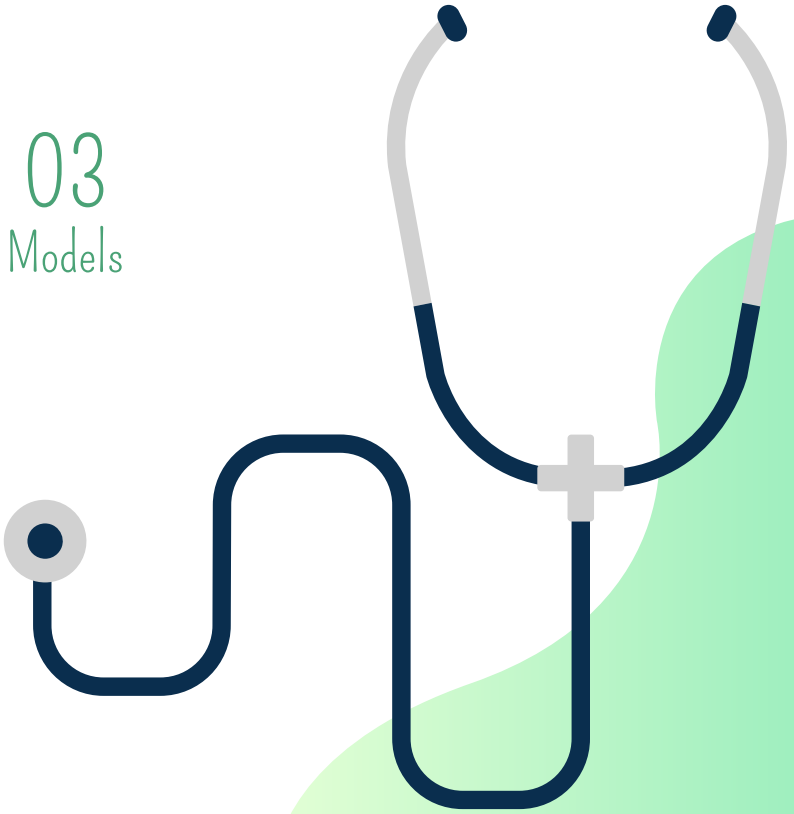
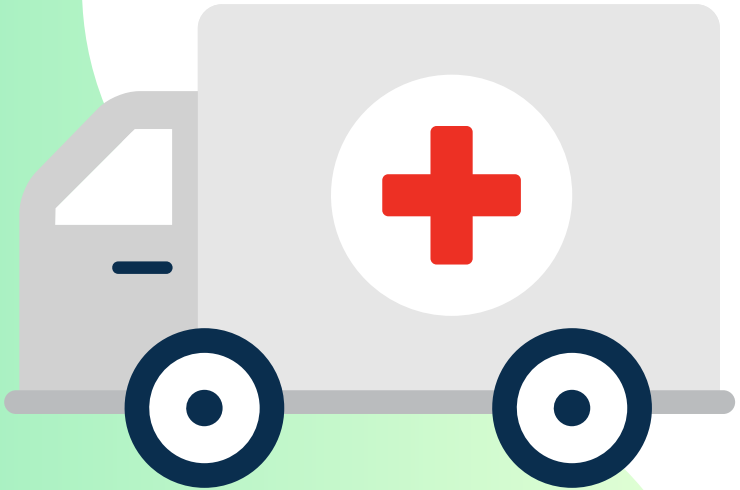
ST4248 Group B2 Final Presentation

Lim Xi Chen Terry (A0199513W)
Chew Bangyao Paul (A0204816J)
Chan Wen Yong (A0201868B)
Low Hon Zheng (A0204803R)

Table of Contents





01

Problem Statement

DIABETES INFOGRAPHICS

Diabetes in numbers

422 MILLION

422 million people have diabetes in the world

that's about 1 out of every 11 people

1 OUT OF 4

do not know they have diabetes

Main types of diabetes

TYPE 1 DIABETES

Body cannot produce insulin

TYPE 2 DIABETES

Body produce insulin but can't use it well

GESTATIONAL DIABETES

A temporary condition in pregnancy (GDM)

Problem Statement

To find out how an individual's profile would determine whether they are likely to be diabetic or not.

