****

**BC2407 Analytics II: Advanced Predictive Techniques**

**Semester 2 Academic Year 20/21**

**Semester Project**

|  |  |  |
| --- | --- | --- |
| Members | Natasha Khoo Mei Hui | U1910526E |
| Ng Chi Hui | U1922243C |
| Lim Wei Jie Spencer | U1810413H |
| Wong Kang Xun | U1810804L |
| Jarrel Kong Kai Jit | U1910901J |
| Seminar Number | 6 | |
| Group Number | 3 | |
| Instructor | Kevin Ngui | |
| Submission Date | 04-April-2021 | |

**Content Page**

[1. Executive Summary](#_hs6blk439gf3) 2

[1.1 Overview](#_ihu02nbel03g) 2

[1.2 Situational Analysis](#_om0swqejcog8) 2

[1.3 Key Issues](#_4s406dmxdm3c) 2

[1.4 Key Findings](#_ndlg7b8swezf) 2

[1.5 Proposed Action Plans](#_467xkzs3hequ) 2

[1.6 Limitations](#_ypyd33g7uqh0) 2

[2. Introduction](#_ugfa1yrm68yq) 3

[2.1 Project aim](#_ebzg07sno0gr) 3

[2.2 Background on Type 2 Diabetes](#_xman04zbydh3) 3

[2.3 Case Justification](#_1cmm1nkhq2v6) 3

[3. Situational Analysis](#_nswme2cgod04) 4

[3.1 PEST Analysis](#_ql7z0rjyxvey) 4

[3.2 Overview of Austin, Texas](#_r9bxosif3kg0) 5

[4. Methods of Data Analysis](#_in7kq82xyxp) 6

[4.1 Data Preparation & Cleaning](#_igfchtx4j42v) 6

[4.2 Exploratory Data Analysis](#_pw59ml1v8qp7) 7

[4.3 Model Building](#_ajhaxeixtn0i) 8

[4.4 Model Evaluation & Selection](#_18x2bld2uatk) 16

[5. Key Findings from data analysis](#_1xvoat66y3eu) 17

[6. Recommendation](#_yiog21yifz96) 18

[6.1 Diabetes Detection](#_c406vybsasb6) 18

[6.2 Diabetes Policies - Promoting Active Lifestyles](#_5n7sfpqn416r) 19

[7. Overall Limitation and Future Work](#_vn3mlz9lxrcg) 20

[7.1 Uncertainty over economic factors](#_c6aowp9386hz) 20

[7.2 Exercise intensity and duration](#_d0cqlhkxbjs4) 21

[7.3 Carbohydrate and sugar consumption](#_u4t2brpixmv5) 21

[7.4 Limited Dataset](#_w9t6lktk0dkb) 21

[8. Conclusion](#_baei4y82vto0) 22

[9. Bibliography](#_90s9ld98sst4) 23

[10. Appendix](#_ykf2oklnxbnv) 29

# 

# 1. Executive Summary

## 1.1 Overview

The following report is prepared to assist the American Diabetes Association (ADA) in their mission to address diabetes in the US. The objective of this report is to present our research, analysis, action plans and detailed descriptions of how we intend to improve the rate of type 2 diabetes in America. The methodology applied in our research and analysis was mainly secondary sources of information coupled with data analysis on our dataset (obtained from secondary sources).

## 1.2 Situational Analysis

Type 2 diabetes is a global pandemic. According to the National Diabetes Statistics Report (2020), 90% to 95% of all diabetes are caused by type 2 diabetes. More than half a billion people worldwide are living with diabetes and this number is projected to increase (Saeedi et al., 2019). This issue is especially pertinent in the US, as it imposes a hefty $327B strain on the already stretched national healthcare system in 2017 (ADA, 2018).

## 1.3 Key Issues

The ADA has done much in terms of diabetes detection (Risk Assessment for Diabetes) and implementing campaigns to curb diabetes. However, there remains room for improvement. From our team’s secondary research, we have found that there were several gaps with their approach.

Diabetes Detection: Little emphasis on social determinants of health (SDOH) during diabetes screening.

Diabetes Policies and Programs: SDOH are also not understood and considered sufficiently. Thus, not enough emphasis is placed on it when developing policies and programs.

## 1.4 Key Findings

Through our data analysis, we found that the most important SDOH factors are Previous Diabetes Education, Exercise, Medical Home Category, and Education level. Among these factors, Previous Diabetes Education and Exercise are SDOH factors that can be influenced by the ADA. Secondary research has also backed our key findings of Previous Diabetes Education and Exercise, where it is shown that people only receive diabetes education after they have been diagnosed with diabetes (Powers M. et. al, 2016), and that exercise lowers the risk of diabetes (Sheri R. et.al, 2010).

## 1.5 Proposed Action Plans

We propose the following to improve their diabetes detection program and policies. In terms of diabetes detection, a Diabetes Risk Assessment weighing important variables can be filled by Americans to predict their risk of diabetes. In terms of diabetes programs, incentivizing participants to exercise and creating targeted programs for the elderly (vulnerable group) is key to prevent diabetes.

## 1.6 Limitations

There are some uncertainties regarding economic factors as our current gauge is only based on insurance categories, which may not be accurate. There are also some variations in terms of exercise intensity, which may be subjective, and duration which are information that we do not have from our dataset. This same variation is also seen in the carbohydrate and sugar consumption. Last, the data set is centered around Austin geographically, which is only a small portion of the US population and may not accurately represent the entire US population. The lack of granular data might result in inaccuracies in our predictions and thus recommendations.

# 2. Introduction

## 2.1 Project aim

Our project aims to assist American Diabetes Association (ADA) in their mission to fight diabetes in the US. We aim to do this through analytics, which can allow us to draw greater insights to specific indicators that aid the ADA in predicting whether segments of the population are likely to develop diabetes, and prioritize early detection and prevention plans for them.

It is established that patients with Type 2 diabetes can be screened and diagnosed by doing several tests such as the fasting plasma glucose, oral glucose tolerance and random plasma glucose test (U.S Department of Health & Human Services, 2004). However, how social, economic, and environmental factors contribute to diabetes is not as well understood. By studying these factors, ADA will be able to

1. Provide a more holistic risk assessment for detecting diabetes on the individual level, and
2. Design more effective policies and programs that reduce the prevalence of diabetes.

## 2.2 Background on Type 2 Diabetes

1. What is Diabetes?

Diabetes is a group of metabolic diseases characterized by hyperglycemia resulting from defects in insulin secretion, insulin action, or both (ADA, n.d.). Our study will focus on type 2 diabetes, as according to the National Diabetes Statistics Report (2020), 90% to 95% of all diabetes are caused by type 2 diabetes (Centers for Disease Control and Prevention, 2020).

1. Diabetes as a global pandemic

Diabetes is a global pandemic. Half a billion people worldwide are living with diabetes and this number is projected to increase to an astonishing 25% in 2030 and 51% in 2045 (Saeedi et al., 2019).

1. Effects of Diabetes

Disease leads to a lower quality of life for the individual, and overall decreased productivity for society. The decrease in productivity and wellbeing has long been a cause for concern for governments, whose role is to maximize the welfare of their people. According to Tuncelli K. (2005), “diabetes affects patients, employers, and society not only by reducing employment but also by contributing to work loss and health-related work limitations for those who remain employed.”

1. Reasons to detect diabetes early on

As with many diseases and medical conditions, early detection is critical when it comes to survivability and recoverability of a patient. This is an especially salient point in diabetes, as diabetes is incurable. Furthermore, diabetes exacerbates other health complications a patient might have, such as premature heart disease, stroke, blindness, limb amputations, and kidney failure. (U.S Department of Health & Human Services, 2004).

## 2.3 Case Justification

1. Background on the Diabetes Situation in the United States

The US is fighting an uphill battle against diabetes. With more than a 100 million cases of diabetes and prediabetes, 30% of the population are at risk of facing complications caused by diabetes (Centers for Disease Control and Prevention, 2018). Despite being a first-world country with one of the best healthcare systems in the world, 90% of its prediabetic citizens and 25% of its diabetic citizens are oblivious to their conditions (Centers for Disease Control and Prevention, 2020). Ignorance prevents individuals from seeking medical aid or changing their lifestyle to prevent/minimize late-stage effects such as premature heart disease, stroke, blindness, and limb amputation. On a macro front, diabetes imposes a $327B strain on the already stretched national healthcare system in 2017 (ADA, 2018).

Diabetes is on an upwards trend from 1999 to 2016. The total number of diabetics amongst adults aged 18 or older grew from 9.5% to 12.0% between the period of 1999 to 2016 (Centers for Disease Control and Prevention, 2020), and this number is only projected to increase.

1. Gap in Current Research

Social determinants of health (SDOH) have been researched to examine its impact on the US population. SDOH are the conditions in the environments where people are born, live, learn, work, play, worship, and age that affect a wide range of health, functioning, and quality-of-life outcomes and risks (Office of Disease Prevention and Health Promotion, 2020).

SDOH includes:

(1) Economic Stability

(2) Education Access and Quality

(3) Health Care Access and Quality

(4) Neighborhood and Built Environment

(5) Social and Community Context

However, there remain gaps in the study of how people of different characteristics (e.g., age groups or socio-economic statuses) may be differentially impacted by SDOH, which in turn affect their risk of contracting diabetes (ADA, n.d.). Therefore, SDOHs will be examined in accordance with different characteristics to contribute to a better understanding of its impact on diabetes risk.

# 3. Situational Analysis

## 3.1 PEST Analysis

Through the PEST analysis, important external factors that should be taken into consideration to assess the viability of our project, and to guide our solution generation. The following factors are identified as important from the PEST analysis:

### 3.1.1 Political factors

As of 2019, the US remains neither very politically stable nor unstable, standing at 0.3 on the political stability index (-2.5 weak; 2.5 strong) (The World Bank, 2019). This is attributed to its developed political and economic system, which enforces stability in the daily lives of US citizens. This means that it is worth exploring collaboration with governmental bodies in the formation of our solutions, since the political climate in the US is quite stable.

America is dominated by 2 political parties - the Republican Party and the Democratic Party. Depending on which political ideology the citizen is subscribed to, their sentiments towards the American government and general beliefs differ (Pew Research Center, 2020). Hence, it might be important to deliver our solution through a politically neutral channel (such as collaborating with neutral health organizations or private health centers), as political ideology affects receptiveness.

### 3.1.2 Economic factors

The US has the largest economy in the world. It is primarily powered by the tertiary industry, attributing 77.4% of the GDP to it, as of 2021 (Santander Trade, 2021). The unemployment rate in the US stands at 6.2% as of 2021 (U.S. Bureau of Labor Statistics, 2021). With most employees in the tertiary sector, solutions that target adults in office jobs may have the greatest outreach.

While the US operates in a mixed economy (Ross, 2020), the concept of the free market is applied to healthcare. Hence, healthcare in the US is not governed by one coherent body but taken care of by multiple systems. This means that while some groups of people are covered, such as US army veterans and employees of firms that provide healthcare insurance, a significant group of people are left vulnerable. In fact, 15% of US citizens aged 18-64 do not have insurance coverage (Duckett S, 2020). Their lack of insurance coverage may deter them from going for diabetes screenings, as they may have reservations against accumulating medical bills from expensive clinic visits. Hence, they may be more receptive to solutions that come at low or no cost to them.

