EBA5007 Capstone Project:

Final Report

Khoo May Sze – A0198537L

10 May 2020

Executive summary

The project aims to provide healthcare professionals and insurance providers with statistically significant non-invasive factors, such as lifestyle, medical history and body measurements, that are more relevant for predicting Type 2 diabetes (T2D) in Asian Americans. Individuals of Asian descent have different body composition and thus, healthcare professionals cannot apply generic considerations towards their prognosis.

The project applies different supervised machine learning techniques to distinguish which type of classifiers best fit these types of features. There are 2 types of classification problems studied – binary and multi-class. The models employed include Logistic Regression, Decision Trees and various boosted variations, Support Vector Machines, k-Nearest Neighbour, Naïve Bayes and Neural Network. Many of the non-invasive features are correlated such as body measurements and physical activity levels of an individual. The boosted trees classifiers offer the most stable performance and overcome the inherent multi-collinearity issues. Further, body measurements, specifically waist circumference and sagittal abdominal diameter, are more likely to be significant than BMI in examining health status for Asian Americans.

Results of the multi-class classification problem highlight different key health indicators and lifestyle choices that are more vital to observe for individuals with diabetes or are pre-diabetic. Specifically, restricting one's alcohol consumption is more relevant for those who are already diabetic. Factors that form a healthy lifestyle, reduce visceral fat stored around the waist and lower cholesterol levels, are more relevant to help lower the risk of pre-diabetic Asian Americans. Doctors and healthcare professionals can adapt the findings of this project to prescribe more effective and personalised advice and treatment for patients who are of Asian descent.

Table of Contents

1.0 Introduction

The project is a longitudinal study of predicting Type 2 diabetes (T2D) using lifestyle factors and other easily collected data without specialized medical testing in Asian Americans over 2 cycles of the National Health and Nutrition Examination Survey (NHANES) survey, 2013–2014 and 2015–2016.

The United States (US) Department of Health & Human Services documents that there are at least 1 in 4 Americans who are unaware they have diabetes. However, this ratio narrows to 1 in 2 for Asian Americans. While there is existing literature on predicting T2D, there are few focusing on ethnic-specific subgroups. Research focusing on the Asian American subgroups tend to point out the trends and differences between Asian Americans and US adults of other races. The papers suggest more relevant factors and different cut-off points for common factors in a T2D diagnosis (McNeely and Boyko, 2004, Yim et al, 2010, Wang et al, 2011, Lee et al., 2011, Hsu et al, 2015, Nguyen et al., 2015). The value of their predictive usefulness is not quite substantiated yet. For example, Nguyen et al. (2015) summarises the ongoing efforts in the collection of T2D prevalence data and the factors that go into prevention of T2D. The paper notes that these factors cannot be applied across the board for all races and suggests potential stronger predictors for Asian Americans. These suggestions include lowering the BMI cut-off points for overweight/obesity and measuring waist circumference. Most papers study diabetes across the entire American adult population and use the normal BMI cut-off point of 25 (for overweight) (Lee et al., 2011, Semerdjian and Frank, 2017, Firouzi et al., 2018).

Business objectives of the project are:

1. providing doctors and healthcare professionals with better insight into alternative factors such as body measurements and lifestyle (other than blood glucose tests) that have two-fold impact
    a. signal increased risk of T2D and help prescribe appropriate preventive lifestyle changes to decrease chances of or delay the T2D diagnosis
    b. an early diagnosis of T2D by identifying individuals who have or at risk of diabetes so that doctors and healthcare professionals can prescribe additional tests to properly diagnose and manage their treatments
2. using the predictive model to further delineate significant factors that affect different stages of the T2D diagnosis, be it undiagnosed, pre-diabetes or high risk

The data analytic objective that maps to the business objectives is to support the usefulness of these proposed variables with statistical reliability to aid doctors and health professionals in T2D prediction for Asian Americans.

Further, the data-driven results can be applied outside of the medical domain by helping insurance companies identify and predict personnel who are at risk of T2D. The project also seeks to study the predictive effect of the updated BMI guidelines for Asian Americans adopted in 2015 by the American Diabetes Association (ADA).

1.1 Summary of Phase 1 report

The first report details the progress of how the final dataset[1] for modelling is collected, consolidated and cleaned. Figure 1.1 summarises the dataset pre-processing journey.