Objectives

Prediction

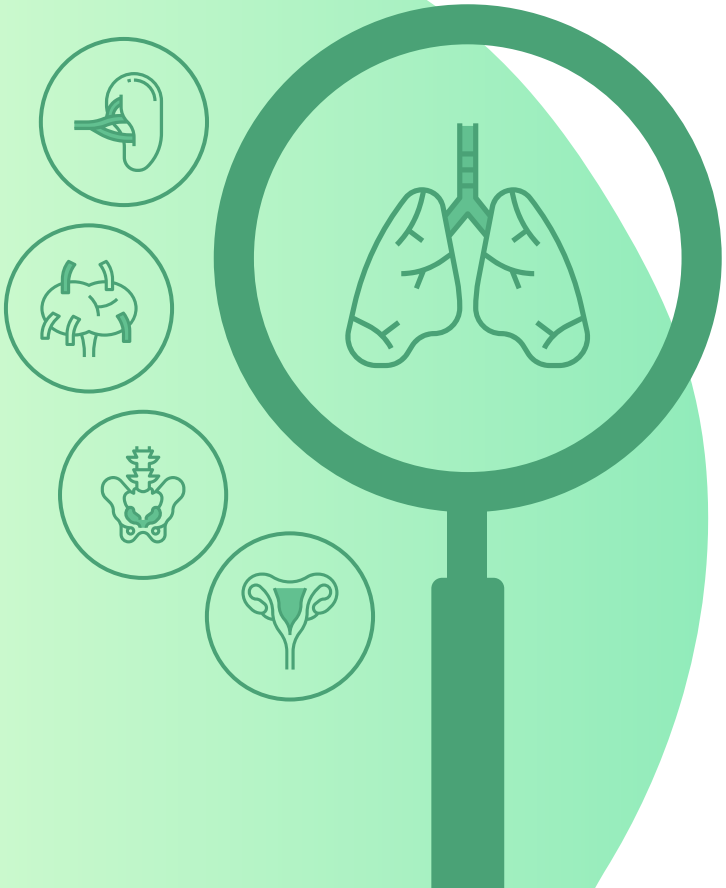
Accurately predicting diabetes using demographics and health – related data

Important Features

Identifying features that contributes the most to diagnosis of being diabetic

Reduced Feature Model

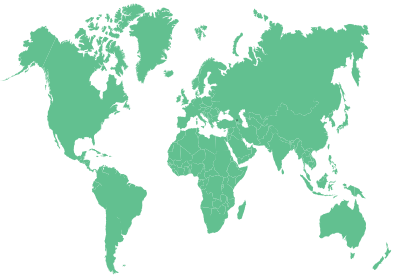
Using a model trained on important features to accurately predict diabetes