### 3.1.3 Social factors

Due to the individualistic nature with which the government approaches healthcare, the same attitudes may apply to the citizens on the ground level. Hence, Americans may be more receptive to solutions that empower the individual to act.

Americans are fiercely protective of their freedom, as embodied by the 1st Amendment in the US Constitution. Hence, solutions that are invasive or disruptive to their current lifestyles may not be well received. This includes policies that enforce a change in the amount of sugar permitted in sugar beverages, for example.

### 3.1.4 Technological factors

Currently, 72.7% of Americans use a smartphone (Statista, 2018) and 89% of households in the US have access to a computer as of 2016 (Ryan, 2018). With about 90% of US citizens being active internet users (Statista, 2021), solutions that leverage the power of the internet can be effectively implemented, and its potential outreach would be significant.

## 3.2 Overview of Austin, Texas

As our dataset pertains to Austin, Texas, analyzing diabetes with respect to the community is important in supplementing our project with insights.

**Demographic details**

The following details are based on data collected in 2018.

|  |  |
| --- | --- |
| **Total population** | 964,243 residents |
| **Median age** | Residents in Austin are relatively young, at 33.6 years old. In comparison, the nation’s median age stands at 37.9 years old. However, the population is getting older, as Austin has increased from a median age of 33 years old in 2017. |
| **Median household income** | Households in Austin earn an income of USD 71,543 annually. In comparison, the median annual income of the United States is USD 61,937. |
| **Racial diversity** | White (non-Hispanic): 48.80%  White (Hispanic): 22.00%  Other (Hispanic): 8.21%  Black or African American (non-Hispanic): 8.13%  Asian (non-Hispanic): 7.59%  Multiracial (non-Hispanic): 2.43%  Multiracial (Hispanic): 1.25%  All others: 1.59%  Apart from non-Hispanic white people, Hispanic people make up a large portion of Austin, with 32.7% being Hispanic. |
| **Income inequality** | Males in Texas have an average income that is 1.37 times higher than the average income of females. Income inequality stands at 0.477 (Gini index), lower compared to the national average. |

Source of data: Data USA <https://datausa.io/profile/geo/austin-tx/>

# 4. Methods of Data Analysis

There are three key steps to our data analysis approach. They are: (1) Data Preparation and Cleaning, (2) Exploratory Analysis, and (3) Model Application.

After analyzing the data, we moved on to analysis of the best model to use.

## 4.1 Data Preparation & Cleaning

Data preparation, which includes data cleaning and visualization, is an important process to be carried out before the identification of key influencing factors and the development of models. The sample dataset chosen to perform the analysis on the diabetes status of the patient is the Austin Health Diabetes dataset. The dataset contains 1,689 records and 21 variables. The variable that is used to assess if the patient is diabetic is the Diabetes Status. Descriptions of the variables can be found in [Appendix 10.1](#_p9y2tdianmk).

### 4.1.1 Method of Data Cleaning

The dataset after cleaning is relatively balanced with about 56.8% of the sample not having diabetes and 43.2% with diabetes. Hence there is no need to balance the dataset using Synthetic Minority Over-sampling Technique (SMOTE) before conducting a train-test split. The details of the methods of cleaning are as below.

1. Removing Redundant Columns

Our team decided to remove the column named “Class” and “Problem Area in Diabetes (PAID) Scale Score”. This is because Class does not aid in our analysis and our study is with regards to the social economic factors that affect diabetes, whereas PAID is a 20-item representative self-reported instrument for measuring diabetes-related emotional distress, that covers a range of negative emotional problems of patients with diabetes (Lee, 2014). This will not be meaningful to our study as we are trying to predict whether patients have diabetes and PAID only applies to those who already have diabetes.

1. Standardizing Responses

Survey participants gave various responses with the same meaning. For example, when asked about “Food Measurement”, some respondents replied with “Not Sure” while others replied with “I don’t know how”. Hence, we standardized these responses across the different columns to simplify the dataset.

1. Sub-setting of Data to replace N.A.s

For columns Age and Race, N.A.s in the columns are replaced with the mode within the column. For the rest of the columns, we subset the data into the respective race ethnicities. Thereafter, split the N.A.s proportionately to the other responses within the column for the sub-set. For example, in the Asian subset, if there are 5 N.A.s in the “TobaccoUse” column, 6 “Yes” and 4 “No”, we would replace 2 of the N.As with “No” and 3 of the N.A.s with “Yes”.

## 4.2 Exploratory Data Analysis

We conducted some exploratory data analysis through visualization of the data that we are working with. The entire visualizations made are in [Appendix 10.2](#_crbhkh30l37g) and the highlights of which are shown in the table below.

|  |  |
| --- | --- |
| 1. Age Distribution against Diabetes | |
|  | * The people with diabetes in this dataset are generally older than those without diabetes. * People with diabetes have a higher median age, but a smaller IQR. |
| 1. Previous Diabetes Education against Diabetes | |
|  | * There is a higher proportion of people with diabetes who have had previous diabetes education, as compared to people who have not had previous diabetes education. |
| 1. Exercise against Diabetes | |
|  | * The proportion of people who have diabetes with differing days of exercise remains relatively consistent, less the population who exercise 5 or more days a week, which has a smaller proportion of people with diabetes. |

## 4.3 Model Building

In this section, we will use various modelling techniques to find out the most significant factors in the determination of Diabetic patients. Various models are built to find a model that best fits our dataset and returns results with a high accuracy rate. The models are logistic regression, CART, random forest, XGBoost and neural networks. Diabetes Status was used as an output, to determine the variables significant in determining if one has Diabetes or not. It is a categorical variable with 2 responses: “Yes” and “No”.

### 4.3.1 Logistic Regression

Logistic regression models find the probabilities for classification problems with two possible outcomes (Molnar, 2021).

1. Modelling Process

In predicting the probabilities of the outcomes, having, or not having diabetes, logistic regression is run twice, first with the entire dataset (log1) and second with only [significant variables](#_a7dmc3vbtuua) (log2). This is to mitigate overfitting. Only significant variables with p-value of 5% or less would be fitted in the second logistic regression model. This includes Age, Gender, Medical Home Category, Education Level, High Blood Pressure, Previous Diabetes Education, Sugar Beverage Consumption, Carbohydrate Count and Exercise. As Income was a derived column, it was natural that it would have a high multicollinearity score of more than 5, hence we removed it from the analysis.

1. Model Accuracy

To evaluate the model, a confusion matrix is performed on the train set first with a threshold of 0.5. This is to obtain the accuracy of the model to determine how well the model is trained using the train set. After which, the trained model is passed through the test set and another confusion matrix is produced to determine the accuracy of the trained model on the test set.

A confusion matrix ([Appendix 10.4](#_ur3r5p6bjvlj)) is used to evaluate the logistic regression model because it can measure sensitivity, specificity, false positive and negative rate, and accuracy. The accuracy of the trained logistic model would tell us how well the model is in predicting individuals with diabetes.

From the confusion matrix that was generated, the logistic regression model gave an accuracy of 0.71398 on the train set and 0.6951 on the test set. The model also returned a false negative rate of 0.2274 and a false positive rate of 0.4067.

1. Model Limitation

Logistic regression presents a few limitations to the analysis. First, logistic regression requires **moderate or no multicollinearity between independent variables**. This would mean no repetition of information can be used in the model as it would lead to wrong training of parameters while minimizing the cost function. Hence, when additional features are added to the dataset, there would be a manual process of finding multicollinearity and removing them before an analysis can be done. This would mean that logistic regression is not flexible with the addition of new data (Grover, n.d.).

Secondly, it is **difficult to capture complex relationships using logistic** regression due to the simplicity of the model. Complex and powerful algorithms such as neural networks can be used to circumvent this (Grover, n.d.).

Thirdly, the model has a higher propensity towards **overfitting on high dimensional data sets**. This problem is especially pronounced when there is little training data. Given that our data set has only 1368 records after cleaning but is evaluated across 19 X-variables, this issue is relevant to our context. Regularization techniques can be considered to overcome this, at the cost of increasing its complexity.

Finally, when conducting logistic regression, **only important and relevant features should be used to build the model**. Incorporating irrelevant features may train the model wrongly causing the accuracy to drop and its predictions made to be incorrect (Grover, n.d.).

### 4.3.2 Neural Network

1. Model Parameters

Our neural network model consists of 65 independent variables trying to predict 1 dependent variable, diabetes status. We derived 64 independent variables by creating dummy variables from biological and economic variables (categorical) we believe will have an impact on predicting one’s diabetes status. The last independent variable is the normalized participant’s age. It is normalized to speed up the learning process for faster convergence (Stöttner, 2019).

**Nominal variables:** tobacco use, heart disease, previous diabetes education, high blood pressure, gender, and race.

**Ordinal variables:** exercise, diabetes knowledge, education level, food measurement, carbohydrate counting, sugar beverage consumption, fruits vegetable consumption, insurance category, medical home category and income.

1. Modelling Process

We created a for-loop to determine the optimum number of hidden layers to use for our model. We tested models from layer (1) to layer (6,3). We adopted the 10-fold cross validation to train and test our model to ensure that our results are free from sample bias and check the robustness of our model (Basalamah, 2019). The models are evaluated based on a confusion matrix derived from cross-validated test set predictions to determine its accuracy, false negative and false positive.

1. Model Accuracy

The optimum neural network model is Layer (1,1) with an accuracy of 70.2% and false negative rate of 35.5%.

1. Model Limitations

Fluctuating model accuracy

Despite the set seed, we observed fluctuations in model accuracy. This is because neuralnet randomly assigns starting weights to the variables, independent of the seed value. We used 10-fold cross validation to create a more stable evaluation method. Producing models free from sample and initial weighting bias.

Time consuming training process

With only 1,368 data points, the training process takes approximately 18 hours to complete. Should the neural network be rolled out on a national scale, the dataset would consist of roughly 300 million data points, resulting in a significant increase in training time. The neural network model would have to be constantly trained to ensure that the latest SDOH and its influence on an individual’s diabetic risk is reflected in the model.

As a non-profit healthcare organization, ADA would not have the technological capacity to build and update the neural network models. It would also be very expensive to do so. Significant amount of time would be needed to train the initial model. It is also unlikely that the model can be kept updated with new data points.

Blackbox

Consensus amongst data science professionals shows that there is little visibility over the neural network’s learning process (Maroto, 2017). While we are aware that the model learns from stochastic gradient descent and backpropagation, there is still uncertainty over how the model learns. The black box impairs the understanding of how the number of hidden layers affects model accuracy and the rationale behind weightings assigned to input factors and nodes.

The lack of explainability would not inspire confidence in our end-users to use our model to predict their diabetic risk. Furthermore, without understanding how the neural network functions, ADA is unlikely to roll it out to the general public. Since the unexplainable false negative cases would have serious implications on a citizen’s life and result in backlash over the organization.