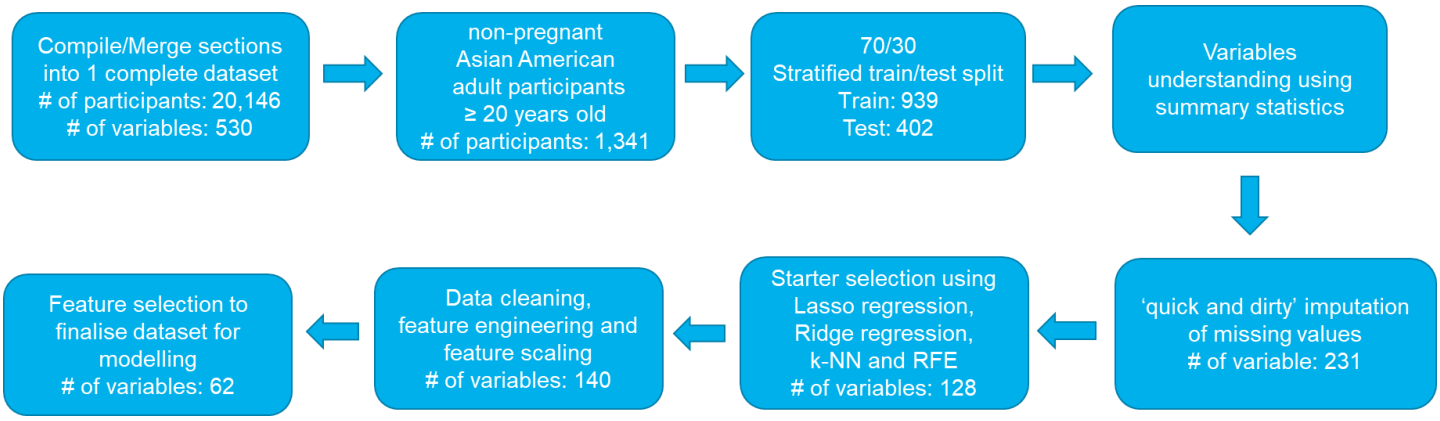


Figure 1.1 Data pre-processing

The data pre-processing is applied to both train and test sets which includes:

i. narrowing down the 530 variables to about 60 features that are statistically significant and engineered based on existing research findings
ii. replacing the NA values of continuous variables with median values, and NA values of categorical variables with mode values
iii. making sure that the values in each variable is meaningful, even 0 values

[1] See Appendix A.

iv. feature engineering of some original variables to better capture the meaningfulness. Variables such as glucose and BMI are categorized[2]

The project also aims to compare the binary classification method (A) and the multi-class classification method (B) in predicting T2D. The outcome variables are:

i. for binary classification method A, *DIQ010A* where 1 indicates that the participant has diagnosed/undiagnosed diabetes, 2 otherwise.

ii. for multi-class classification method B, *DIQ010B* where: **1 indicates that the participant has diagnosed diabetes** – answered '1' (yes) or '3' (borderline) to DIQ010 question of whether a doctor has advised that he/she has diabetes, as well as a recorded fasting plasma glucose (FPG) reading >= 126; **2 indicates the participant has undiagnosed diabetes** – answered '2' (no) to DIQ010 question of whether a doctor has advised that he/she has diabetes, as well as a recorded FPG reading >= 126; **3 indicates that the participant has prediabetes** – FPG reading between 100 – 125; **4 indicates that the participant has no diabetes** – FPG reading <= 100

The project takes on a heuristic analytical approach to solving the 2 classification problems. Supervised machine learning classifiers such as Decision tree and various boosted variations are applied in the modelling phase. Ensemble classifiers are not considered because the project serves to provide easy-to-understand techniques that can be implemented for field practitioners and place more emphasis on the findings and analysis from the model results. Further, prior research find that boosted tree classifiers perform better than ensemble classifiers (Semerdjian and Frank, 2017). The models employed for the project are discussed in Chapters 2 and 3.

1.2 Project software

The project is conducted in **R**, using ***caret*** for most of the models trained. Complementary packages to ***caret*** for model training such as ***gbm***, ***fastAdaboost***, ***kknn***, ***neuralnet***, ***randomForest***, ***klaR***, ***caTools*** and ***naivebayes*** are also employed. The model performance evaluation phase also includes the use of packages ***pROC*** and ***PRROC***. These are further discussed in the following chapters and referenced in the appendices. Common data science libraries such as ***tidyverse*** and ***dplyr*** are used, alongside graph plotting libraries such as ***ggplot2***.

[2] See Appendix B.

Packages employed for the data pre-processing and feature selection phases include ***SASxport*** to convert and read the original dataset files, ***mice***, ***VIM***, ***splitstackshape***, ***mlbench*** and ***fastmatch***. The data balancing process for the binary and multiclass classification problems uses packages ***DMwR*** and ***UBL*** respectively. The source codes are provided in separate folder as part of the report submission.

2.0 Binary classification

The project employs various classifiers to validate the importance of features selected in the prediction of T2D in Asian Americans. Given that there are many good performing models documented in existing literature to predict T2D, the project's main focus is to offer healthcare professionals and insurance providers better insight on a person's existing health and lifestyle factors that may indicate likelihood of T2D diagnosis.

The predictive models are built using k-fold cross-validation on the training dataset. As there are imbalanced classes (11.3% diabetic in training dataset), the Synthetic Minority Oversampling Technique (SMOTE) is employed to analyse model performance following existing literature (Nguyen et al., 2019). The models are trained using the balanced dataset to prevent it from focusing more on the prediction accuracy of the majority class. For the binary classification, the dataset balancing is achieved using the package *DMwR*.