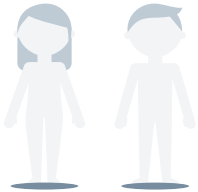
02

Data Exploration

BEHAVIOURAL RISK FACTOR SURVEILLANCE SYSTEMS 2020



America



401,958
respondents

279
questions

Data Cleaning

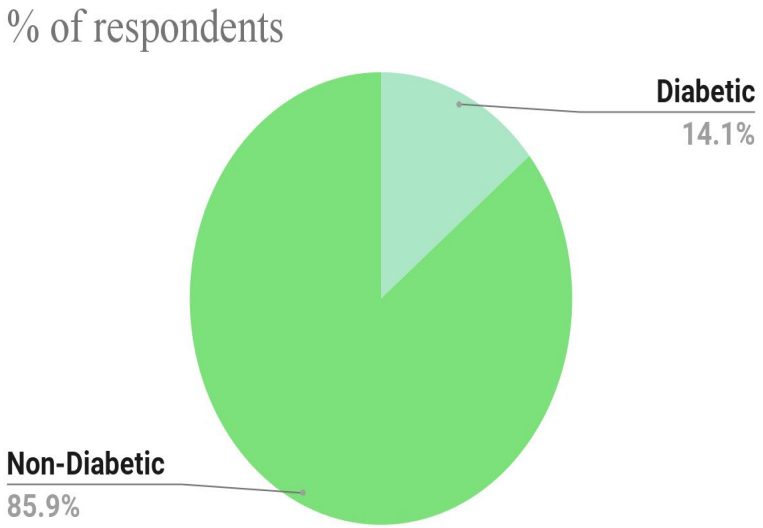
Select features

Converting values

Removal of rows with blanks

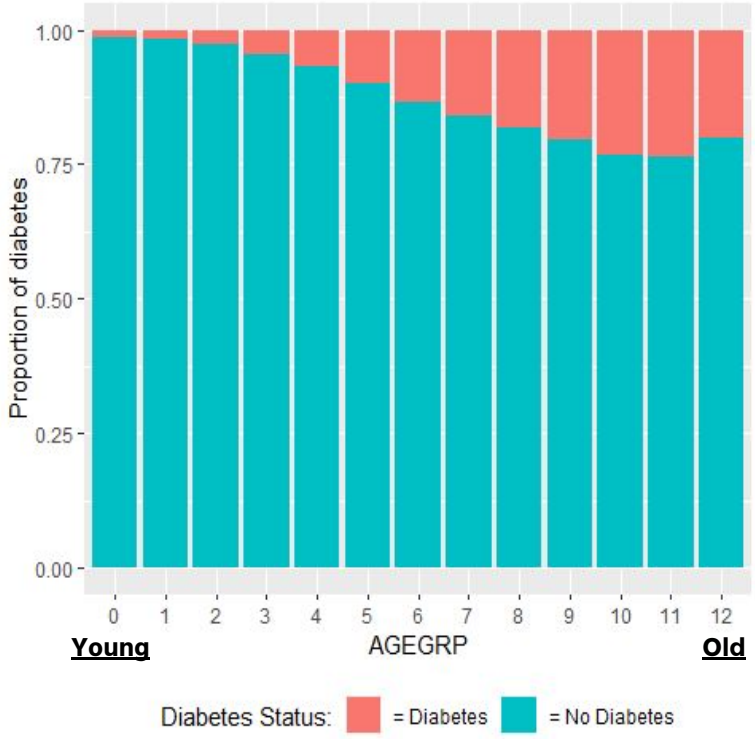
Removal of rows with ambiguous data

Imbalanced Data

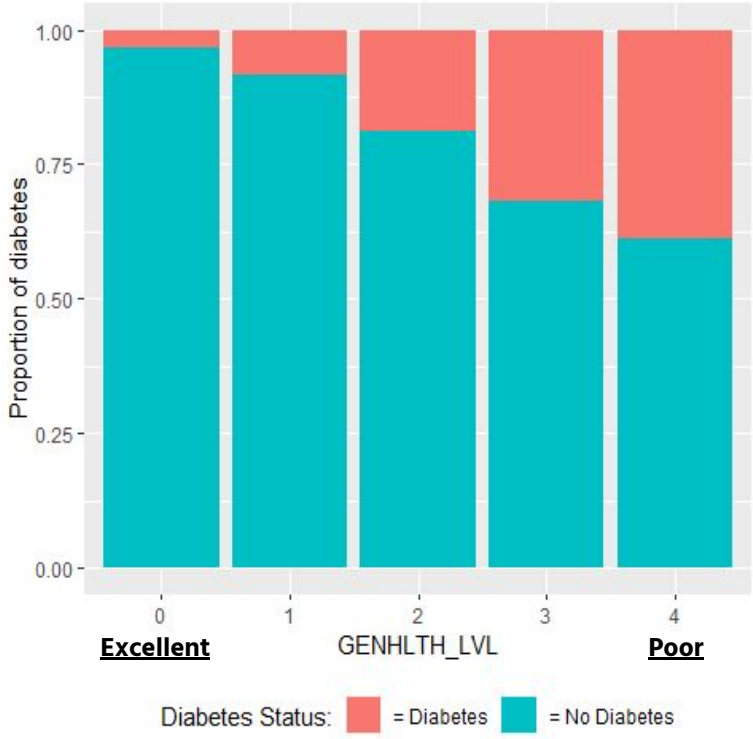


EDA

Age Group

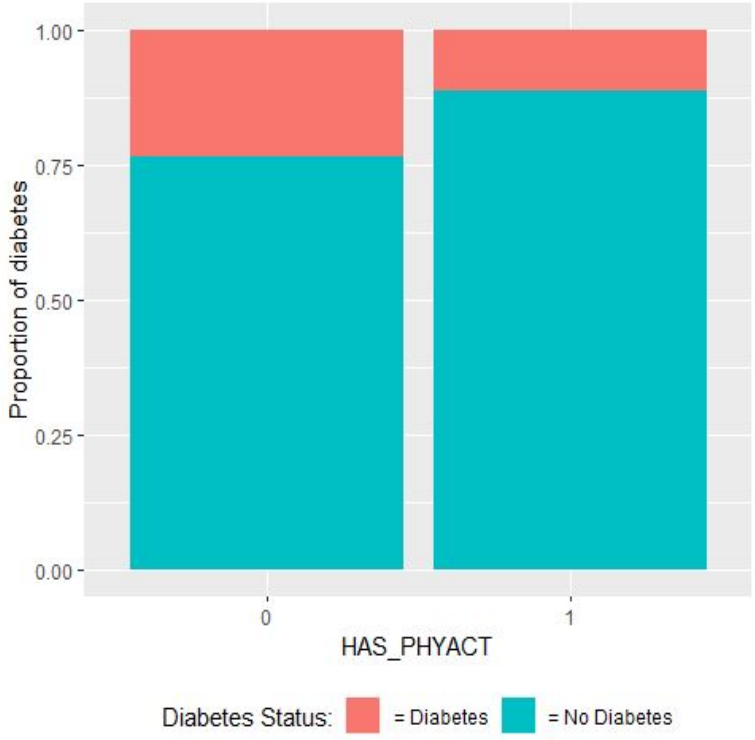


General health

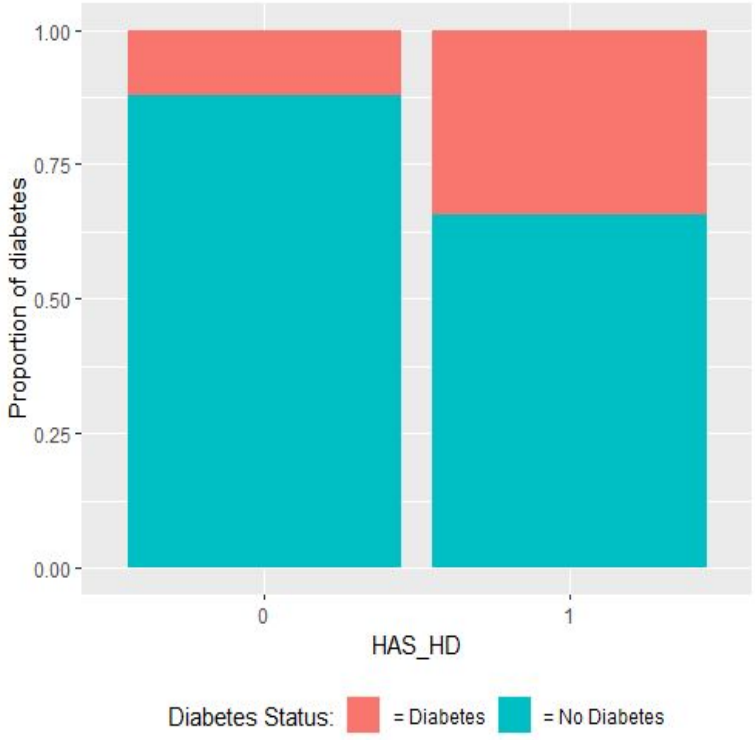


