Type 2 Diabetes in Austin, Texas

Presented by:
Ng Chi Hui (U1922243C)
Lim Wei Jie Spencer (U1810413H)
Natasha Khoo Mei Hui (U1910526E)
Jarrel Kong (U1910901J)
Wong Kang Xun (U1810804L)

TABLE OF CONTENTS

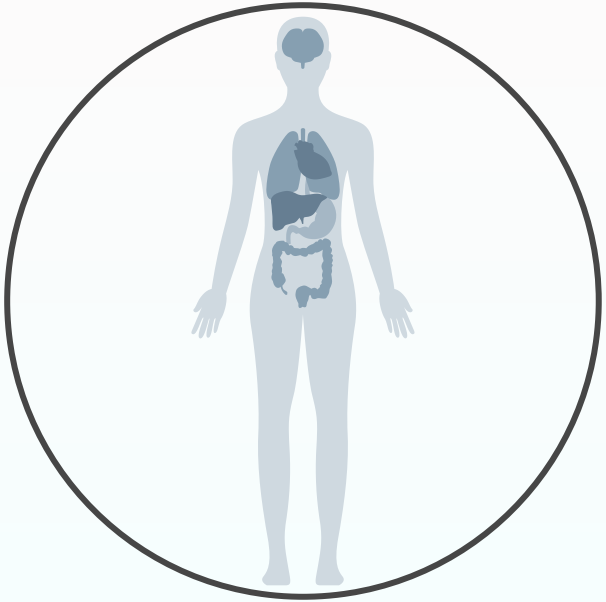
Jarrel

01

Introduction

Problem Statement

How can we improve the diabetes (type 2) situation in the US?



Background: Why Type 2 Diabetes?

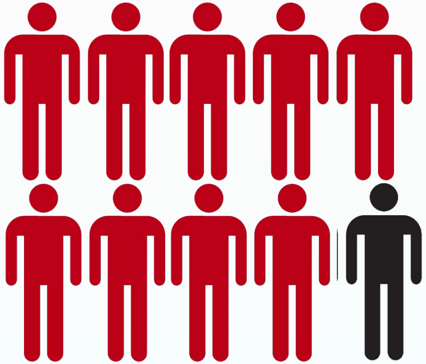
90-95%

of all diabetes are caused by Type 2 diabetes

Background: Type 2 Diabetes Mellitus

Global Pandemic

- Half a billion people worldwide are living with diabetes
- Projected to increase to an astonishing 25% in 2030 and 51% in 2045



Diabetes in U.S.

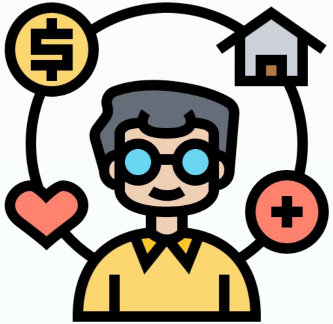
- National Diabetes Statistics Report that 9 in 10 diabetes are caused by type 2 diabetes

Jarrel

Background: Effect of Diabetes

Individual

- Lower Quality of Life



Societal

- Overall decreased productivity for the society

Situational Analysis - Climate in the US

Political
- 0.3 on political stability index
- Dominated by 2 political party

Economic
- Largest Economy in the world
- Healthcare - adopts concept of free market

PEST
Analysis

Social
- Individualistic nature
- Fiercely protective of their freedom

Technological
- 72.7% of Americans use a smartphone
- 89% of household have access to computer

Jarrel

Why does US have a **HIGH** Diabetes Rate despite having an Advance Healthcare System?

Jarrel

Key Issues

Key Issues

Diabetes Detection

Purely Medical/Lab tests

Diabetes Policies

SDOH factors not a key component

SDOH factors needs to be a sufficiently considered before the Diabetes Situation in US can be improved,