Table 2.1 Dataset Class Proportion – *DIQ010A*

	1	**%**	**2**	**%**
Original	123	13.1	816	86.9
After SMOTE	861	50	861	50

The following classifiers are employed to predict T2D in the project:

1. Logistic Regression (LogR)
2. Boosted Logistic Regression (Logitboost)
3. Decision Tree (DT)
4. Random Forest (RF)
5. Extreme Gradient Boosting Trees (XGB)
6. Stochastic Gradient Boosted Trees (GBM)
7. Adaptive Boosting Trees (Adaboost)

8. Neural Network
9. Support Vector Machines (SVM)
10. Naïve Bayes (NB)
11. k-Nearest Neighbours (k-NN)

The following sections will document the modelling approach, evaluate performance of the models, detail the sensitivity analysis and robustness tests undertaken, and conclude on the weighted importance of features selected by these models.

2.1 Classification models

Model performances are evaluated using a combination of metrics derived from the confusion matrix, Area Under Curves (AUC) of Receiver Operating Characteristics (ROC) curve and Precision-Recall (PR) curve. This is summarised in Table 2.3.

2.1.1 Logistic regression and boosted variation

The logistic regression model is a common model utilised for binary (two-class) classification problems. On top of the classic model, a gradient boosting technique – Extreme Gradient Boosting (XGB) – is applied to learn a linear model for this classification problem.

The logistic regression model trained using *glm()* function, performs the poorest in terms of the recall metric, 48.1% and has the smallest ROC-AUC at 0.308. This may be due to multi-collinearity issues within the dataset. The variables that are highly correlated are not removed in the data pre-processing stage because of the project objectives in determining useful features. Further, despite its multi-collinearity violation, these features are retained for modelling because domain knowledge in existing literature established the features' importance and predictive ability in the T2D prediction.

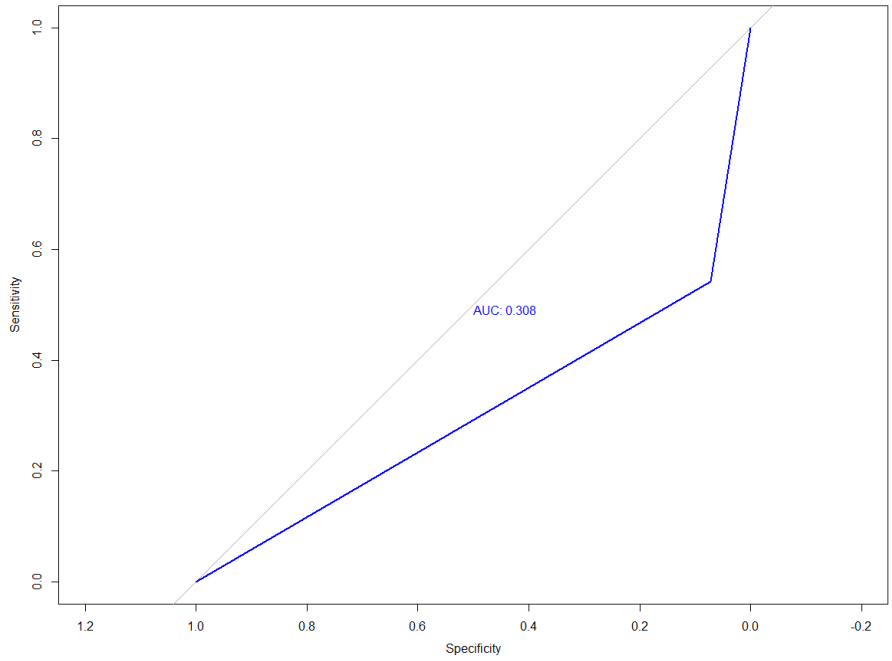


Figure 2.1 ROC-AUC of LogR (Binary)

Stepwise (backwards) elimination is also adopted whereby all independent variables are first entered into the regression formula and one variable is progressively removed if its p-value exceeds 0.05 (i.e. not statistically significant)[3].

[3] See Appendix E for resulting model.

The boosted logistic regression model, with a random selection of tuning parameter combinations[4] performed better with a recall of 80.0% and a much better ROC-AUC illustrated in Figure 2.2.