### CART

1. Model Description

The CART model is used to predict a categorical Y (Diabetes Status). We used the entire dataset with all the other X variables to grow a maximal tree, which we had to prune thereafter with an optimal complexity parameter value using the 1 SE rule to obtain the optimal tree. This will give us the final CART model to predict diabetes status.

1. Model Accuracy

To determine model accuracy, a train-test set was used with a split ratio of 0.7. The train set provided an accuracy of **0.6733** (4 d.p.) with the confusion matrix shown in *figure 4.1.* The test set provided an accuracy of **0.6732** (4 d.p.) with the confusion matrix shown in *figure 4.2*.

|  |  |
| --- | --- |
|  |  |
| Figure 4.1 Train Set Confusion Matrix | Figure 4.2 Test Set Confusion Matrix |

The closeness in accuracy between the 2 models shows that there was no over-fitting or under-fitting of the data, which implies that the train-test split was appropriate. However, with accuracies of about 0.673, the model would hardly be considered accurate, especially with a high false negative rate of 0.435, which carries harsher consequences compared to having more false positives in our case.

1. Model Limitations

Other than the CART model being inaccurate for the data we have, there are some other limitations to the CART model.

First, **the tree structure is unstable,** since small changes in the data set might result in a completely different tree being produced. Another disadvantage in using CART is that decision trees tend to overfit quickly at the bottom, which will cause poor decisions to be made if there are only a few observations in the last nodes (Yadav, 2019).

One of the solutions to these limitations is to use Random Forest, a more stable model that is an ensemble of many decision trees. This ensemble method also increases accuracy and prevents the data from being overfitted.

### 4.3.4 XGBoost

XGBoost, short for eXtreme Gradient Boosting, is an ensemble modelling method. It works sequentially, growing one tree after. The previous tree’s predictive power is improved upon when growing the next tree (by reducing the misclassification in the next iteration). This algorithm thereby converts weak learners to strong learners. XGBoost is used in this case, as our problem is a supervised learning problem that requires the prediction of a target categorical variable.

1. Modelling Process

The data set was split into a 70-30 train-test split. Then, the model is trained on its default parameters with the xg.train() function. The objective was set to “multi:softprob”, and the eval\_metric was set to “mlogloss”, as the predicted variable was categorical.

To improve the model, the following parameters were adjusted: nrounds, eta, max depth, subsample and colsample\_bytree. Using Bayesian Optimization for hyperparameter tuning, an appropriate set of hyperparameters were chosen to produce a model that was more optimal for the problem at hand.

The following limitations were imposed on the parameters, which were guided by earlier exploration of potential suitable model hyperparameters: *max\_depth* [2, 10], *min\_child\_weight* [1, 25], *subsample* [0.25, 1], *colsample\_bytree* [0.25, 1], *nrounds* [400, 900].

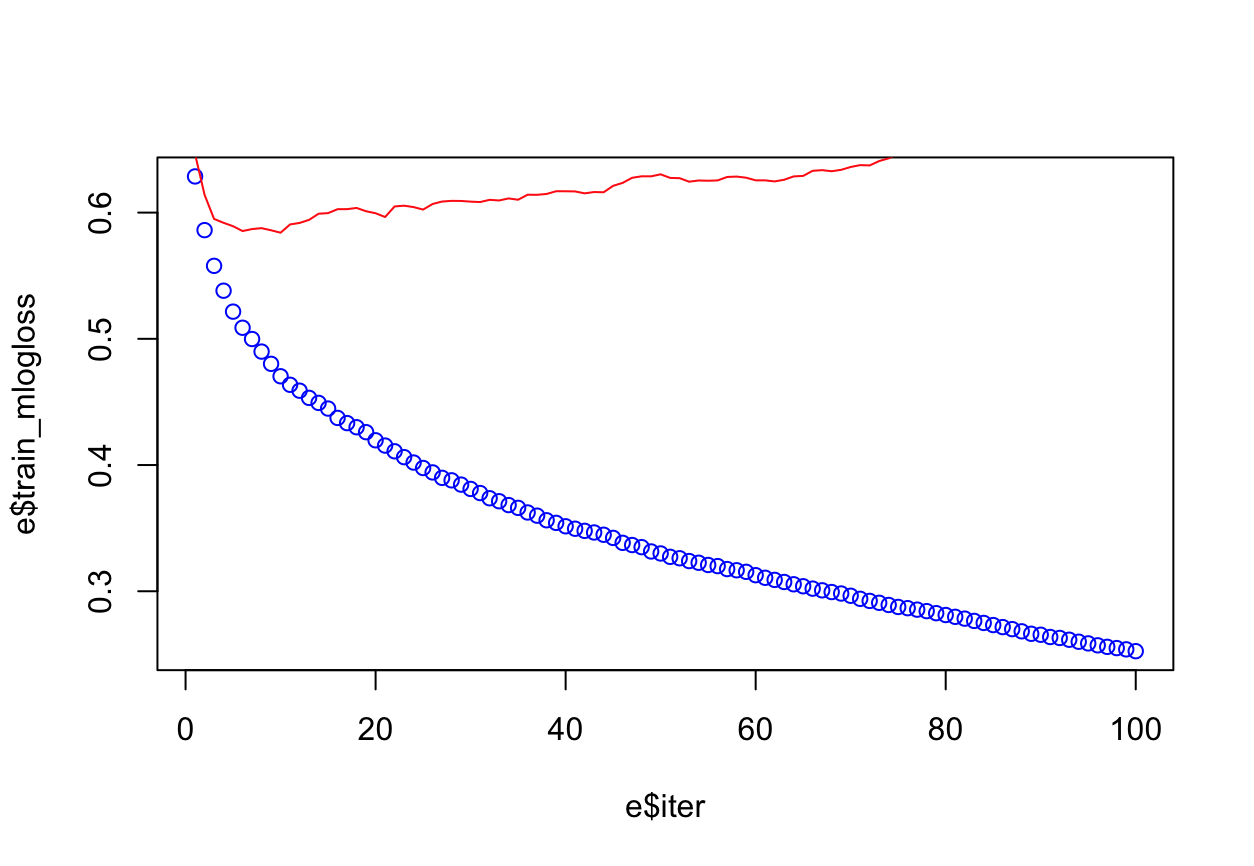
1. Model Results

With the default parameters ([Appendix 10.6](#_ob0b2xcnzlzj)), the model produced the following results (full results in [Appendix 10.7](#_69k9kfiwucvv)):

|  |  |  |
| --- | --- | --- |
|  | Train set Accuracy | Test set Accuracy |
| Overall accuracy | 0.9342 | 0.6683 |
| Sensitivity | 0.9632 | 0.6910 |
| Specificity | 0.8961 | 0.6384 |
| False Negative Rate | 0.1039 | 0.3616 |

While the train set accuracy is very high, the test set accuracy is much lower at 66.83%. Hence, the model does not predict well for unseen data with its current parameters and may be an overfit for its training data.

The model’s log-loss to the number of its iterations is plotted on a graph as seen below (figure 4.3). The red line represents the log-loss from the test set, and the blue plot represents the log-loss from the train-set. The greater the error in prediction, the higher the log-loss value. For the train set, the log-loss decreases as the number of iterations increase, but the test set does not reflect the same pattern, showing that the model does not generalize well on the test set.

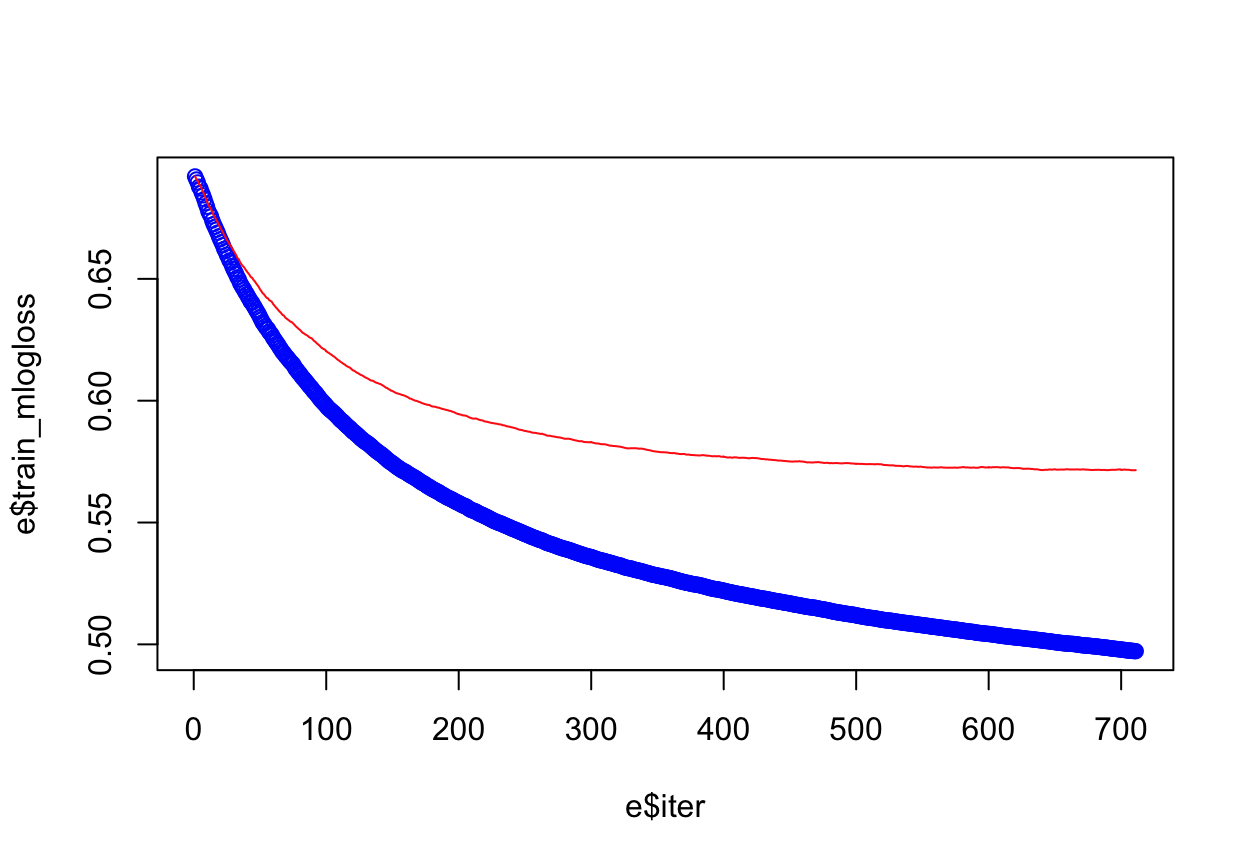


*Figure 4.3: Log-loss plot - under fitted*

1. Model Improvements

The Bayesian Optimization for hyperparameter tuning takes into account previous evaluation results, which is used in forming a probabilistic model “mapping hyperparameters to a probability of a score on the objective function” (Koehrsen, 2018). With the optimal parameters determined through Bayesian Optimization ([Appendix 9.8](#_kvy7uicffqug)), the model produced the following results (full results [Appendix 9.9](#_rfdj4tqp910g)):

|  |  |  |  |
| --- | --- | --- | --- |
|  | Train set Accuracy | Test set Accuracy | Improvement |
| Overall accuracy | 0.7704 | 0.7073 | 0.039 |
| Sensitivity | 0.8272 | 0.7511 | 0.0601 |
| Specificity | 0.6957 | 0.6497 | 0.0113 |
| False Negative Rate | 0.3043 | 0.3503 | 0.0113 |



*Figure 4.4: Log-loss plot after hyperparameter tuning*

Not only did the model’s performance improve on its validation set, but the model also generalizes much better on the test set compared to the first model (figure 4.4). This indicates that the model is better fitted after hyperparameter tuning.

