statistics/about-diabetes

Assumptions of Logistic Regression. (n.d.). Statistics Solutions. Retrieved April 15, 2024, from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/

CDC. (2023, April 18). *Type 2 Diabetes*. Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/basics/type2.html

National Center for Health Statistics. (2023, June 29). *2022 National Health Interview Survey*. https://www.cdc.gov/nchs/nhis/2022nhis.htm

Symptoms & Causes of Diabetes—NIDDK. (n.d.). National Institute of Diabetes and Digestive and Kidney Diseases. Retrieved April 30, 2024, from https://www.niddk.nih.gov/health-information/diabetes/overview/symptoms-causes

Sklearn.ensemble.randomforestclassifier. scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Appendix A: Github Link

https://github.com/MarkusHoyt/Team1_capstone

Appendix B: Data Dictionary

Link to the codebook for selected variables:

https://drive.google.com/file/d/11X_KtGCyEtBbJ4GAvGWVhnZAG4fR92Nk/view?usp=sharing

Green Highlight - demonstrates that a feature was numeric but had categoric placeholders for NA that had to be removed

- SRVY_YR: Survey year, indicating the year in which the data was collected.
- URBRRL: Urban/rural classification.
- REGION: Geographic region of the participant.
- AGEP_A: Age of the participant.
- SEX_A: Gender or sex of the participant.
- HISP_A: Hispanic or Latino ethnicity status of the participant.
- RACEALLP_A: Race or racial identity of the participant.
- MLTFAMFLG_A: Multi-family household flag.
- PHSTAT_A: Physical health status.
- HYPEV_A: Hypertension or high blood pressure evaluation.
- CHLEV_A: Cholesterol evaluation.
- ANGEV_A: Angina or chest pain evaluation.
- MIEV_A: Myocardial infarction or heart attack evaluation.
- STREV_A: Stroke evaluation.
- ASEV_A: Asthma evaluation.
- NUMCAN_A: Number of cancers evaluated.
- HEIGHTTC_A: Height of the participant in inches.

- WEIGHTLBTC_A: Weight of the participant in pounds.
- BMICAT_A: Body Mass Index (BMI) category.
- DISAB3_A: Disability status.
- NOTCOV_A: Health insurance coverage status.
- PAYWORRY_A: Payment worry for healthcare services.
- URGNT12MTC_A: Urgent care visit within the last 12 months.
- EMERG12MTC_A: Emergency care visit within the last 12 months.
- ANXLEVEL_A: Anxiety level.
- DEPLEVEL_A: Depression level.
- SMKCIGST_A: Smoking status (cigarettes).
- SMKECIGST_A: Smoking status (e-cigarettes).
- LEGMSTAT_A: Legal marital status.
- PARSTAT_A: Parental status.
- CITZNSTP_A: Citizenship status.
- SCHCURENR_A: School enrollment status.
- POVRATTC_A: Poverty ratio.
- FSNAP12M_A: Food stamp participation within the last 12 months.
- FDSCAT4_A: Food security category.
- HOUTENURE_A: Housing tenure.
- CHDEV_A: Childhood development.
- DIA_STATUS: Diabetes status.