EDA

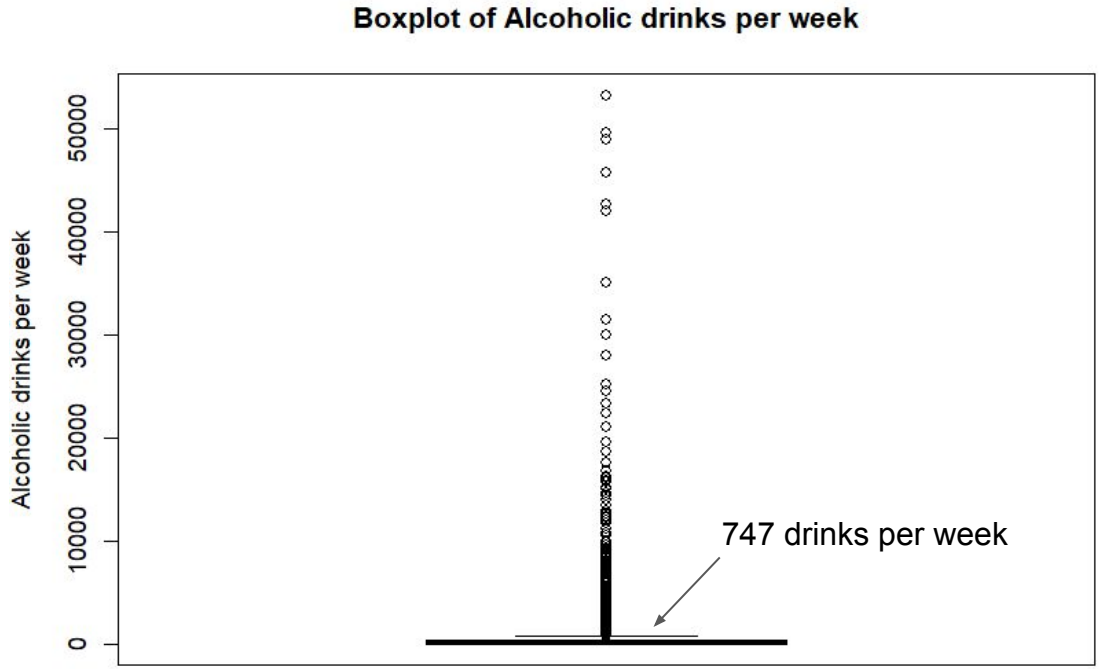
Physical Activity



Heart Disease



Outlier removal

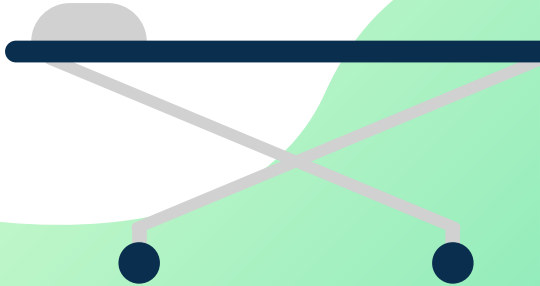


Correlation

Comparison	Measure	Analysis	Result
Continuous variables	Pearson Correlation Coefficient	Moderate correlation found between: - NUM_POORHLTH and NUM_POORPHYHLTH (0.611) - NUMPOORHLTH and NUM_POORMENTHLTH (0.451)	Remove NUMPOORHLTH
Binary Categorical variables	Cramér's V	All values < 0.5 and are close to 0.	No correlation found