1. Model Advantages

XGBoost is faster than other boosting methods and can handle large datasets. It is highly popular due to its tendency to be able to yield highly accurate results. This is because XGBoost makes use of systems optimization and algorithmic enhancements, such as:

* parallelized tree building
* efficient handling of missing data
* tree pruning using ‘depth-first’ approach
* built-in cross-validation capability (at each iteration)
* regularization through both LASSO (L1) and Ridge (L2) for avoiding overfitting

1. Model Limitations

While XGBoost is a popular model used in a wide variety of applications, it has its limitations. Firstly, it is bad at extrapolation. While this is not a significant issue for classification problems, it will affect its predictive power for regression problems, where the output is a continuous variable. By and large, when making predictions, XGBoost models (and tree models in general) cannot extrapolate beyond the limits of its training data (Mavuduru, 2020).

XGBoost models are also sensitive to outliers, as newer iterations are built based on fixing the errors of previous iterations. Hence, a data set with significant outliers may hinder the accuracy of the model (Corporate Finance Institute, n.d.).

### 4.3.5 Random Forest

The Random Forest is a supervised learning algorithm that builds multiple decision trees that are trained with a bagging method (Bootstrap Aggregation) which are then “merged together” to get a more accurate and stable prediction.

1. Modelling Process

We first ran the Random forest model against all other variables in the dataset. To optimize the model, we plot the error rates of the model with 25, 100, 500, 1000, 2000 trees with varying RSF values of 1, 4 and 20. 1, 4, 20 are chosen by a fixed formula, the results showed that the best parameters to use are 1000 trees with an mtry value of 4. Following this, to determine the model accuracy, a train test split of 70% train and 30% test was carried out.

1. Model Results

The train set provided an accuracy of **0.7296** (4 d.p.) with the confusion matrix shown in *figure 4.5.* The test set provided an accuracy of **0.6805** (4 d.p.) with the confusion matrix shown in *figure 4.6.*

|  |  |
| --- | --- |
|  |  |
| *Figure 4.5 Confusion Matrix for train Set* | *Figure 4.6 Confusion Matrix for test set* |

The test set accuracy is considerably low at 0.6805, the model would hardly be considered accurate. There is also a relatively high false negative rate of 17% and false positive rate of 15%.

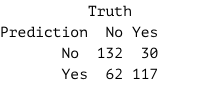
1. Model Improvements (Hyperparameters Tuning)

To improve the accuracy of the model, hyperparameter tuning was being carried out with the help of tidymodels metapackage. We first split the data into train and test sets using stratified sampling. Following that, we build a recipe for data preprocessing where (Julia Silge, 2020)

1. Step\_downsample: Downsample the slightly imbalanced dataset
2. Step\_dummy: convert nominal data to one or more numeric binary terms
3. Step\_corr: remove variables that have large absolute correlation with others

Next, we created a model specification for random forest where we tune the mtry (number of predictors to sample at each split) and min\_n (number of observations needed to keep splitting nodes), and the number of trees was set at 1000. A workflow is created to help us pre-process the data and tune it. 10-fold cross validation resamples were used for tuning with a grid of 20 grid points chosen each time.

After the initial tuning, it appears that lower mtry return high AUC values and higher min\_n likely returns higher AUC, though AUC fluctuates from throughout the range of 2 - 40. Using this information, we set the narrower range of hyperparameters we want to try, based on results from our initial tune. (mtry = range 1-20 and min\_n 2 - 40) and tune again. This returns the value of mtry = 5 and min\_n = 2. We then fit this final model on the entire training set and evaluate it on a test set which returns us an accuracy of 0.73% and an auc value of 0.800. With a false negative rate of 9% and false positive rate of 18%. There were significant improvements in accuracy, FNR and FPR.



1. Model Advantages
2. Ensures that the Models Diversify/uncorrelated from each other

Decision trees are very sensitive to the data that they are trained on where small changes to the train set can result in very different tree structure and accuracy. Random forest takes advantage of this by allowing each individual tree to randomly sample from the dataset with replacement; this process is known as bagging (Yiu, 2019).

1. Feature Randomness

In CART, the Gini coefficient is used to decide the best binary split. Whereas, in Random Forest different features are also used to make the splitting decisions (mtry). Each tree in a random forest can only be picked from a randomly selected subset of features variables are considered for each split. This helps in decorrelating the trees and generating different trees for use which ultimately results in lower correlation across trees and more diversification (Yiu, 2019).

1. Model Limitations
2. Computationally Intensive

Random forest creates a lot of trees and then combines their output. By default, it creates 500 trees in R. This will require much more computational power and resources than using a normal decision tree. Compared to a decision tree, random forest will require much more time to train as it generates a lot of trees instead of one tree and decides based on the majority votes. Currently, since our dataset is quite small, the time taken is not that huge. However, if a larger dataset were to be used, which it likely is, as ADA will have the statistics and dataset to further apply our models it might be very time consuming to train the models.

1. Black Box
2. Lack of explainability

Random Forest can also be considered a type of black box modelling (Stöttner, 2019). Though the idea behind Random Forest can be considerably simple for data scientists. However, to the regulators (ADA) who are untrained in data science, the complex algorithm can be very challenging to understand. Compared to Decision trees where a “real” tree can be printed out to visualize the model, however, the same cannot be done for random forests. The lack of explainability in interpreting this model (Walker, 2020) can be very challenging though easier than that of neural networks. Regulators (ADA) might not be confident of using the model when they are unable to fully understand the internal processes. If ADA regulators were to adopt these models, time would have to be taken out to learn the inner workings which might then come at an opportunity cost.

1. Unseen Problems that might impact the output.

With the inability to see the training process clearly, there could be unseen problems like over fitting, spurious correlates - which are impossible to catch due to the lack of understanding of the model underlying operation (Walker, 2020). Over time, this inability to comprehend the model might result in further inaccuracies.

## 4.4 Model Evaluation & Selection

### 4.4.1 Model Evaluation

Models are judged based on: Accuracy, False Negative Rate, Speed and Explainability.

Results of the various models are shown in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | False Negative Rate | Speed | Explainability |
| Logistic Regression | 69.5% | 22.7% | Fast | High |
| Neural Network | 70.2% | 35.5% | Slow | Low |
| CART | 67.3% | 43.5% | Fast | High |
| Random Forest | 73.0% | 19.0% | Medium | Low |
| XGBoost | 70.7% | 35.0% | Medium | Low |

### 4.4.2 Model Selection

Based on the table, the random forest model has the highest accuracy and lowest false negative rate. The random forest model takes a decent amount of time to be trained compared to logistic regression and CART. There is little explainability regarding how it selects variable importance and how these variables affect diabetic risk. However, the trade-off between speed and explainability for accuracy and false negative rate is well-justified given the significantly lower false negative rate of the random forest model.

False negative rate is an especially important metric for our case since a false negative prediction gives our users who actually have diabetes false assurance. Resulting in them not seeking medical attention and continuing their unhealthy lifestyle.

We shall deploy the Random Forest model as part of our diabetes-predicting solution to ADA.

# 5. Key Findings from data analysis

Random Forest is stochastic in nature; thus, we have set random seeds so that the results will be reproducible. To analyze random forest, we need to understand that it is an ensemble classifier that only makes sense of results from a multitude of decision trees. It would not make sense to combine n trees into one final tree and plotting the individual trees will not be meaningful. It is an ensemble classifier that “uses different samples of the same dataset in order to train multiple versions of the same model” (Miller & Forte, 2017). Hence, we will not be able to get a final decision tree like in CART.

However, we can analyze the model based on the model’s variable importance. We can identify the variables with the best predictive power (Diabetes Status). In our team’s implementation, the permutation method is used for determining the variable importance. This method measures the mean decrease in classification accuracy after if the randomly permuted X-variable is used instead of the original X-variable.

The variable importance ([Appendix 10.10](#_4q2jnaflnl9)) can be summarized into 3 categories, health factors, SDOH factors and demographic factors ([Appendix 10.11](#_t256p84wmjx)).

To further support our analytics model findings, research from other sources need to be analyzed how these factors positively or negatively impact one’s Diabetes Status. This is because variable importance does not tell us the relationship between the factor and the predicted Y. Thus, we do not know how an important variable influences the prediction. The variable importance measures the increase in error / reduction in accuracy of the prediction when its randomly permuted variable X is used instead of the original X. Hence, we can only conclude how much the factor influences the model’s accuracy.

The research findings below will focus on the factors that ADA can influence diabetes patients to change.

1. Previous Diabetes Education

The visualizations provided for in our exploratory data analysis suggests that there is a positive correlation between previous diabetes education and one’s diabetic risk. However, it is not intuitive to suggest that having previous diabetes education causes diabetes. Rather, the causation effect is the other way round. Firstly, diabetes patients are more likely to seek diabetes education to learn more about their illness and disease management methods. Secondly, Diabetes patients are usually educated on the diseases at diagnosis, as it helps with diabetes management. In fact, the ADA position statement maintains that all patients receive diabetes self-management education and support upon diagnosis and as needed thereafter (Powers et al, 2016).

1. Exercise Levels

Research by the ADA and American College of Sports medicine shows that increased exercise has a positive effect on diabetes risk. Increased physical activity enables muscle cells to use insulin and glucose more efficiently, thus lowers diabetic risk (Sheri et al, 2010). Physical activity is a key element in both the prevention and the management of type 2 diabetes. This [trend](bookmark://lkeb5yt8k999) is also observed during the exploration of the dataset.

# 6. Recommendation

### 6.1 Diabetes Detection – Early Diabetes Detection via Risk Assessment

According to Abid et al. (2016), preventative solutions to combat diabetes today are underpinned by early detection and prevention, with a focus on high-risk individuals. High risk individuals are identified via several risk factors. These include individuals who:

* Have a body mass index (BMI) exceeding 25kg/m^2
* Have a blood pressure of 135/80mmHg or above
* Have hypertension
* Do not regularly exercise
* Are 45 years old or older

However, despite these guidelines established by ADA, **millions of Americans are undiagnosed** (Vieira G, 2018). This means that a significant number of diabetics could have prevented their condition if there was early intervention.

This could be attributed to the fact that diabetes screening, such as fasting blood glucose tests, are optional, require taking some time off, and cost money. Many might choose not to go for screening, thinking they are “healthy enough” or simply do not want to know if they have a medical condition that they might need to spend money on. According to our PEST analysis, the estimated 15% of uninsured Americans between 18 to 64 may be hesitant to go for screening as well. Therefore, our solution aims to be **easily accessible** and **provides results immediately**.

Our team intends on crafting a survey that can be published by the ADA for anyone to take online. This survey acts as a risk assessment for diabetes.