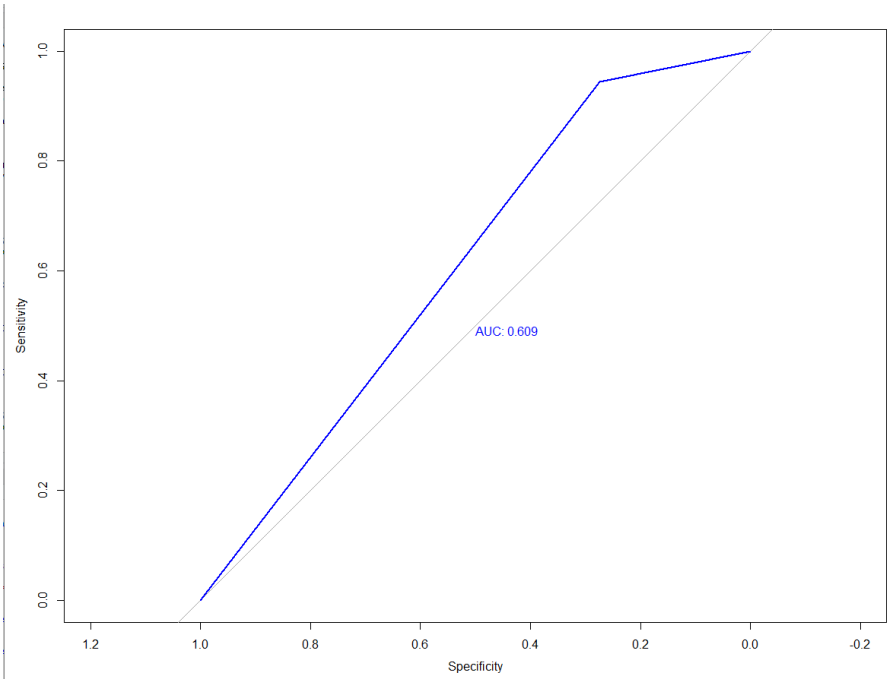


Figure 2.2 ROC-AUC of Logitboost (Binary)

2.1.2 Classification trees

There are 5 types of classification trees – Decision trees, Random forest, XGB trees, Stochastic gradient boosted Trees and Adaptive boosting trees. Given the inherent multi-collinearity within the dataset which can lead to skewed or misleading results, algorithms of decision trees and their boosted variations are immune to multicollinearity by nature. When they decide to split, the tree will choose only one of the perfectly correlated features.

Decision trees recursively split features based on their target variable's impurity. The best split is chosen by minimising the Gini index. The Gini index measures how often a randomly chosen attribute from a set is incorrectly labelled[5].

[4] See Appendix D for R packages used and tuning parameters.
[5] See Appendix F for the resulting decision tree and the other ROC-AUC plots.

The boosted trees use the decision trees as base learners by combining many weak learners to form a strong learner. The algorithms train individual models in a sequential way and each individual model learns from mistakes made by the previous model.

Random forest involves building multiple trees by taking bootstrap samples and random sampling of features. Thus, Random forest is acknowledged as a model that is unlikely to overfit. Gradient boosting involves creating new models that predict the residuals of prior models. The predictions from new models and that of prior models are then added together to establish the final prediction. It utilises a gradient descent algorithm to minimize the loss when adding new models. Adaptive boosting applies greater weights on data points that are difficult to predict. As the tree grows, the weights of difficult-to-classify observations are increased, and the weights of easy-to-classify observations are decreased. The final model is the weighted sum of the predictions made by the previous tree models.

The XGB trees and Random forest perform the best out of the 5 types, and out of all the models undertaken in the project. XGB trees achieve the highest recall at 82.7% and Random forest achieve the biggest ROC-AUC of 0.705[6].

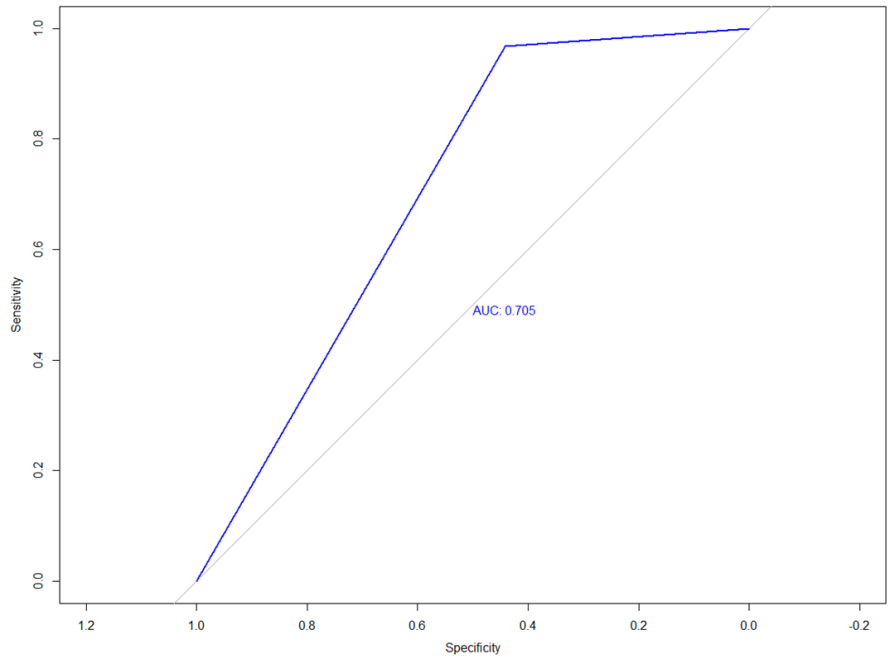


Figure 2.3 ROC-AUC of RF (Binary)

[6] See Appendix D.

2.1.3 Neural network

A feed forward neural network is employed in the initial modelling phase as part of feature selection. It results in a 119-5-1 network with 606 weights for the *DIQ0101A* model. 2 of the models utilised for feature selection, neural network and k-NN classifier are also undertaken in the modelling phase to allow comparison of features. The resulting feed forward neural network in this phase is a 73-9-1 network with 676 weights[7].