Non-Binary Categorical variables	Phi Coefficient	All values are either < 0.5 or > -0.5 and are close to 0.	
Continuous & Binary Categorical variables	Point-Biserial Correlation Coefficient	All values are either < 0.5 or > -0.5 and are close to 0.	

03

Models



Models

Logistic Regression (Baseline)

No Down-sampling

No scaling

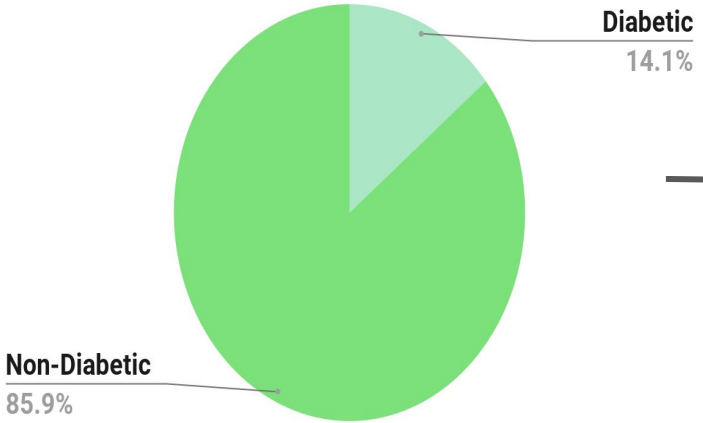
Parametric method

Comparison

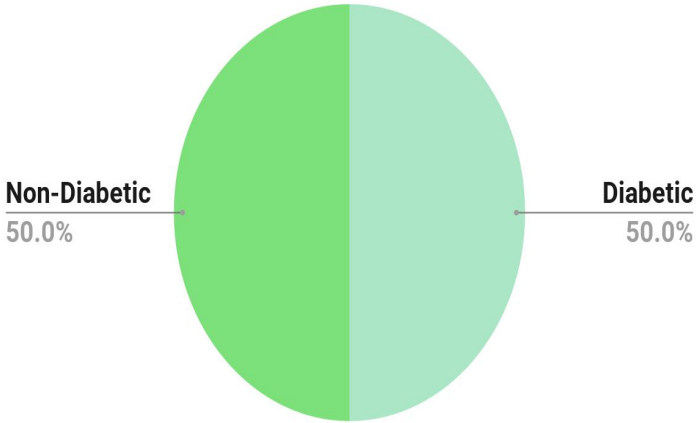
	Logistic Regression
Accuracy	0.862
Precision	0.540
Recall	0.149
F1 score	0.233
AUC	0.564