### Design of Risk Assessment

Health indicators for predicting diabetes have long been cemented through the application of this knowledge in screening tests. Thus, our solution focuses on identifying and using **other non-health indicators** (while trying to rely less on these established health indicators, as some health data may not necessarily be conveniently accessible to survey-takers) to predict the patient's diabetes status to provide a more holistic picture of one’s diabetes risk status.

Depending on its variable importance, the assessment would rely on each indicator to different extents. For example, age with high variable importance would have a high weightage in predicting the possibility of diabetes. On the contrary, the race category would have a lower weightage in predicting the possibility of diabetes. This follows the methodology of variable importance using mean decrease in accuracy, where the factors that result in a greater error when randomly permuted would have greater importance.

### Implementation

The random forest model developed previously will be used to predict an individual’s diabetic risk. ADA can implement the predictive model accessible to the public via a webpage or mobile application. Users can then fill in their relevant health indicators, SDOH and demographic-based inputs into the webpage to obtain their diabetic risk (high or low). Based on the predicted risk and the risk factors they exhibit, ADA can provide more tailored medical advice for these individuals (e.g., Advising them to seek proper diagnosis in their local clinic, change their dietary or physical lifestyle, etc.).

The main aim of this self-help early diabetes detection tool is to strongly encourage individuals with high diabetic risk to get themselves properly diagnosed by medical professionals. Early detection allows them to take preventative measures before they develop type 2 diabetes, which is incurable.

Moreover, the findings of our variable importance analysis from random forest and exploratory data analysis shows that most people have diabetes education only after being diagnosed with diabetes. The random forest model also reveals that this is an important variable in predicting one’s diabetic risk. This suggests that the current education program is more reactive in nature, where individuals are more likely to seek such education after they are diagnosed with diabetes.

Therefore, apart from identifying individuals with high diabetic risk, the self-help early diabetes detection system can provide preventative education as well. After completing the risk assessment, individuals with low diabetic risk but display certain risk factors for diabetes can be given guidance on lifestyle changes to reduce diabetes risk. This will reduce the probability of these individuals getting type 2 diabetes as they grow older.

The effectiveness of this solution will be assessed via the total number of individuals who have taken the risk assessment, and the clickthrough rate of links to helpful resources that recommend lifestyle changes to reduce diabetes risk.

## 6.2 Diabetes Policies and Programs – Promoting Active Lifestyles

As mentioned in the key issue, current diabetes policies fail to consider SDOH factors (Hill, 2013). From our key findings, exercise is a key SDOH factor in diabetes detection. However, currently, the economic, social, cultural, and physical environment in the USA is not sufficiently supportive of behaviors that prevent obesity or type 2 diabetes (O'Brien & Kandula, 2016). According to the National Health Interview Survey, only about 23% of adults ages 18 to 64 are getting enough exercise (Ducharme, 2018).

When the broader environment is largely unsupportive, behavioral change requires individuals to be aware, motivated, and strongly supported (O'Brien & Kandula, 2016).

Thus, the main reason for lack of exercise among Americans points to the ***lack of incentive to exercise*** *and* ***age*** being a key deterrent.

## Motivating through incentives

Polls reveal that 35% of Americans do not exercise due to the lack of motivation (Schmall, 2019). Individuals face great inertia and resistance to kickstart the exercising process. We aim to remove this inertia by motivating individuals by providing incentives for them to kickstart their active lifestyle.

CDC only provides passive strategies to incentivize Americans to adopt a healthy lifestyle (Centre for Disease Control and Prevention, 2020). CDC should adopt more active strategies to increase its conversion rate, getting more individuals to take up an active lifestyle.

According to the PEST analysis, Americans perceive healthcare as an individual responsibility. Given the nature of healthcare in America, where there is no one coherent system and most services are privatized, it would be more effective to collaborate with private fitness centers or gyms to motivate individuals to exercise.

Through funding discount programs with these centers, Americans may be more motivated to exercise as facilities become more accessible than before. ADA can collaborate with fitness centers and gyms to motivate people to exercise. Participating citizens can visit certain gyms at discounted rates and earn points to obtain further discounts. ADA can model their new initiatives after programs deployed by other countries. An example would be the National Steps Challenge conducted by Singapore’s Health Promotion Board where participants were given discounts from participating merchants, most of whom have products and services relating to a healthy lifestyle. Success metrics include the number of sign ups and the number of people who show up to exercise at these fitness centers.

## Targeting vulnerable age groups

A report by New York Post finds that “two in five Americans feel too old to work out, with 41 officially being the age Americans feel too old to exercise” (Schmall, 2019). As age catches up, older adults are fearful of working out for fear of hurting themselves, especially if they have a pre-existing condition. Those who do not work out consistently also do not know where to start (Mastroianni, 2018).

US Governmental bodies such as the National Institute of Health have attempted to encourage the elderly to lead an active lifestyle through launching various campaigns. One such example is the Go4Life campaign, which focuses on fitness for older adults. The campaign made use of a website to create awareness and tips for exercising. However, the website is no longer functioning. Information such as posters, videos and articles about this campaign are now embedded deep in the health.gov website. This makes access to information for the elderly extremely difficult. There is no one central location for easy access of information. Overall, the lack of interactivity and difficult access to promotional materials meant that widespread awareness about exercising among the older adults is present (Go4Life, 2018).

### Solution for encouraging older adults to exercise:

#### One Stop Platform

Creating one coherent website that compiles all the elderly healthcare and exercises resources helps to ensure the elderly can easily access relevant resources for help. In this age of information overload, scattered resources, albeit of high quality, will not effectively reach its intended audience. This website aims to distill information about exercising for the elderly in America, acting as a one-stop portal for all the resources they might need.

This campaign will not be very costly, since it taps on existing high-quality resources, so the project is not built from the ground up. This method of a one stop platform has been tried and proven in other countries where it has successfully encouraged the older adults to exercise more often. One such example is the ActiveSG website implemented by the Singapore government, where Singaporeans can find all information required to exercise in one place.

The success metrics for this platform includes the website’s search engine optimization rankings based on keywords such as “Exercise for older people” and the number of visitors to the website per month.

#### Encouraging formation of Communities

Providing a platform for the elderly to form local exercise communities on the website will also help promote and potentially popularize exercising amongst the elderly. Similar age group participants can come together to exercise and share the activities in forum like pages - like creating a social network. Research has shown that community base social support can help to greatly increase exercise levels and encourage positive behavior change (Dailey, 2018). The more elderly persons who sign up to be part of these committees, the more effective this campaign will be.

# 7. Overall Limitation and Future Work

## 7.1 Uncertainty over economic factors

Our model approximates an individual’s income level based on their insurance status. This poses 2 limitations over our model.

Firstly, we are only able to identify low-income level individuals taking on welfare insurance schemes such as Medicaid. We were not able to identify middle and high-income earners and the effect economic risk has over their diabetic risk.

Secondly, we lack the exact income figures to provide a more holistic picture over an individual’s economic status. Exact numbers allow income to be modelled as a continuous variable to capture the nuances between low and extremely low-income levels or high and extremely high-income levels.

In the future, we hope to partner with other relevant agencies such as Inland Revenue Service (IRS). This will allow us to obtain numerical income levels to build a more robust model capturing the effects of economic status on diabetic risk.

## 7.2 Exercise intensity and duration

The project collects an individual’s self-reported number of days exercised in a week. We are not able to ascertain the duration and intensity for each exercise activity to calculate the number of calories burnt per session. Since the data is based on the self-reported figures, there are discrepancies between each user’s definition of exercise. The model is thus not able to accurately detect the relationship between an individual’s physical activities and their diabetic risk.

To provide an in-depth analysis to each individual’s exercise routine, with the user’s consensus, we aim to partner with companies that have lifestyle devices (e.g., Smartwatches) and/or smartphone applications (e.g., Run tracker) to monitor their physical activities. Data collected from these sources will significantly improve our understanding on the relationship between physical activity and diabetes risk. For users without these devices and applications, our survey can include more specific questions (e.g., average exercise duration, type of exercise). This will allow us to develop more accurate models relying on an individual’s physical lifestyle to assess their risk of contracting diabetes.

## 7.3 Carbohydrate and sugar consumption

Similar to exercise, an individual’s carbohydrate and sugar consumption is based on their self-reported figures on the number of times consumed per week. This results in inaccuracy in determining the amount of glucose they ingest per week.

The presence of excess glucose is the primary cause of diabetes. As such, carbohydrate and sugar beverage consumption are significant metrics in developing a model to predict an individual’s diabetes risk based on their dietary habits. We would want the data relating to total glucose consumption to be as precise as possible to build an accurate diabetes-predicting model.

To gain better clarity on an individual’s dietary habits, we can link up with calorie and diet tracking apps, where users watching their diet would log down their daily meals. If users’ consent, these applications can provide us with estimates of an individual’s glucose intake based on the amount and types of food they consume. For users who do not use such applications, our survey can include more specific questions (e.g., average number of meals per day, types of dishes, type of sugar beverage, size of sugar beverage). This would provide us with greater insight on their dietary habits to develop a more accurate model.

## 7.4 Limited Dataset

The dataset that we have used is centered around Austin geographically, and is considered relatively small compared to the entire US population. The lack of granular data may not provide an accurate picture on the exact lifestyle variables that are important to predict diabetes. This would cause some variance in terms of the accuracy of the model, especially if we compare it to running it with a larger dataset having the entire US population.

Furthermore, “Recent reports have also highlighted a gap in high-quality research to evaluate diabetes prevention policies.” (Konchak et al., 2016, 2). Should ADA decide to pursue this analytical route in their fight against diabetes, a larger data set should be used either by region, or by the entire US population. This will ensure that the models will have enough data points which should be able to train a model with higher accuracy.

# 8. Conclusion

The diabetes situation in the US, particularly Austin as focused on in our report, is highly complex. We believe that the potential for growth and usage of our recommendations and models can prove to be highly effective in the fight against Diabetes, especially with a larger dataset comprising a better composition which represents the entire population of the US better.

While the Risk Assessment and the one-stop platform and their key success metrics are relatively straight forward, the success metrics for incentive-based programs promoting exercise may be more difficult to measure, as collaboration is required with third party companies for data collection.