02

Data Pipeline

Step 1: Data Cleaning

Remove Redundant Columns

Standardize Response

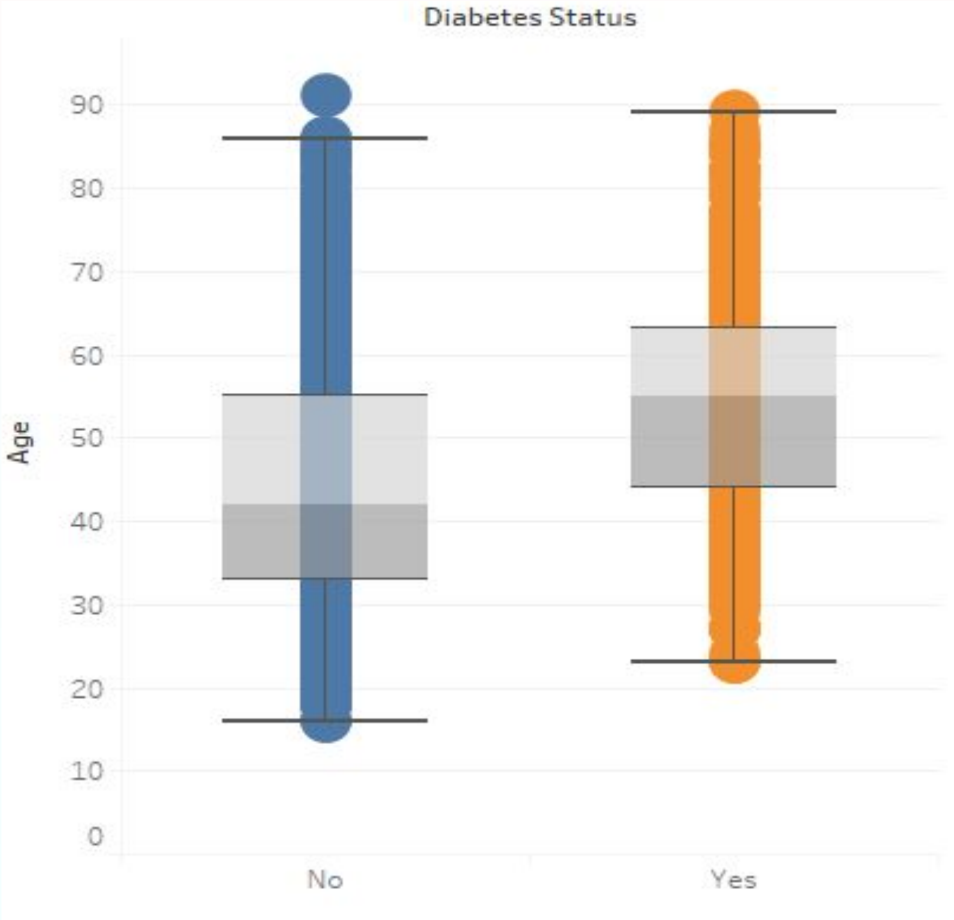
Subsetting of Data to Replace N.As

Spencer

Step 2: Data Visualization

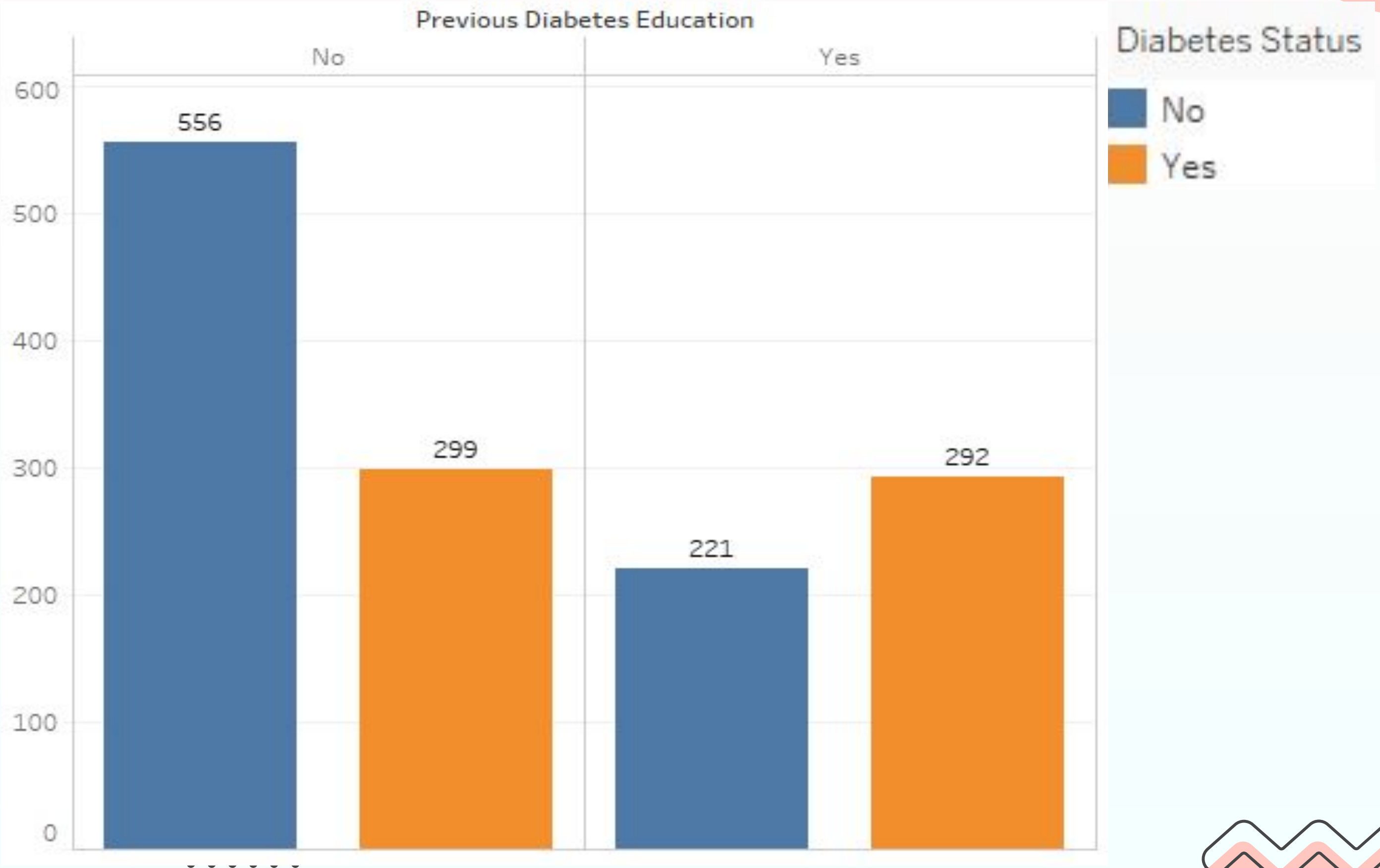
Spencer

Age Distribution against Diabetes



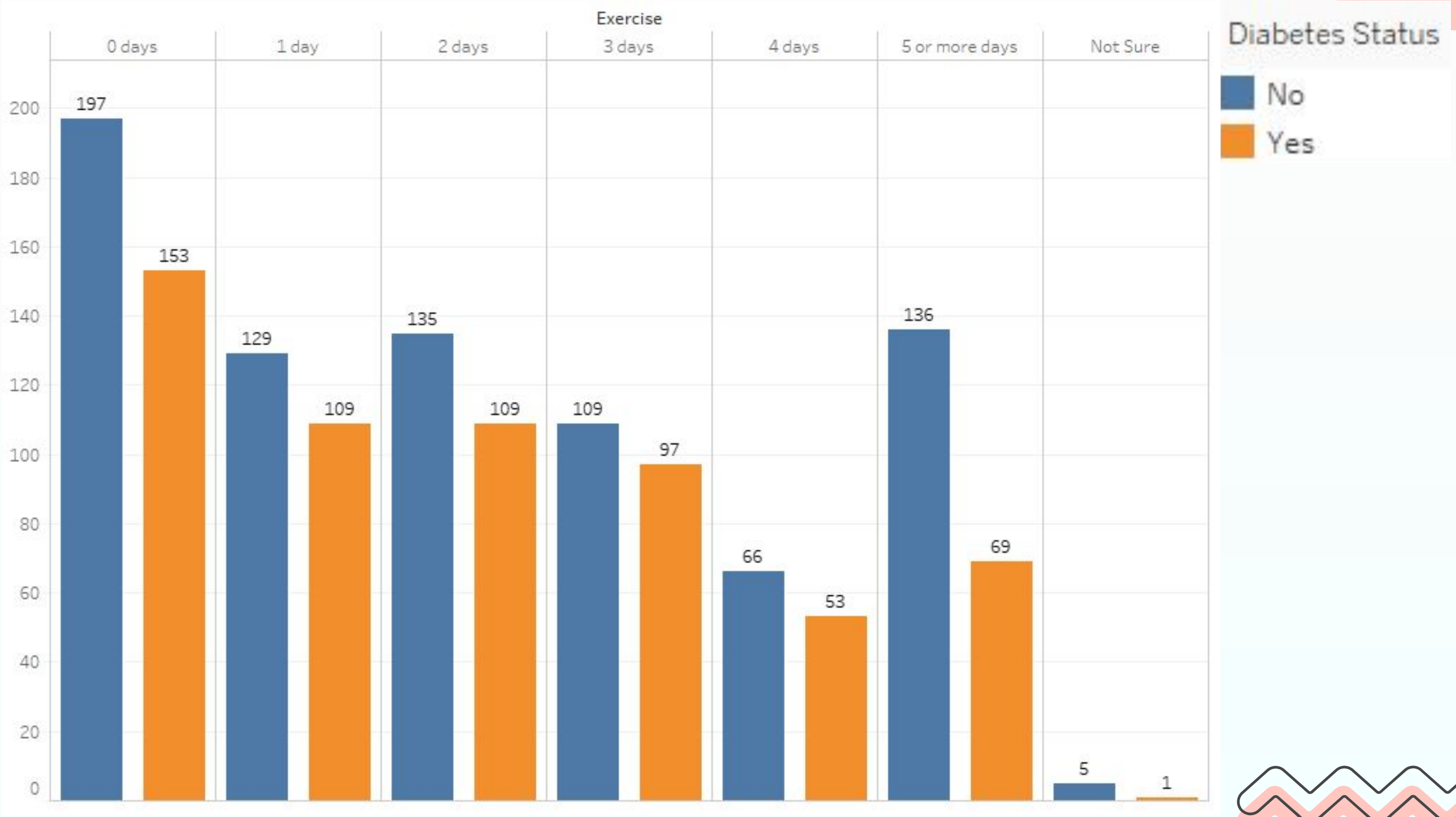
Spencer

Previous Diabetes Education against Diabetes



Spencer

Exercise against Diabetes



Spencer

Step 3: Model Building