Imbalanced Data

% of respondents



% of train data



Models

Logistic Regression (Baseline)

No Down-sampling

No scaling

Parametric method

L1 Regularised Logistic Regression

Down-sampling

No scaling

Parametric method

Comparison

	Logistic Regression	L1 Regularised Logistic Regression
Accuracy	0.862	0.731
Precision	0.540	0.313
Recall	0.149	0.755
F1 score	0.233	0.442
AUC	0.564	0.741

Models

Logistic Regression (Baseline)

No Down-sampling

No scaling

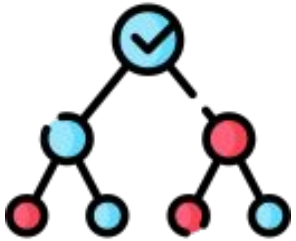
Parametric method

L1 Regularised Logistic Regression

Down-sampling

No scaling

Parametric method



XGBoost

Down-sampling

No scaling

Non parametric method

Comparison

	Logistic Regression	L1 Regularised Logistic Regression	XGBoost
Accuracy	0.862	0.731	0.747
Precision	0.540	0.313	0.326
Recall	0.149	0.755	0.741
F1 score	0.233	0.442	0.452
AUC	0.564	0.741	0.744

Models

Logistic Regression (Baseline)

No Down-sampling

No scaling

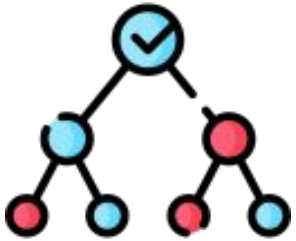
Parametric method

L1 Regularised Logistic Regression

Down-sampling

No scaling

Parametric method

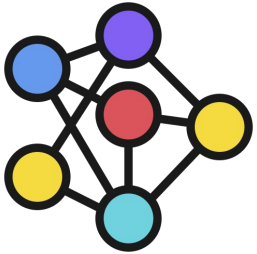


XGBoost

Down-sampling

No scaling

Non parametric method



NN (3 hidden layers)

Down-sampling

Standard scaling

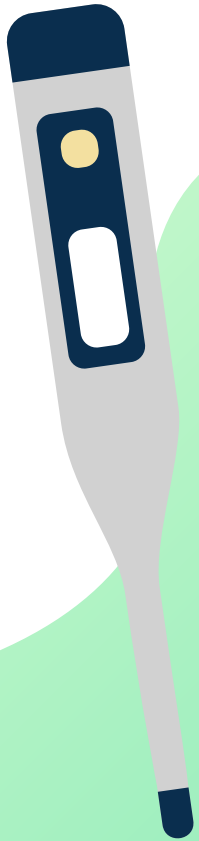
Parametric method

Comparison

	Logistic Regression	L1 Regularised Logistic Regression	XGBoost	NN
Accuracy	0.862	0.731	0.747	0.747
Precision	0.540	0.313	0.326	0.325
Recall	0.149	0.755	0.741	0.734
F1 score	0.233	0.442	0.452	0.451
AUC	0.564	0.741	0.744	0.742

04

Variable Selection



Reduced feature list

CATEGORICAL VARIABLES

- HAS STROKE
- HAS HEART DISEASE
- HAS PHYSICAL ACTIVITY
- HAS FINANCIAL ISSUES
- RACE
- AGE GROUP
- SEX
- DAYS SINCE LAST CHECK UP
- GENERAL LEVEL HEALTH

CONTINUOUS VARIABLES

- BMI
- NUMBER OF DRINKS DRANK PER WEEK

Reduced feature model

	Accuracy	Precision	Recall	F1 score	AUC
Logistic Regression with reduced features	0.723	0.309	0.780	0.443	0.747
L1 Regularised Logistic Regression	0.731	0.313	0.755	0.442	0.741

05

Future Works

Additional Data Collection

- Blood pressure
- Cholesterol
- Diet

THANK YOU!