# 

# 9. Bibliography

Abid, A., Ahmad, S., & Waheed, A. (2016, June 13). Screening for Type II Diabetes Mellitus in the United States: The Present and the Future. *Clinical Medicine Insights: Endocrinology and Diabetes*, *9*, 19-22. 10.4137/CMED.S38247

American Diabetes Association. (n.d.). *The path to understanding diabetes starts here.* American Diabetes Association. https://www.diabetes.org/diabetes

American Diabetes Association. (2018). *Statistics About Diabetes*. American Diabetes Association. https://www.diabetes.org/resources/statistics/statistics-about-diabetes

Basalamah, S. (2019, January 10). *Cross Validation of an Artificial Neural Network*. Salsabila Basalamah. https://salsabilabasalamah.medium.com/cross-validation-of-an-artificial-neural-network-f72a879ea6d5

Centers for Disease Control and Prevention. (2018). *Health, United States, 2018 – Data Finder*. Centers for Disease Control and Prevention. https://www.cdc.gov/nchs/hus/contents2018.htm

Centers for Disease Control and Prevention. (2020). *Lifestyle Change Program Details*. Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/prevention/lcp-details.html

Centers for Disease Control and Prevention. (2020). *National Diabetes Statistics Report 2020 Estimates of Diabetes and Its Burden in the United States*. Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf

Centre for Disease Control and Prevention. (2020, June 9). *Strategies to Increase Physical Activity*. CDC. https://www.cdc.gov/physicalactivity/activepeoplehealthynation/strategies-to-increase-physical-activity/index.html

Corporate Finance Institute. (n.d.). *Boosting*. Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/knowledge/other/boosting/#:~:text=One%20disadvantage%20of%20boosting%20is,almost%20impossible%20to%20scale%20up

Dailey, S. (2018, September 5). *Celebrate Go4Life Month in September*. health.gov. https://health.gov/news/healthgov-blog/201809/celebrate-go4life-month-september

Ducharme, J. (2018, June 28). *Only 23% of Americans Get Enough Exercise, a New Report Says*. Times. https://time.com/5324940/americans-exercise-physical-activity-guidelines/#:~:text=Only%2023%25%20of%20Americans%20Get%20Enough%20Exercise%2C%20a%20New%20Report%20Says,-Share

Duckett, S. (2020, November 19). *More expensive but less effective: the US healthcare system explained*. Grattan Institute. https://grattan.edu.au/news/more-expensive-but-less-effective-the-us-healthcare-system-explained/#:~:text=The%20United%20States%20does%20not,multiple%20systems%2C%20with%20no%20coherence.&text=If%20you%20are%20over%2065,Medicaid%20system%20is%20your%20opti

Go4Life. (2018, March 26). *Community-based social support for physical activity*. https://health.gov/news/healthgov-blog/201809/celebrate-go4life-month-september

Grover, K. (n.d.). *Advantages and Disadvantages of Logistic Regression*. OpenGenus Foundation. https://iq.opengenus.org/advantages-and-disadvantages-of-logistic-regression/

Hill, J. (2013). Understanding the Social Factors That Contribute to Diabetes: A Means to Informing Health Care and Social Policies for the Chronically Ill. *The Permanente Journal*, *17*(2), 67-72. 10.7812/TPP/12-099

Julia Silge. (2020, 03 26). *Tuning random forest hyperparameters with #TidyTuesday trees data*. https://juliasilge.com/blog/sf-trees-random-tuning/

Koehrsen, W. (2018, June 24). *A Conceptual Explanation of Bayesian Hyperparameter Optimization for Machine Learning*. Towards Data Science. https://towardsdatascience.com/a-conceptual-explanation-of-bayesian-model-based-hyperparameter-optimization-for-machine-learning-b8172278050f

Konchak, J. N., Moran, M. R., O'Brien, M. J., Kandula, N., & Ackermann, R. T. (2016, May). The State of Diabetes Prevention Policy in the USA Following the Affordable Care Act. *Current Diabetes Reports*. https://www.researchgate.net/deref/https%3A%2F%2Flink.springer.com%2Farticle%2F10.1007%2Fs11892-016-0742-6

Lee, Y. W. L. K.-W. L. Y. S. K. a. M.-S. N.-H. L. (2014, 10 29). *Measurement of diabetes-related emotional distress using the Problem Areas in Diabetes scale: psychometric evaluations show that the short form is better than the full form*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4215007/#:~:text=The%20Problem%20Areas%20in%20Diabetes%20(PAID)%20scale%20is%20a%2020,USA%20for%20patients%20with%20diabetes.

Maroto, M. (2017, April 21). *Neural Networks = Black Box?* Towards Data Science. https://towardsdatascience.com/neural-networks-black-box-b20723f9a417

Mastroianni, B. (2018, December 10). *For Aging Bodies, Endurance and Resistance Training Can Help*. Healthline. https://www.healthline.com/health-news/endurance-not-resistance-training-helps-with-aging

Mavuduru, A. (2020, November 10). *Why XGBoost can’t solve all your problems.* Towards Data Science. https://towardsdatascience.com/why-xgboost-cant-solve-all-your-problems-b5003a62d12a

Miller, J. D., & Forte, R. M. (2017). *Mastering Predictive Analytics with R* (2nd ed.). Packt Publishing.

Molnar, C. (2021, March 29). *A Guide for Making Black Box Models Explainable.* Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book/logistic.html

O'Brien, M. J., & Kandula, N. (2016, May). The State of Diabetes Prevention Policy in the USA Following the Affordable Care Act. *Current Diabetes Reports*, *16*(6), 3-5. https://www.researchgate.net/publication/301945167\_The\_State\_of\_Diabetes\_Prevention\_Policy\_in\_the\_USA\_Following\_the\_Affordable\_Care\_Act

Office of Disease Prevention and Health Promotion. (2020). *Social Determinants of Health*. U.S. Department of Health and Human Services. https://health.gov/healthypeople/objectives-and-data/social-determinants-health

Pew Research Center. (2020, September 14). *Americans’ Views of Government: Low Trust, but Some Positive Performance Ratings*. U.S. Politics & Policy. https://www.pewresearch.org/politics/2020/09/14/americans-views-of-government-low-trust-but-some-positive-performance-ratings/

Powers M. et. al, P. (2016, 04). *Diabetes Self-management Education and Support in Type 2 Diabetes: A Joint Position Statement of the American Diabetes Association, the American Association of Diabetes Educators, and the Academy of Nutrition and Dietetics*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4833481/

Ross, S. (2020, October 13). *Is the United States a Market Economy or a Mixed Economy?* Investopedia. https://www.investopedia.com/ask/answers/031815/united-states-considered-market-economy-or-mixed-economy.asp

Ryan, C. (2018, August). *Computer and Internet Use in the United States: 2016*. American Community Survey Reports. https://www.census.gov/content/dam/Census/library/publications/2018/acs/ACS-39.pdf

Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., Colagiuri, S., Guariguata, L., Motala, A. A., Ogurtsova, K., Shaw, J. E., Bright, D., & Williams, R. (2019, November). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Research and Clinical Practice*, *157*. https://doi.org/10.1016/j.diabres.2019.107843

Santander Trade. (2021). *United States: Economic and Political Outline*. Santander Trade Markets. https://santandertrade.com/en/portal/analyse-markets/united-states/economic-political-outline

Schmall, T. (2019, Januray 13). *This is why most Americans don’t exercise more*. New York Post. https://nypost.com/2019/01/13/this-is-why-most-americans-dont-exercise-more/

Sheri R. er.al. (2010, 12). *Exercise and Type 2 Diabetes*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2992225/

Statista. (2018, May). *Smartphone penetration rate as share of the population in the United States from 2010 to 2021*. Statista. https://www.statista.com/statistics/201183/forecast-of-smartphone-penetration-in-the-us/

Statista. (2021, February). *Digital population in the United States as of January 2021*. Statista. https://www.statista.com/statistics/1044012/usa-digital-platform-audience/

Stöttner, T. (2019, May 16). *Why Data should be Normalized before Training a Neural Network*. Towards Data Science. https://towardsdatascience.com/why-data-should-be-normalized-before-training-a-neural-network-c626b7f66c7d

Tunceli, K., Bradley, C. J., Nerenz, D., Williams, L. K., Pladevall, M., & Lafata, J. E. (2005, November 28). The Impact of Diabetes on Employment and Work Productivity. *Diabetes Care*, 6. https://doi.org/10.2337/diacare.28.11.2662

U.S. Bureau of Labor Statistics. (2021, March 26). *State Employment and Unemployment Summary*. U.S. Department of Labor. https://www.bls.gov/news.release/laus.nr0.htm

U.S Department of Health & Human Services. (2004). *Diabetes: A National Plan for Action. The Importance of Early Diabetes Detection*. https://aspe.hhs.gov/report/diabetes-national-plan-action/importance-early-diabetes-detection#:~:text=Early%20detection%20and%20treatment%20of,limb%20amputations%2C%20and%20kidney%20failure

Vieira, G. (2018, September 26). *Are You 1 of the 7 Million Americans Who Don't Know They Have Diabetes?* Healthline. https://www.healthline.com/health-news/are-you-one-of-the-8-million-americans-who-doesnt-know-they-have-diabetes

Walker, C. (2020, Jan 31). *White Box vs Black Box Models: Balancing Interpretability and Accuracy*. https://blog.dataiku.com/white-box-vs-black-box-models-balancing-interpretability-and-accuracy

The World Bank. (2019). *Worldwide Governance Indicators*. Databank. https://databank.worldbank.org/source/worldwide-governance-indicators

Yadav, A. (2019, January 11). *DECISION TREES*. Towards Data Science. https://towardsdatascience.com/decision-trees-d07e0f420175

Yiu, T. (2019, Jun 12). *Understanding Random Forest*. https://towardsdatascience.com/understanding-random-forest-58381e0602d2

# 

# 10. Appendix

## 10.1 Data Dictionary

|  |  |
| --- | --- |
| Columns | Meaning |
| Class | Class type (APH = Austin Public Health; ARCF = Abundant Rain Christian Fellowship; EB= El Buen Samaritano; PCHW = Promotores Community Health Workers) |
| ClassLanguage | Language class was taught in |
| Age | Age of participant |
| Year | Year of class |
| Gender | Gender of participant |
| Insurance Category | Insurance type of participant |
| Medal Home Category | Medical home of participant |
| Race/Ethnicity | Race/ethnicity of participant |
| Education Level | Education Level of participant |
| Diabetes Status (Yes/No) | Diabetes Status (Yes/No) of participant |
| Heart Disease (Yes/No) | Heart disease diagnosis (yes/no) |
| High Blood Pressure (Yes/No) | High blood pressure diagnosis (yes/no) |
| Tobacco Use (Yes/No) | Tobacco user (yes/no) |
| Previous Diabetes Education (Yes/No) | Previous diabetes education reported by participant (yes/no) |
| Diabetes Knowledge | Self-reported knowledge of diabetes (poor/fair/good) |
| Fruits & Vegetable Consumption | Fruits & Vegetable Consumption eaten each week |
| Sugar-Sweetened Beverage Consumption | Sugar-Sweetened Beverage Consumption each week |
| Food Measurement | Number of time food was measured each week |
| Carbohydrate Counting | Number of times carbohydrates were counted each week |
| Exercise | Number of days participant exercised each week |
| Problem Area in Diabetes (PAID) Scale Score | The PAID score is a measure of difficulty in managing one's diabetes. It ranges from 0-100, with higher scores indicating more problems. |