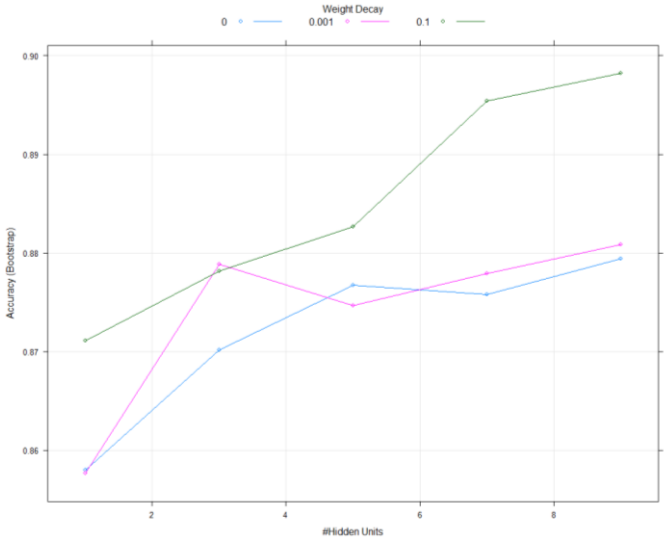


Figure 2.4 Neural Network (Binary)

While the model accuracy is high at 80.6%, the recall metric only scores at 50%.

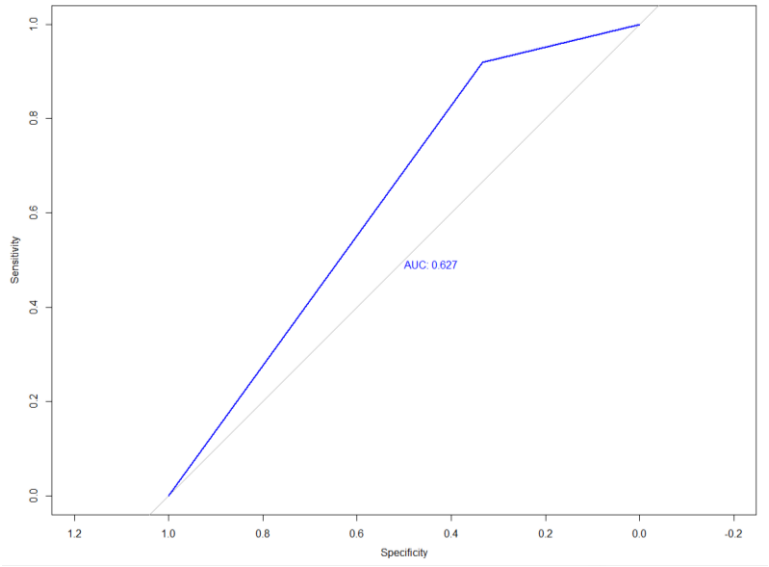


Figure 2.5 ROC-AUC of Neural Network (Binary)

[7] See Appendix D.

2.1.4 k-Nearest Neighbours

The k-NN classifier employs Euclidean distance as its method. The results[8] of initial modelling (feature selection) using k-NN classifier present meaningful variables supported by existing literature in the T2D prediction. It also proposes variables to consider for dietary and lifestyle factors. The same algorithm is employed for the modelling phase and the resulting k-NN classifier performs respectably scoring 63.5% for recall and a ROC-AUC of 0.635.

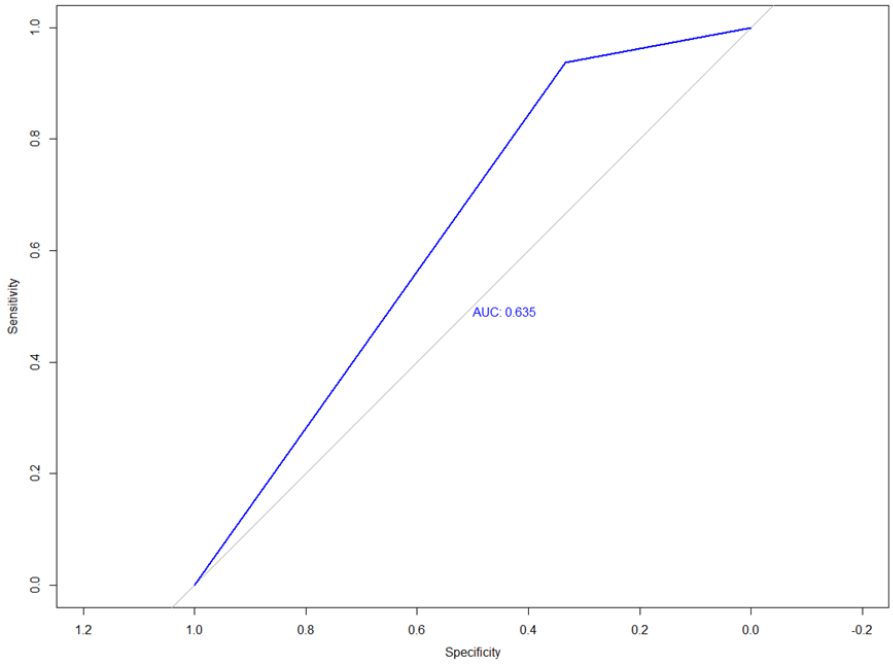


Figure 2.6 ROC-AUC of k-NN (Binary)