Model Classes

Ensemble Methods

1) CART
2) Random Forest
3) XGBoost

Regression

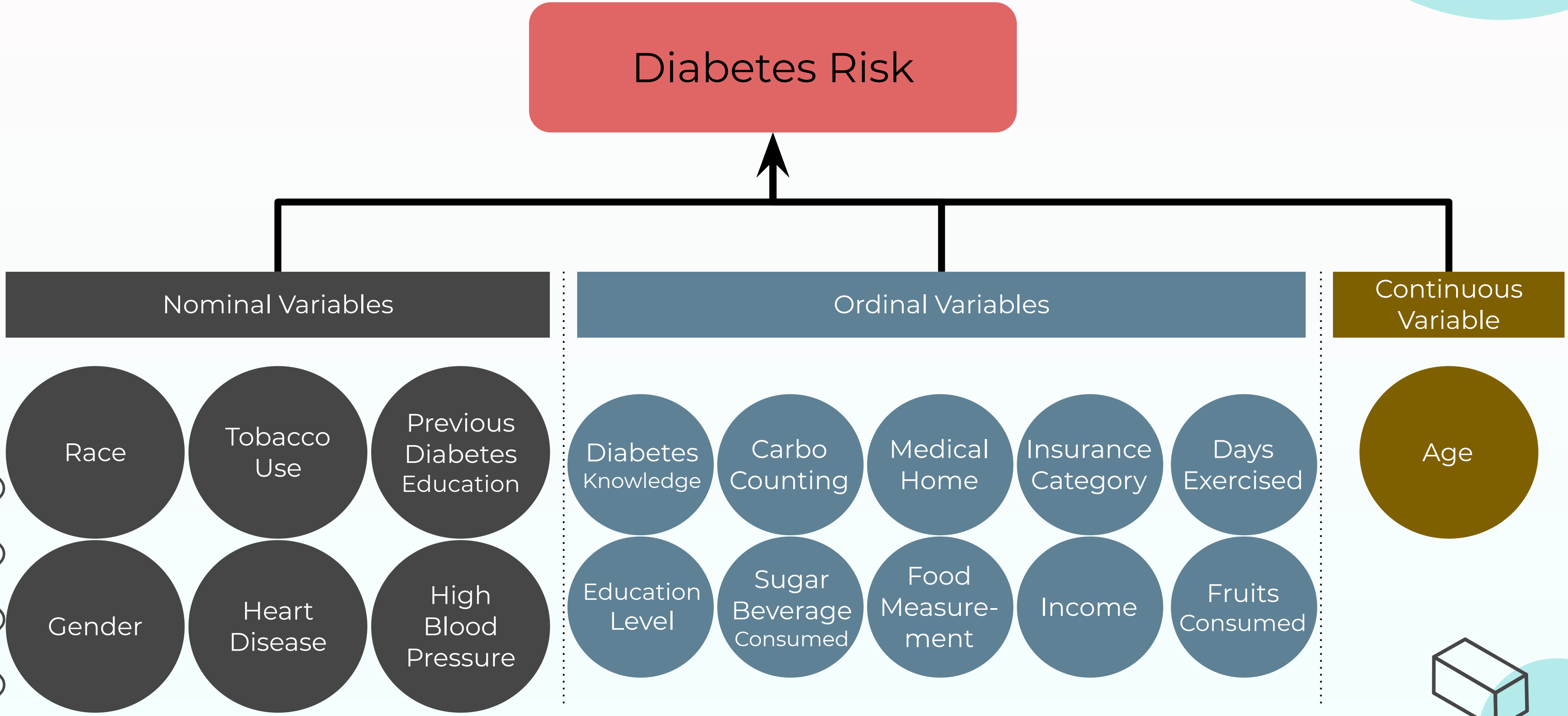
1) Logistic Regression

Neural Network

1) Neural Network

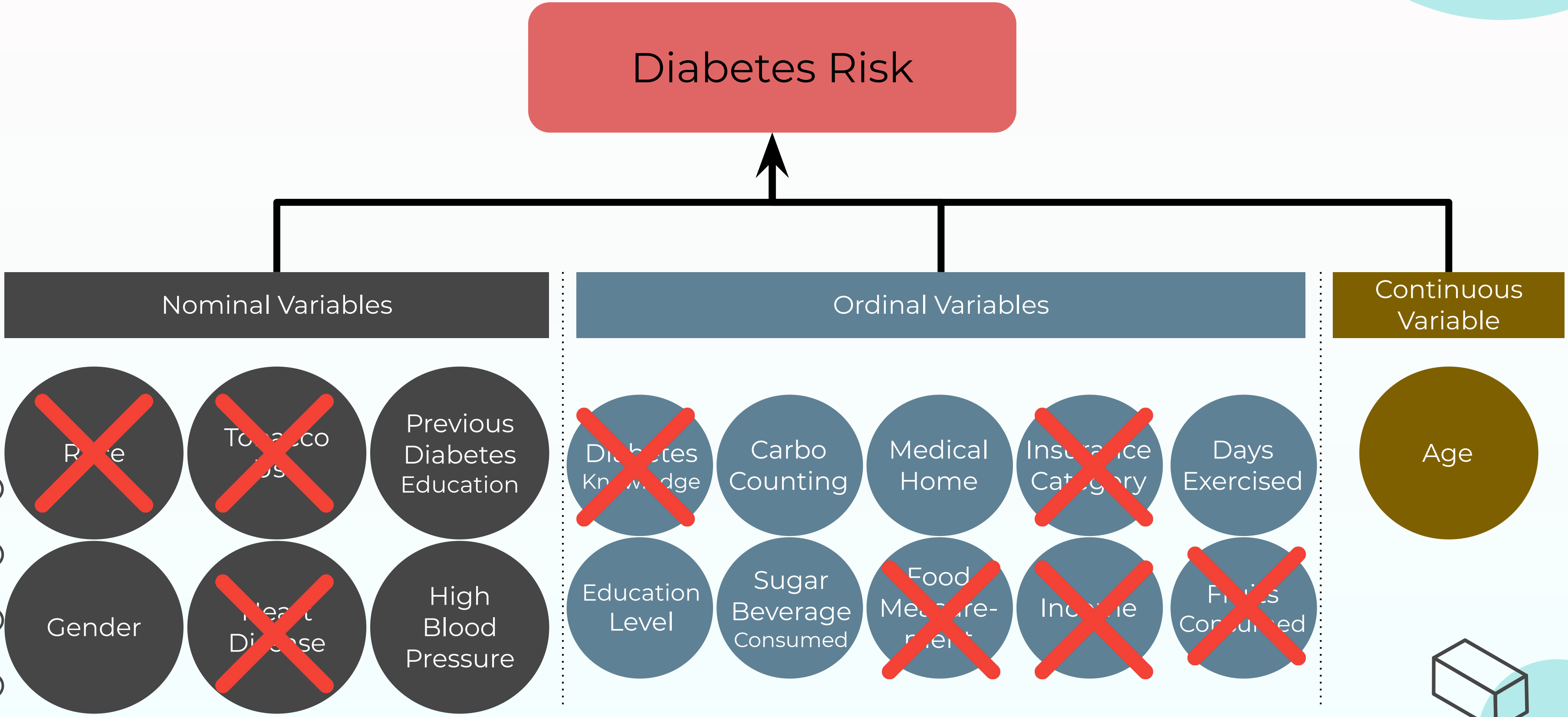
Spencer

Model Parameters



Spencer

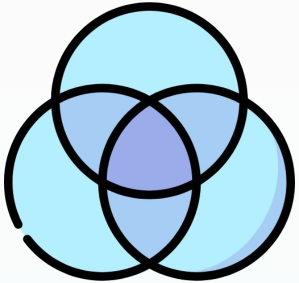
Logistic Regression: Modelling



Logistic Regression: Model Accuracy

Accuracy	False Negative Rate
69.8%	23.2%

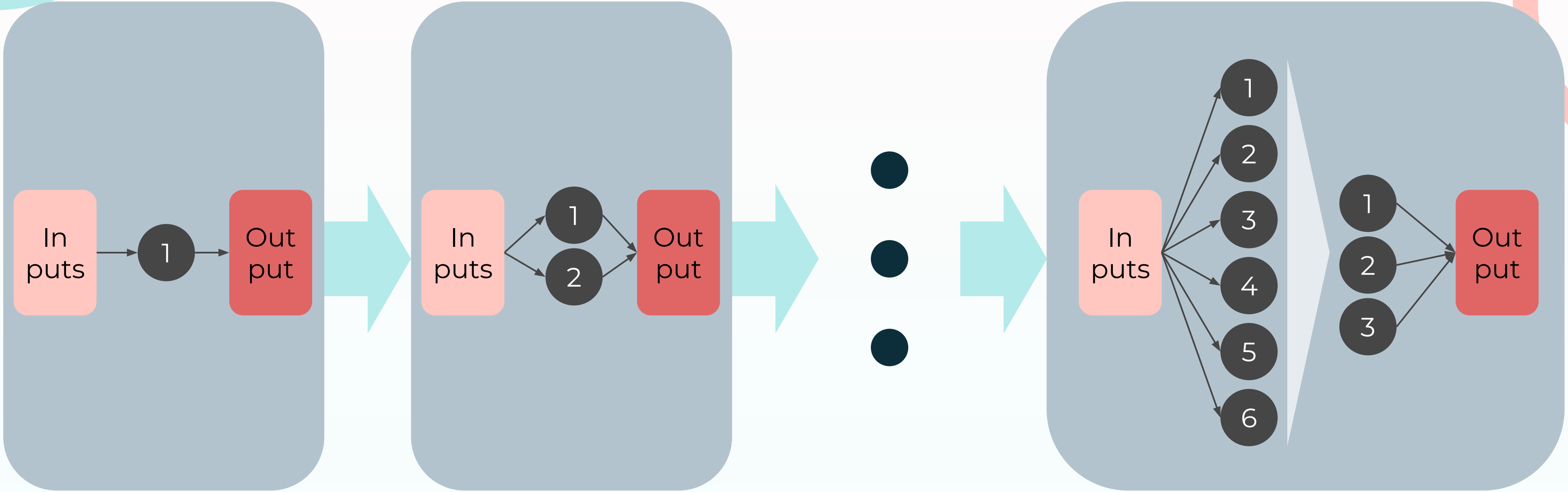