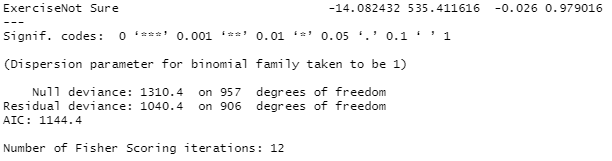
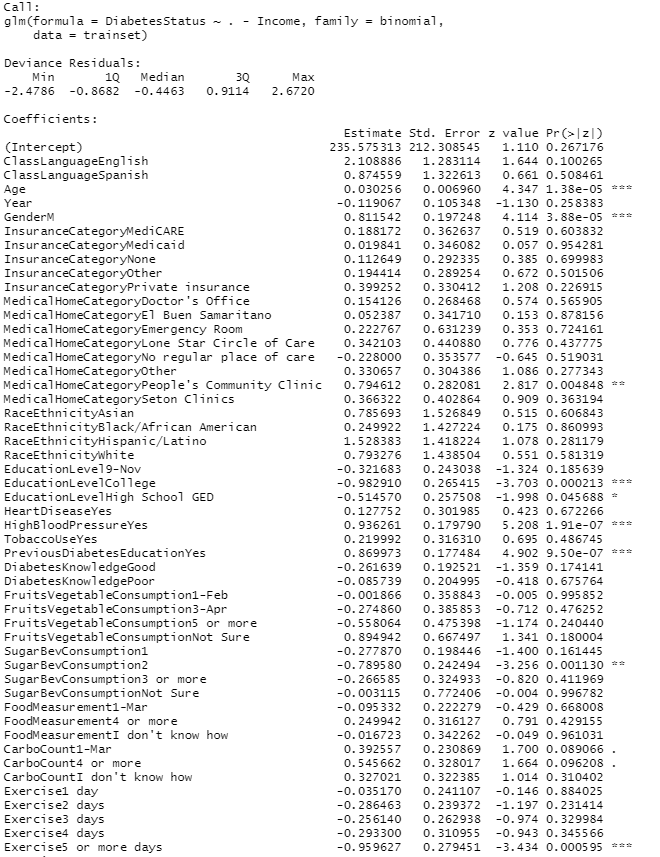
## 

## 10.2 Data Visualisations

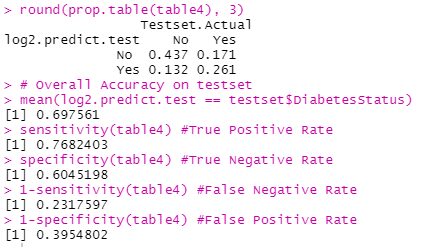
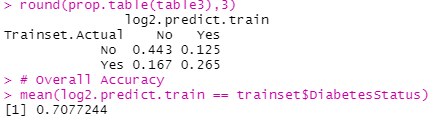
|  |  |
| --- | --- |
| 1. Age Distribution against Diabetes | |
|  | * The people with diabetes in this dataset are generally older than those without diabetes. * People with diabetes have a higher median age, but a smaller IQR |
| 1. Age Distribution against Insurance Category | |
|  | * The population with no medical insurance seems to have a lower median compared to all the other categories |
| 1. High Blood Pressure against Diabetes | |
|  | * There is a higher proportion of people with diabetes who also have high blood pressure, as compared to people without high blood pressure |
| 1. Previous Diabetes Education against Diabetes | |
|  | * There is a higher proportion of people with diabetes who have had previous diabetes education, as compared to people who have not had previous diabetes education |
| 1. Gender against Diabetes | |
|  | * There is a larger proportion of males with diabetes compared to females with diabetes |
| 1. Class Language against Diabetes | |
|  | * In terms of proportion, the English speakers have a higher concentration of people with diabetes. * There is a small sample size of Chinese/English speakers in this dataset |
| 1. Exercise against Diabetes | |
|  | * The proportion of people who have diabetes with differing days of exercise remains relatively consistent, less the population who exercise 5 or more days a week, which has a smaller proportion of people with diabetes |
| 1. Race Ethnicity against Diabetes | |
|  | * There is a larger population of Hispanic/Latino in the dataset, however, they are the only sub-population with a smaller number of people with diabetes |
| 1. Education Level against Diabetes | |
|  | * The sub-population of people with High School GED have the highest proportion of people with diabetes, whereas the sub-population of people who have been to college has the smallest proportion |
| 1. Insurance Category against Diabetes | |
|  | * Younger demographic does not have an insurance plan * Explaining the smaller proportion of people without insurance have no diabetes |
| 1. Carbohydrate Count against Diabetes | |
|  | * The more an individual tracks their carbohydrate intake, the more likely they have diabetes |
| 1. Sugar Beverage Consumption against Diabetes | |
|  | * The more an individual consumes sugar beverage, the more likely they have diabetes |

## 

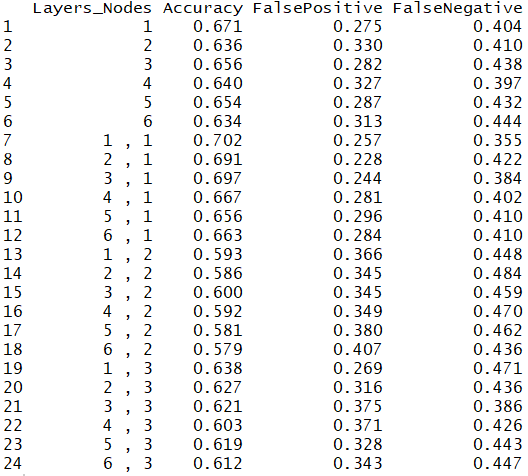
## 10.3 Logistic Regression Significant Variables



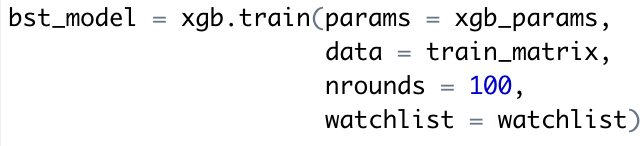
## 10.4 Logistic Regression Results



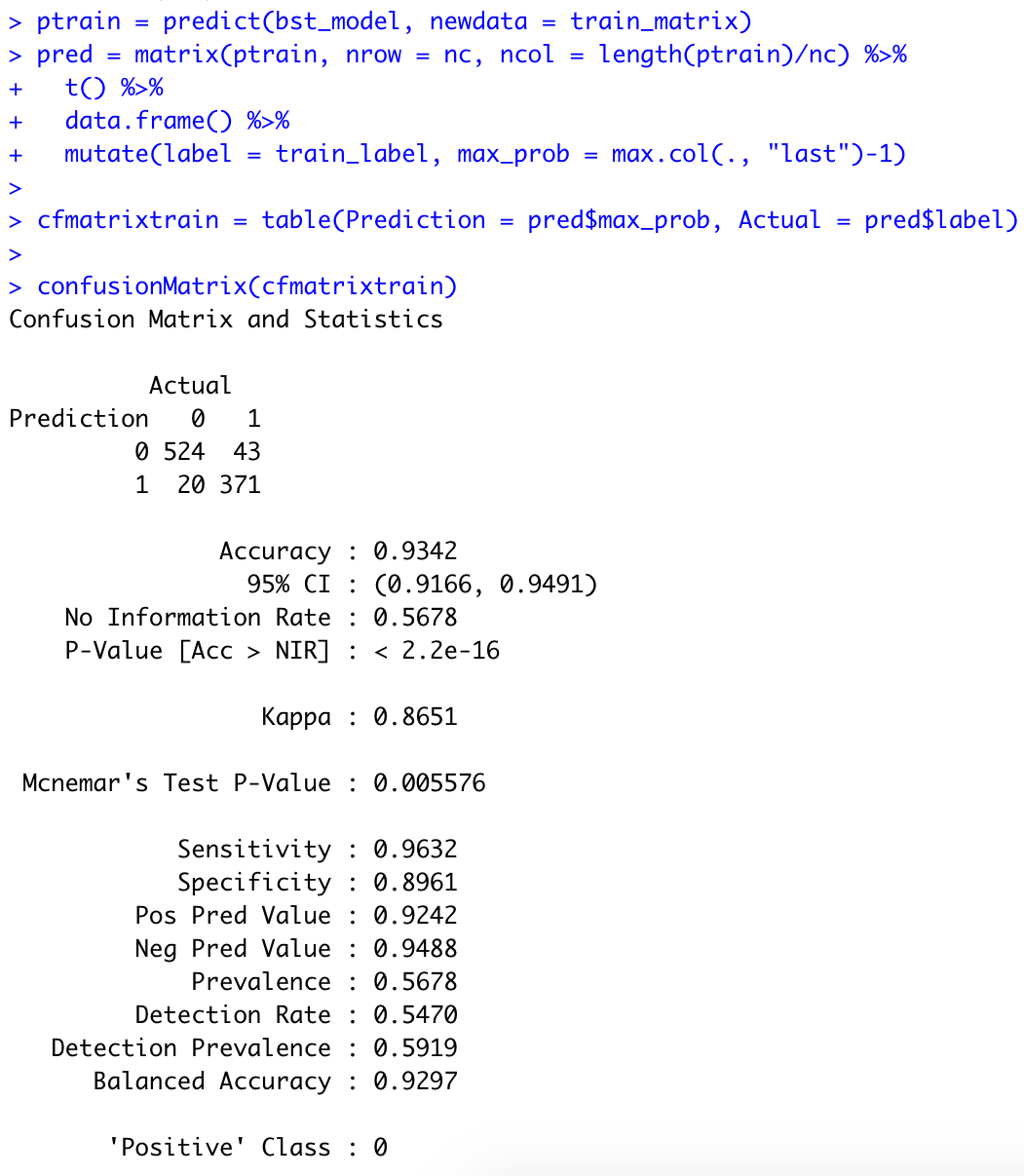
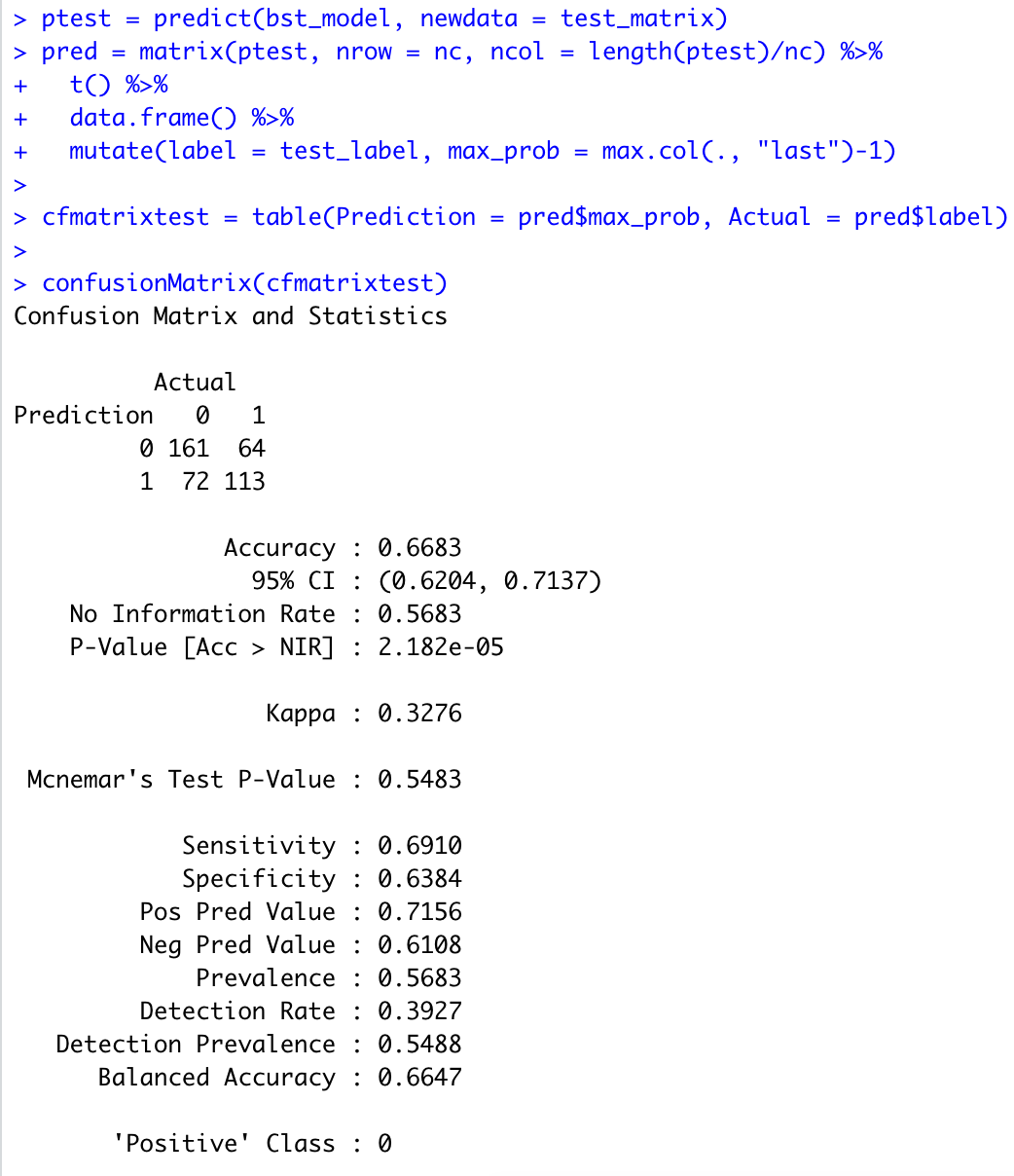
## 10.5 Neural Network Results