2.1.5 Support Vector Machines

Support Vector Machines (SVM) is considered a good supervised machine learning model for classification problems because it produces significant accuracy with less computation power. Its use can also be extended to multi-class scenarios, thus making it applicable to this project so that there is basis for comparison. Each data item is plotted as a point in n-dimensional space (where n = number of features) with the value of each feature being the value of a particular coordinate. SVs are co-ordinates of each individual observation. SVM classifier is a boundary that segregates the 2 classes (hyperplane). Thus, classification is performed by finding the

[8] See Appendix C.

hyper-plane in the n-dimensional space that distinctly differentiates the 2 classes and data points.

The SVM with linear kernel before any hyperparameters tuning score 57.7% in recall and a ROC-AUC of 0.695. The SVM method can handle both linear and non-linear class boundaries. A non-linear separation boundary is generally true of real-life data. The most commonly used kernel transformations are polynomial kernel and radial kernel for SVM. These 2 variations of SVM model are also trained. Recall scores for these 2 models are 57.7% for SVM polynomial kernel and 51.9% for SVM radial kernel.

The hyperparameters are then tuned for SVM linear kernel using the default grid search routine specified. The best resulting SVM[9] scores 69.2% in recall, and a ROC-AUC of 0.689 which is comparable to that of the boosted classification trees.

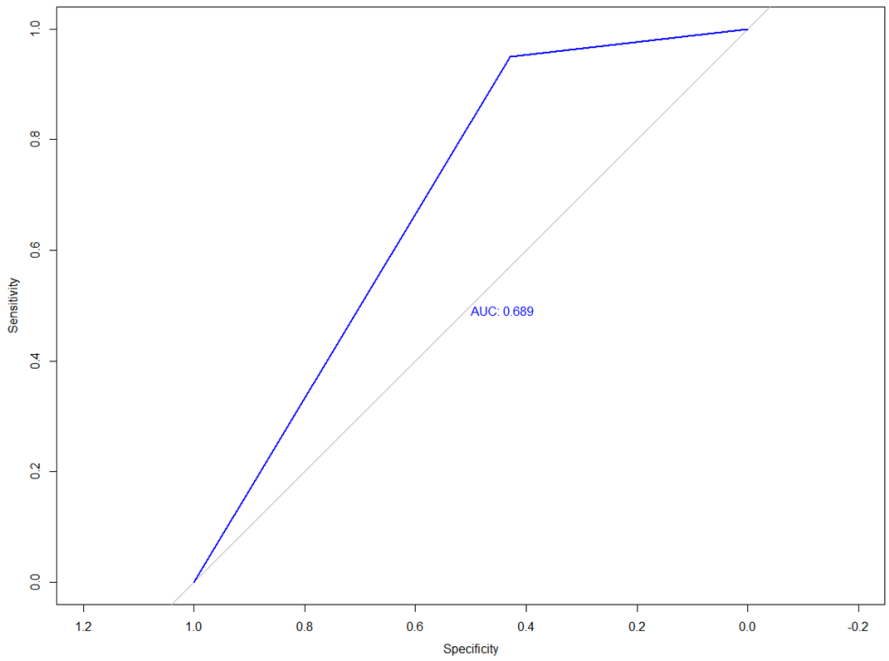


Figure 2.7 ROC-AUC of SVM (Binary)

2.1.6 Naïve Bayes

The primary assumption of Naïve Bayes (NB) classifier is that all the variables in the dataset are not correlated (i.e. independent). The NB classifier functions by simplifying the calculations of the probabilities for every variable and selects the outcome with highest

[9] See Appendix D.

probability. Its simplistic assumption and calculations makes it a simple and popular model to employ in classification problems; but render the model a bad estimator so its main use is to derive the base accuracy of the dataset, instead of taking the probability outputs seriously. This model is undertaken for the project because it can be applied for binary and multi-class classification problems, thus providing a basis for comparison.

While the NB models trained[10] present high accuracy for the test set (83 – 86%), the recall metric scores very low between 15 – 44%. The better NB model scoring 44% recall is trained using the *klaR* package and user defined *expand.grid()* function for tuning the hyperparameters.