Logistic Regression: Limitation



Variable Selection	Simplistic model	Overfitting
• No multicollinearity between variables • Only important and relevant variables should be used	• Difficult to capture complex relationships	• Prone to overfitting on high dimensional dataset

Spencer

Neural Network: Tuning Process



For loop to **determine the optimal combination** of hidden layers and nodes to produce the model with best accuracy

Neural Network: Model Accuracy

Optimal Model

(1,1)
2 hidden layer
1 node each

Inputs → 1 → 1 → Diabetic Risk

Accuracy	False Negative Rate
70.2%	35.5%

Neural Network: Model Limitations

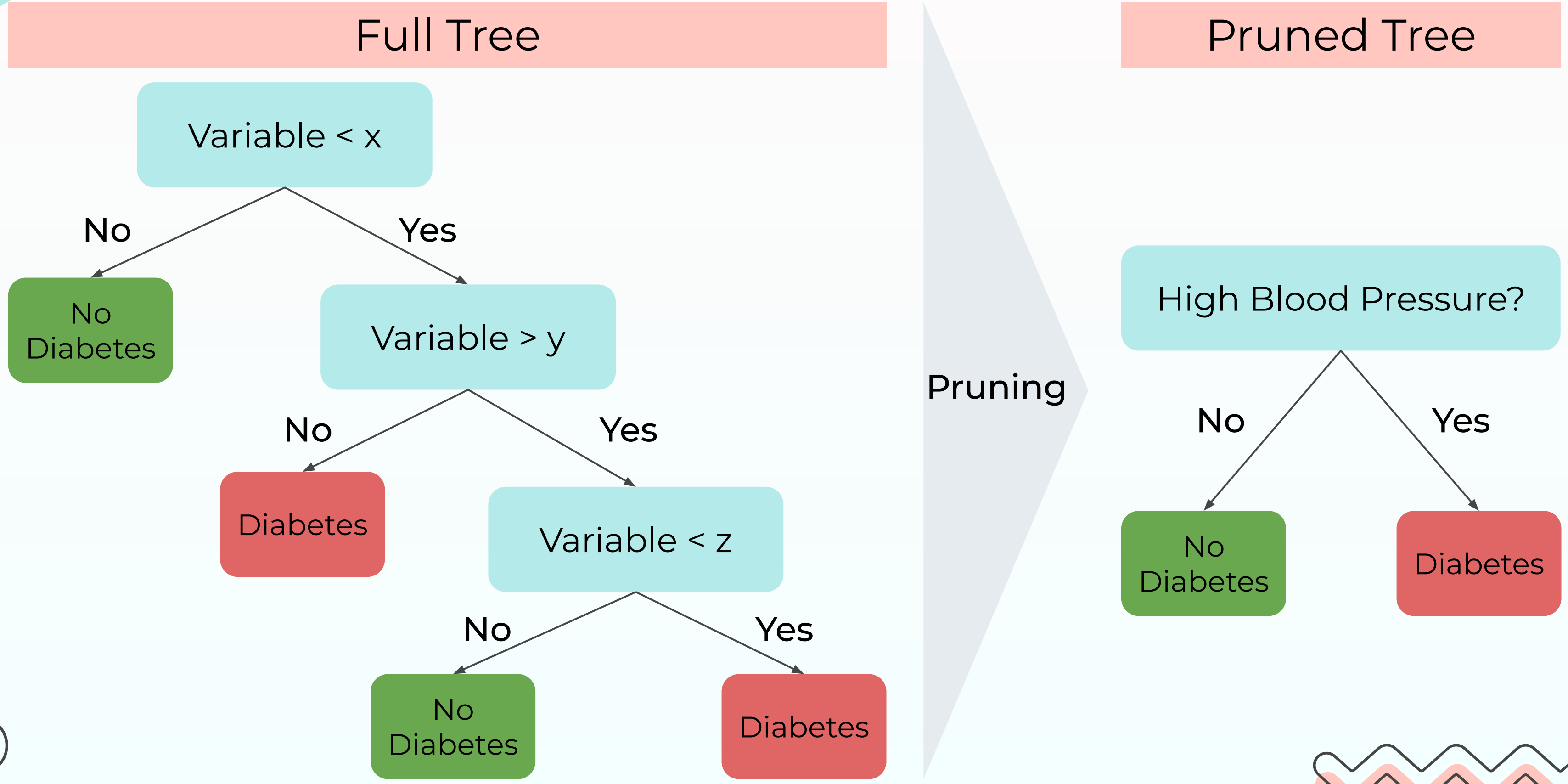
Time consuming training process

- Each network takes a long time to develop
- Develop multiple networks to optimize

Blackbox

- Unknown how different configuration of hidden layers affect accuracy
- Does not explain how variables affect diabetes