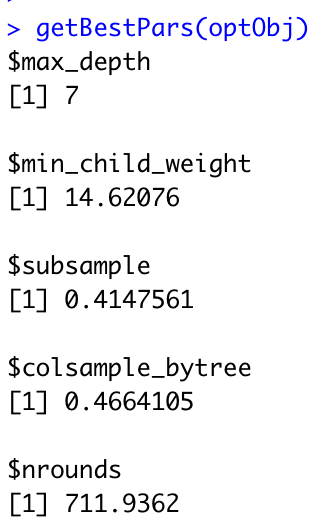
## 10.6 Default Parameters of XGBoost Model



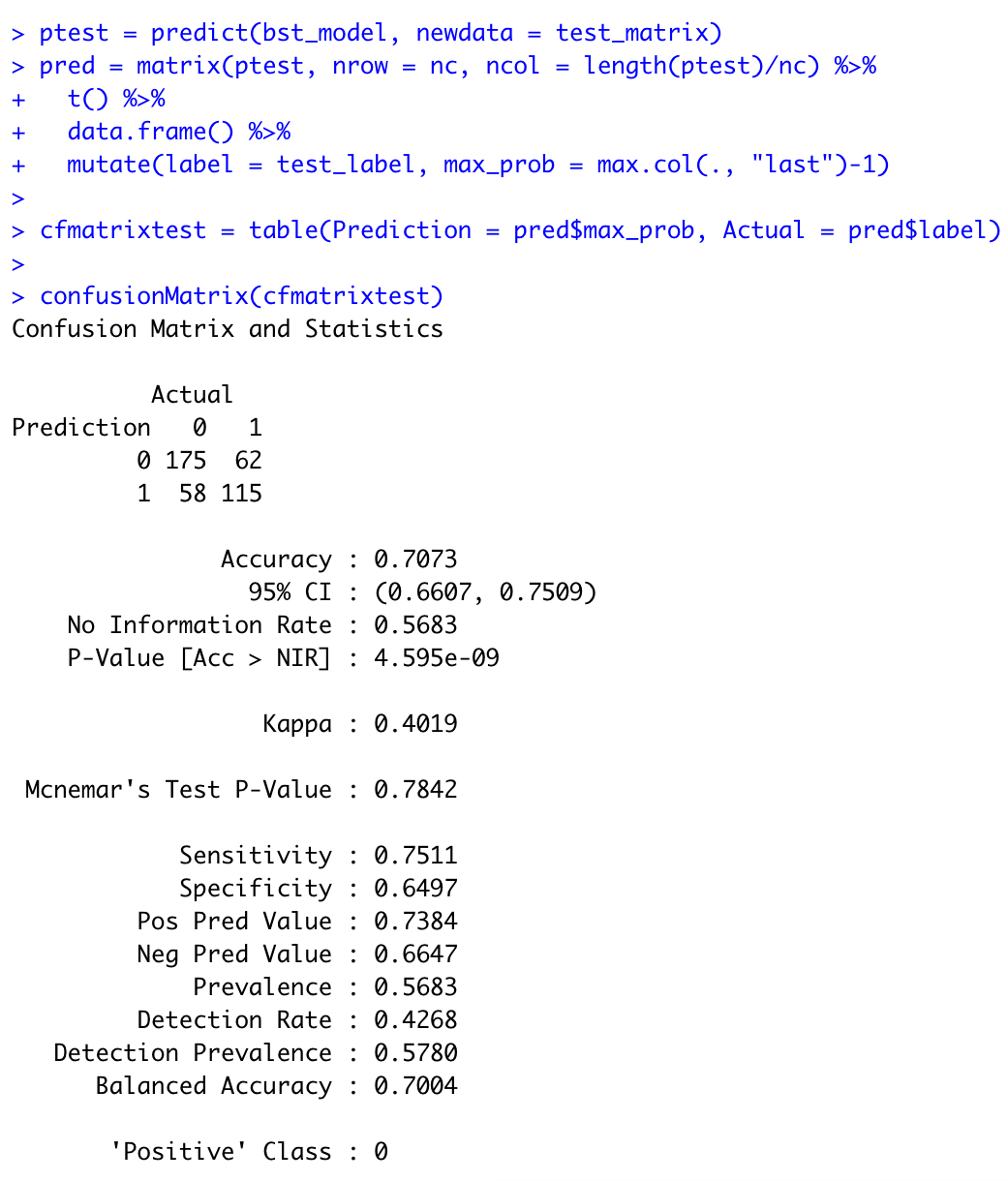
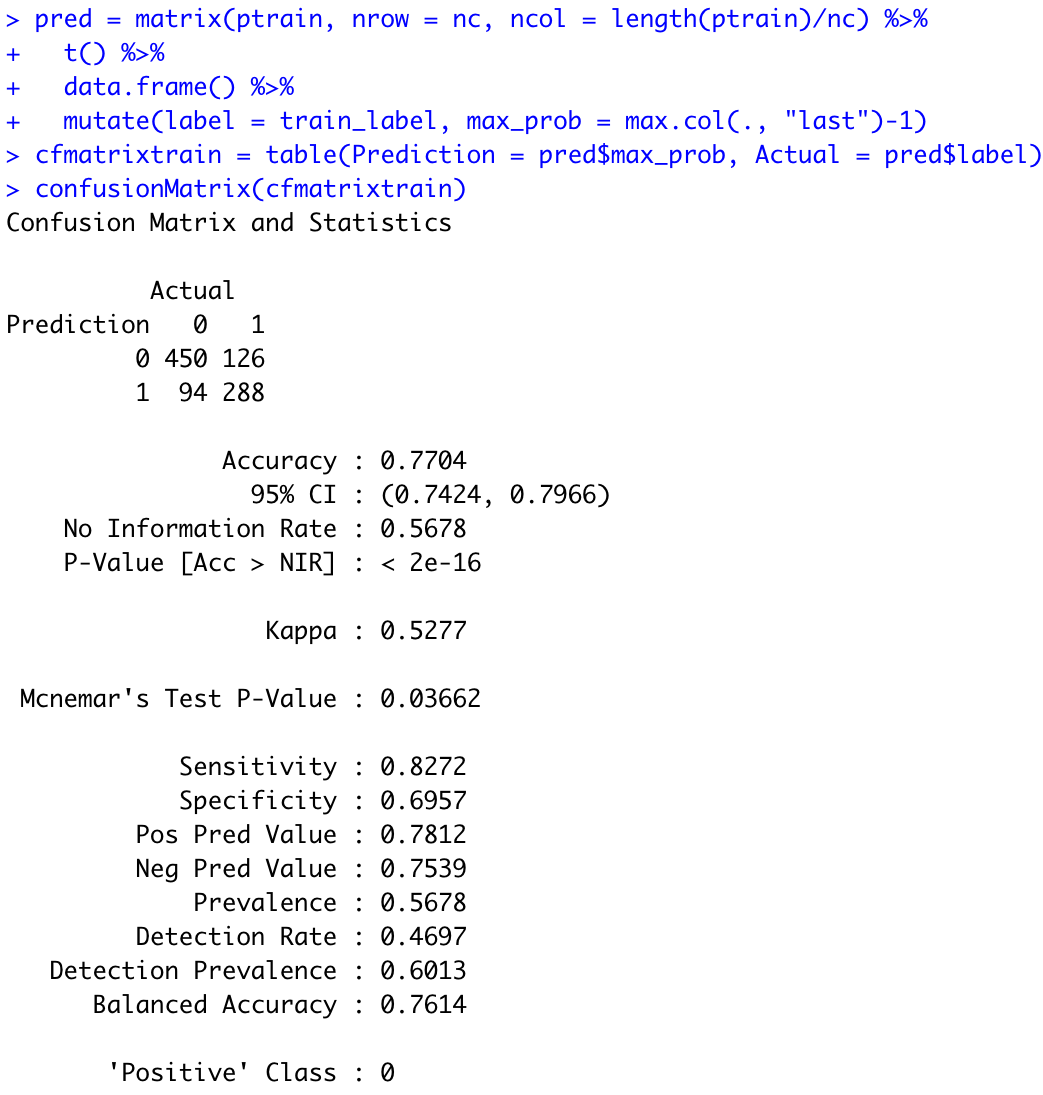
## 10.7 XGBoost Results without Hyperparameter Tuning



## 10.8 Optimal Hyperparameters used in XGBoost.



## 10.9 XGBoost Results after Hyperparameter Tuning



## 10.10 Random Forest Most Important Variables

|  |
| --- |
|  |
|  |

## 10.11 Variable Importance Factors Category

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Importance Factors**  *Top to bottom (most important to less important)* | **Health  Factors** | **SDOH Factors**  *(i.e Social, Economic, Environment)* | **Demographic Factors** |
| **Age** |  |  |  |
| **High Blood Pressure (Yes)** |  |  |  |
| **Previous Diabetes Education (Yes)** |  |  |  |
| **Gender (Male)** |  |  |  |
| **ClassLanguage (English)** |  |  |  |
| **Exercise (5 or more days)** |  |  |  |
| **Race Ethnicity (Hispanic Latino)** |  |  |  |
| **Medical Home Category (Doctor’s office)** |  |  |  |
| **Education Level (College)** |  |  |  |
| **Race Ethnicity (Black African American)** |  |  |  |