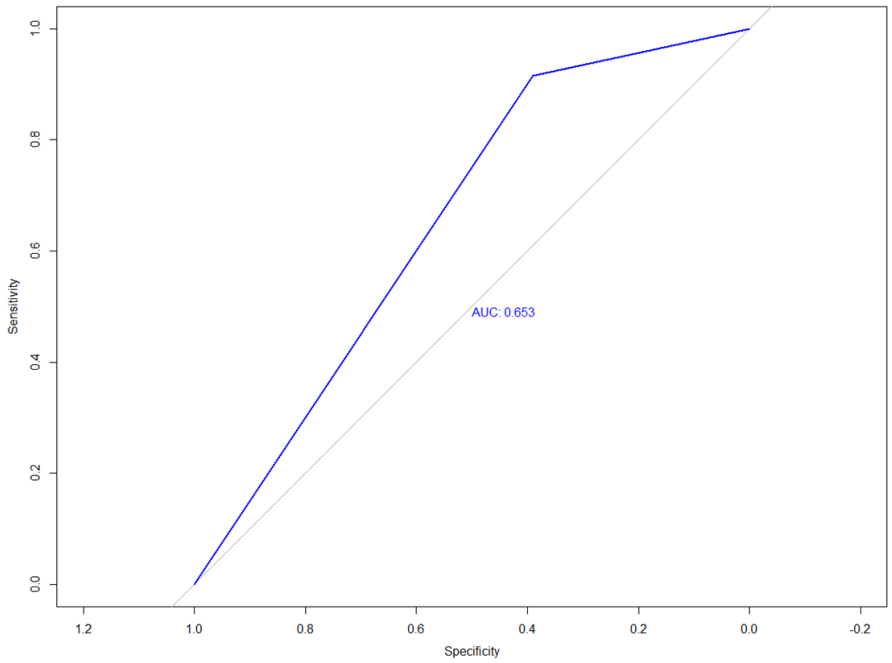


Figure 2.8 ROC-AUC of NB (Binary)

[10] See Appendix D for the tuning parameters of the best resulting NB model in terms of recall. The models are trained using *'nb'* and *'naïve_bayes'* methods within 2 R packages, *klaR* and *naivebayes*.

2.2 Validation criteria

The above discussion has already briefly discussed 2 evaluation metrics the project focuses on – Recall and ROC-AUC. In medical diagnosis, recall (or, sensitivity) is the ability of a test to correctly identify those with the disease (true positive rate). Whereas, test specificity is the ability of the test to correctly identify those without the disease (true negative rate). As the dataset experiences class imbalance (i.e. majority of patients do not have T2D), it is highly likely that the models will correctly predict this majority class (in confusion matrix context, true negative).

Greater emphasis is placed on minimising the number of false negatives (predict the patient has diabetes but actually does not) in a medical context. From a prediction standpoint, it is crucial that the model selected for implementation can maximise the predictions that accurately identify patients with T2D and doctors can provide timely advice and treatment. The recall metric takes these into account and is thus selected. The prediction probability threshold value is set at 0.5. The balance between precision and recall is conveyed in the F1 score. It is a measure of a test's accuracy.

ROC-AUC is a threshold-independent metric that illustrates the trade-off between recall and specificity. A larger AUC indicates a better overall performance across all thresholds (Ghanvatkar and Rajan, 2019; Harutyunyan et al, 2019). The ROC-AUC metric is also considered to be more informative when faced with highly skewed datasets. ROC curves are generated based on the predicted outcome and true outcome. The AUC for the test data sets are calculated and used to compare the discriminative power of the different models. This curve plots two parameters – TP rate (TPR/Recall) and FP rate (FPR):

$$\text{TPR} = \frac{TP}{TP+FN} \qquad \text{FPR} = \frac{FP}{FP+FN}$$

where, TP, FP, TN and FN represent the number of true positives, false positives, true negatives and false negatives respectively.

Another threshold-independent metric PR-AUC that illustrates the trade-off between precision and recall is also employed in the model evaluation phase. This metric is suitable for imbalanced dataset as it focuses on the correct prediction of the minority class (true and false positives). It is typically used to evaluate the results of k-fold cross-validation.

Thus, Table 2.2 summarises the formulae of the metrics used to evaluate the models' performances:

1. **Test dataset**: The project employs the 70-30 train/test dataset split. Validation using test dataset avoid the potential bias of the performance estimate due to overfitting of the model to training data sets. The model is also predicted using the train set to compare model accuracy and other evaluation metrics to those of test set as a form of validation.
2. **ROC-AUC**
3. **PR-AUC**
4. **Confusion matrix**: Evaluation measures derived from the confusion matrix include sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) are calculated based on the following formulae:
    a. **Sensitivity/*Recall* = TPR**
    b. **F1 score**
    c. Specificity = FPR
    d. PPV = $\frac{TP}{TP+FP}$
    e. NPV = $\frac{TN}{TN+FN}$

    Recall is the main metric used for model evaluation.

5. **Sensitivity analysis and robustness tests**: Given the project objective of proposing significant features, different subsets of variables are processed to ensure that the features deemed important by the classifiers remain consistent or if its information value has been lost in categorising it. For example, using continuous BMI values and original BMI cut-off points, *BMXBMIO* instead of *BMXBMIAA*. There are also other variations of variables described in Appendix A that can be used in replacement of another. For example, categorical variable *SMQ915A* is employed instead of the continuous variable, *SMQ915*. This will be further discussed in Chapter 2.4.