CART: Modelling Process

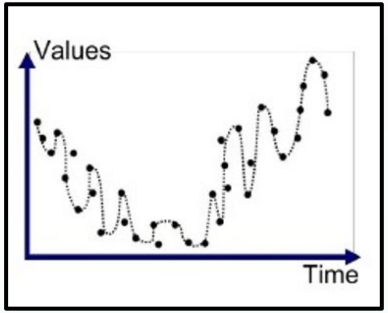


CART: Model Accuracy

Accuracy
67.3%

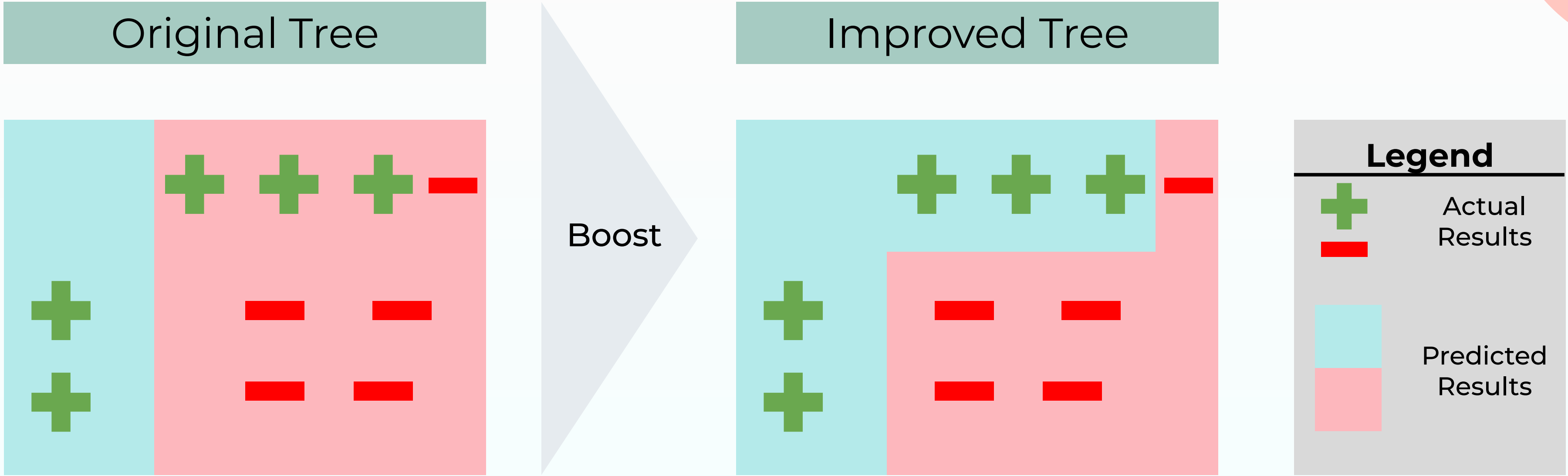
False Negative Rate
43.5%

CART: Model Limitations



Highly dependent on dataset	Overfitting
• Small change in dataset will cause tree to be unstable • Creates a completely new and different tree	• Tendency to overfit quickly at the bottom • Poor decisions if there are too few observation in the tree's lower nodes

XGBoost: Modelling Process



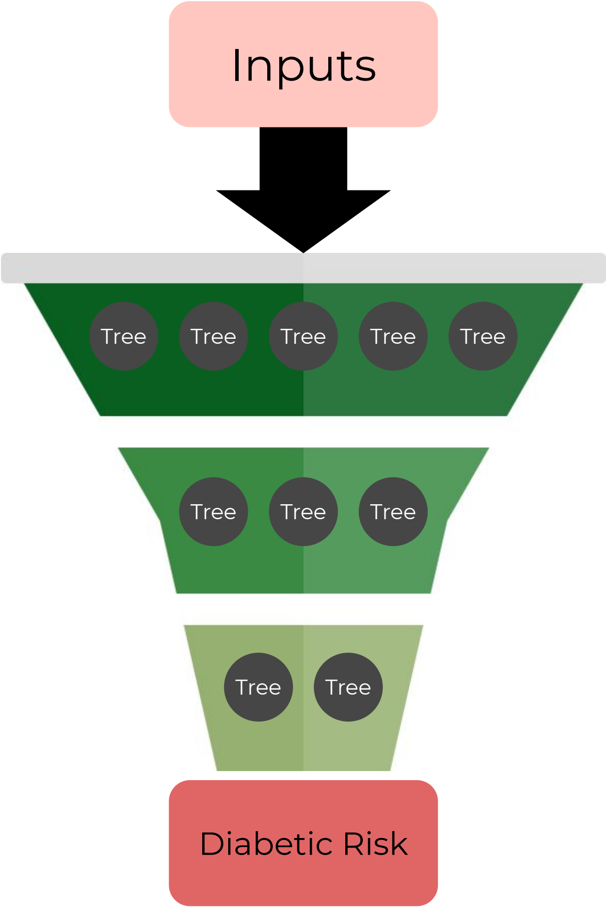
XGBoost : Model Accuracy

Accuracy	False Negative Rate
70.7%	35.0%

XGBoost : Model Limitations

Sensitivity
• Sensitive to outliers • Newer iterations are built by fixing previous errors, overcompensate the outliers

Random Forest: Modelling Process



Explanation
• The random forest model trains out a **large number** of CART trees • Input variables will be **evaluated by multiple CART models** within the random forest model • Each CART tree will provide an **independent decision** regarding diabetic risk • The random forest model will **collate responses** from individual trees • Assign diabetic risk based on **majority rule**

Chi Hui

Random Forest: Model Accuracy

Accuracy	False Negative Rate
73.0%	19.0%

Random Forest: Model Limitations

Time consuming training process

- Train 500 CART models
- Takes significantly longer when rolled out nationally

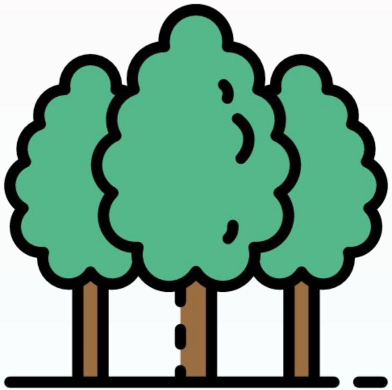
Blackbox

- Lack in explainability
- Decides by majority vote from individual trees

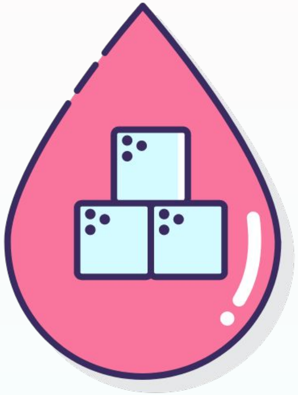
Model Evaluation

Model	Accuracy (%)	FNR (%)	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

Model Selection

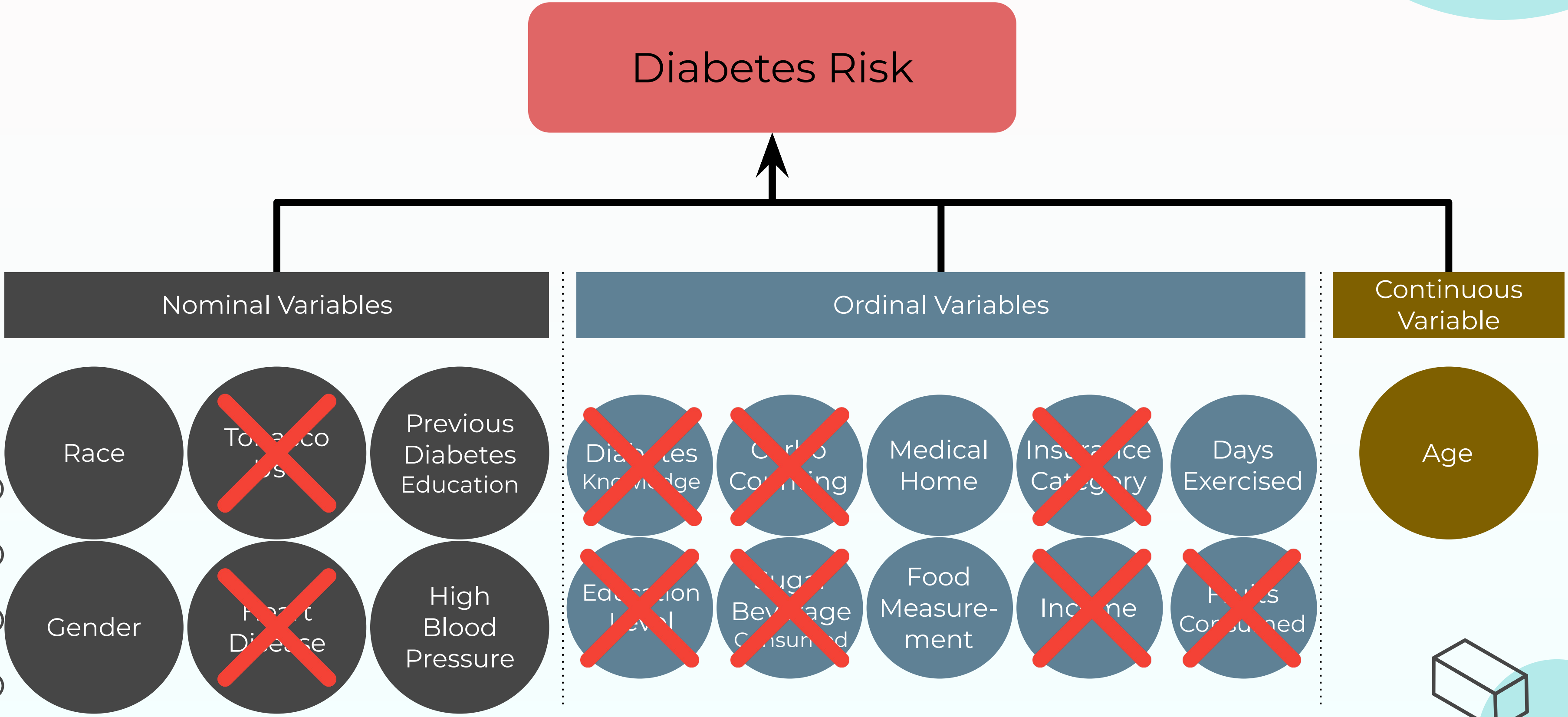


- Highest accuracy and lowest false negative rate
- Little explainability on variable importance and how these affect diabetic risk



- False negative rate is an important metric
- Gives our users who actually have diabetes false assurance that they do not

Variable Importance



03

Recommendations

Summary of Recommendation

Key Issues

- Diabetes Detection
  - Diabetes Risk Assessment
- Diabetes Programs / Policies
  - Promoting active lifestyles

Detection

Through taking the Diabetes Risk Assessment...

Early Education and Diabetes Risk Management

#1: Early Diabetes Detection via Risk Assessment

Reason	How?
Millions of Americans are undiagnosed	Predict diabetes risk based on easily obtainable information
Diabetes tests take time and money. Uninsured population may be hesitant.	↓
↓	Random forest model to predict diabetes risk based on survey input
Prevention through early detection that is quick and convenient	

#1: Self-Help Early Diabetes Detection

User takes risk assessment

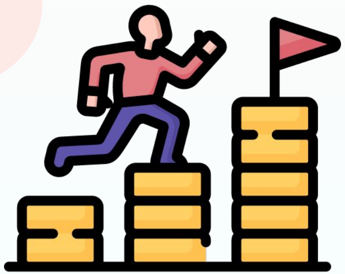
Based on inputs provided, random forest predicts user's diabetes risk (high / low)

At the end of the survey, provide links to diabetes education resources (for all users)

Kang Xun

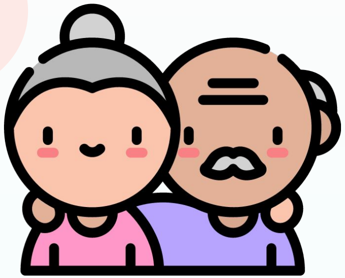
#2: Diabetes Policies and Programs – Promoting Active Lifestyles

1



Motivating (Americans to exercise) through incentives

2

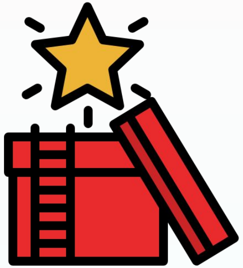


Targeting vulnerable age groups (the elderly population)

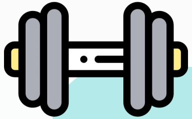
Natasha

Motivating through incentives

35% of Americans do not exercise due to lack of motivation



Provide incentives to motivate adoption of active lifestyle



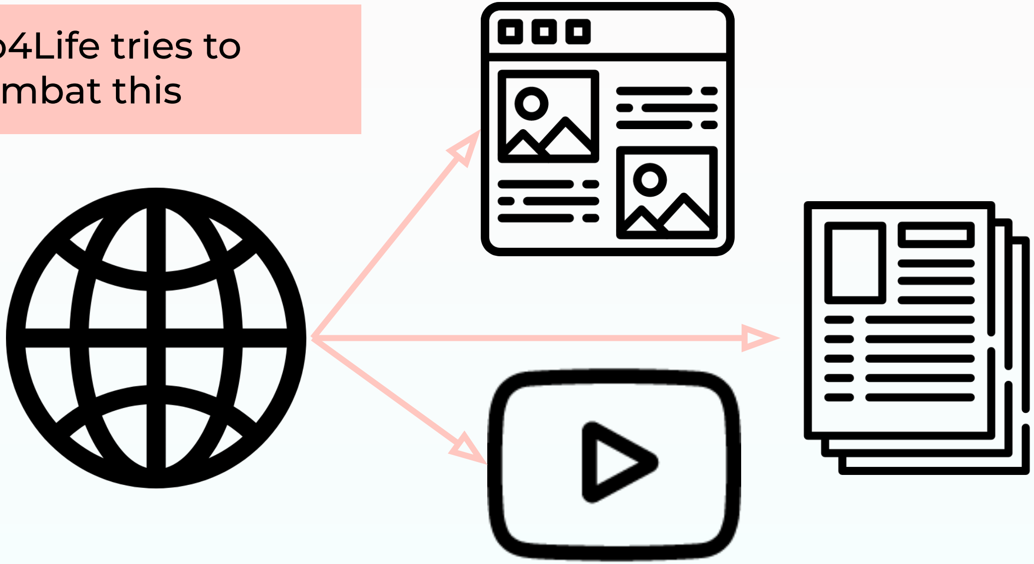
Collaborate with fitness centers and gyms to motivate people to exercise

Discount rates and points system

Targeting Vulnerable Age Groups (elderly)

Go4Life tries to combat this

Elderly feel “too old” to exercise



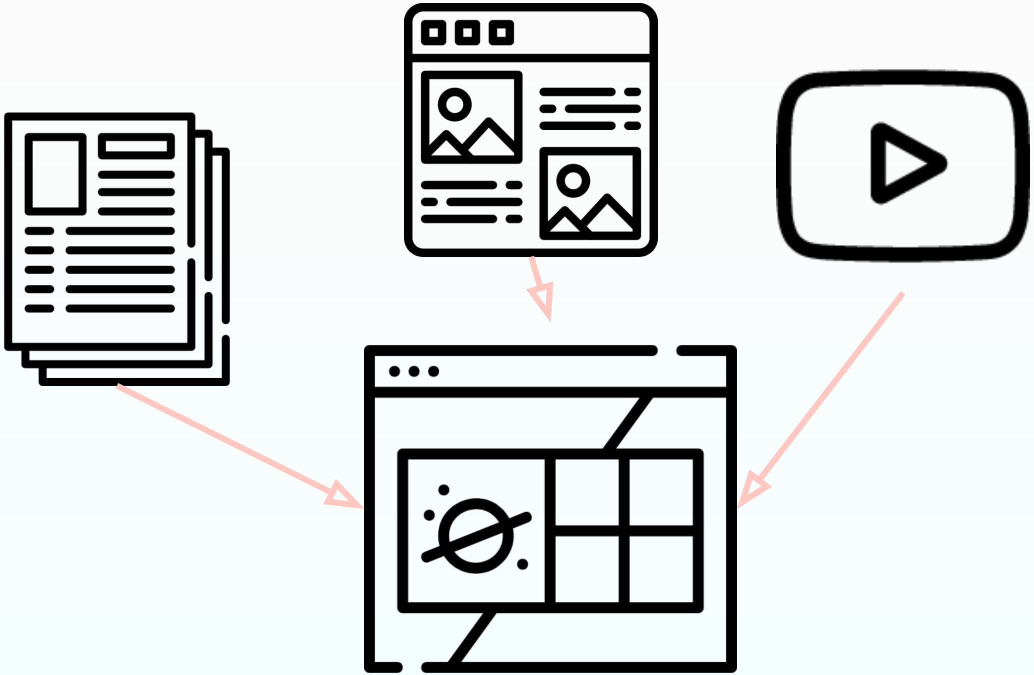
However, instead of one coherent website, Go4Life resources on exercising for the elderly is scattered and hard to search up

One-stop platform

Centralise resources

Easy and cheap to implement

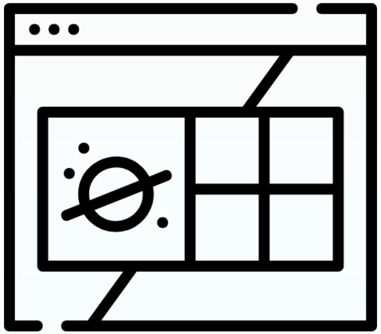
SEO Performance and amount of web traffic



Natasha

Encourage elderly exercise communities

Popularise exercise among the elderly through community building



Tap on website

To bring the elderly together and form exercise communities

The more elderly sign up (which is tracked via the website), the more effective the communities are

04

Conclusion

LIMITATIONS

Uncertainty over Economic Factors

Exercise Intensity & Duration

Carbohydrate & Sugar Consumption

Limited Dataset

Natasha

Summary of Recommendations

Diabetes Detection

Early Diabetes Detection
via Risk Assessment

Diabetes Programs and Policies

Promoting Active
LIfestyles