Table 2.2 Definition and Formula for Evaluation Metrics

Metric and Definition	Formula
Recall (Sensitivity) Measures proportion of correctly identified True Positives among actual positives	$\frac{TP}{TP + FN}$
Area Under Curve – Receiver Operating Characteristics (AUC-ROC) Plot of Recall against (1 – Specificity) for different threshold settings	$\frac{Recall}{1 - Specificity}$ Specificity = $\frac{FP}{TN+FP}$
Precision Measures proportion of correctly identified True Positives among predicted positives	$\frac{TP}{TP + FP}$
Area Under Curve Precision Recall (AUC PR) Plot of Precision against Recall for different thresholds settings.	$\frac{Precision}{Recall}$ Precision = $\frac{TP}{TP+FP}$
F1 Score Measures the balance between Precision and Recall	$2 \text{ x } \frac{Precision * Recall}{Precision + Recall}$

2.2.1 Evaluation of models

Table 2.3 summarises the 3 evaluation metrics calculated for each model. The top 3 models for the binary prediction of T2D are boosted variations – XGB, RF and Logitboost, all scoring at least 80.0%. The overall best model is RF as it also illustrates the highest ROC-AUC at 0.705. These models are thus employed for sensitivity analysis and robustness tests.

Table 2.3 Summary of Model Evaluation Ranked Based on Recall

Rank	Model	Recall (%)	ROC-AUC	PR-AUC	True Positives	False Positives	False Negatives	True Negatives
1	**XGB trees**	**82.7***	0.650	0.449	43	9	86	264
2	RF	80.8	**0.705***	0.471	42	10	53	297
3	Logitboost	80.0	0.609	0.445	36	9	95	201
4	Adaboost trees	73.1	0.689	0.473	38	14	69	281
5	SVM	69.2	0.689	0.479	36	16	48	302
6	GBM trees	69.2	0.655	0.469	36	16	54	296
7	DT	67.3	0.670	0.475	35	17	54	296
8	k-NN	63.5	0.635	0.469	33	19	66	284
9	NN	50.0	0.627	0.483	26	26	52	298
10	LogR	48.1	0.308	0.330	25	27	318	32
11	NB	44.2	0.653	**0.495***	23	29	36	314

Figure 2.9 illustrates the top 3 ROC-AUC where RF, Adaboost and SVM are ranked as such. Technically DT is within the top 4 ranked ROC-AUC but as it is clear that the classification trees ranked highly, another model variation with the next highest ROC-AUC illustrated in Figure 2.9 is the NB classifier.

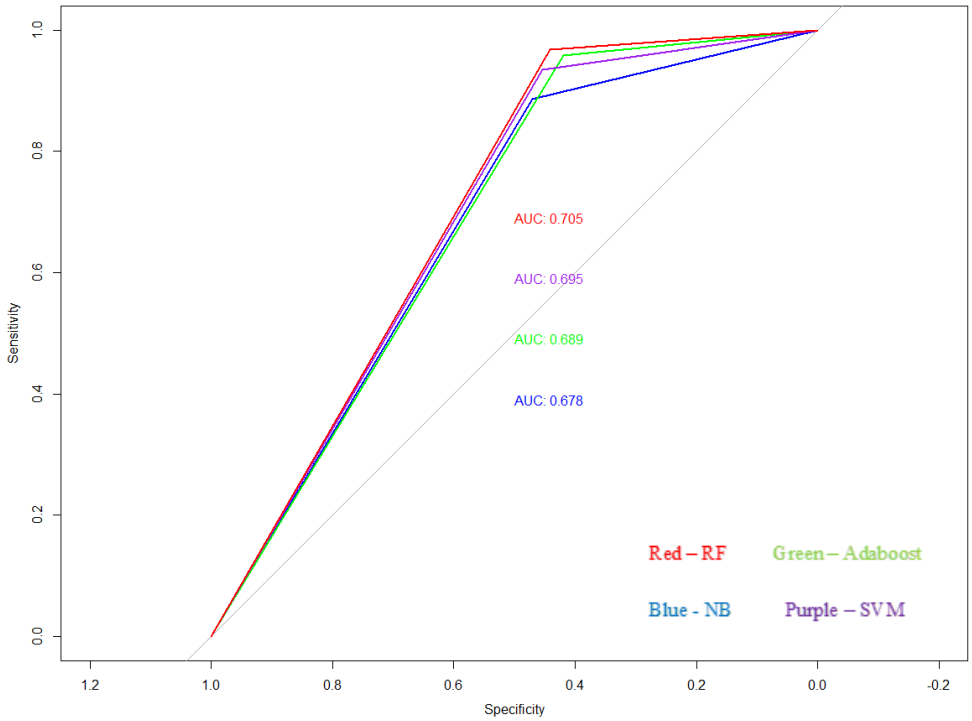


Figure 2.9 Top 4 ROC-AUC (Binary)

Figure 2.10 illustrates the top 4 PR-AUC. 10 models, excluding LogR, present very similar results for the PR-AUC within a close range of 0.445 to 0.495 (insignificant difference of 0.05).

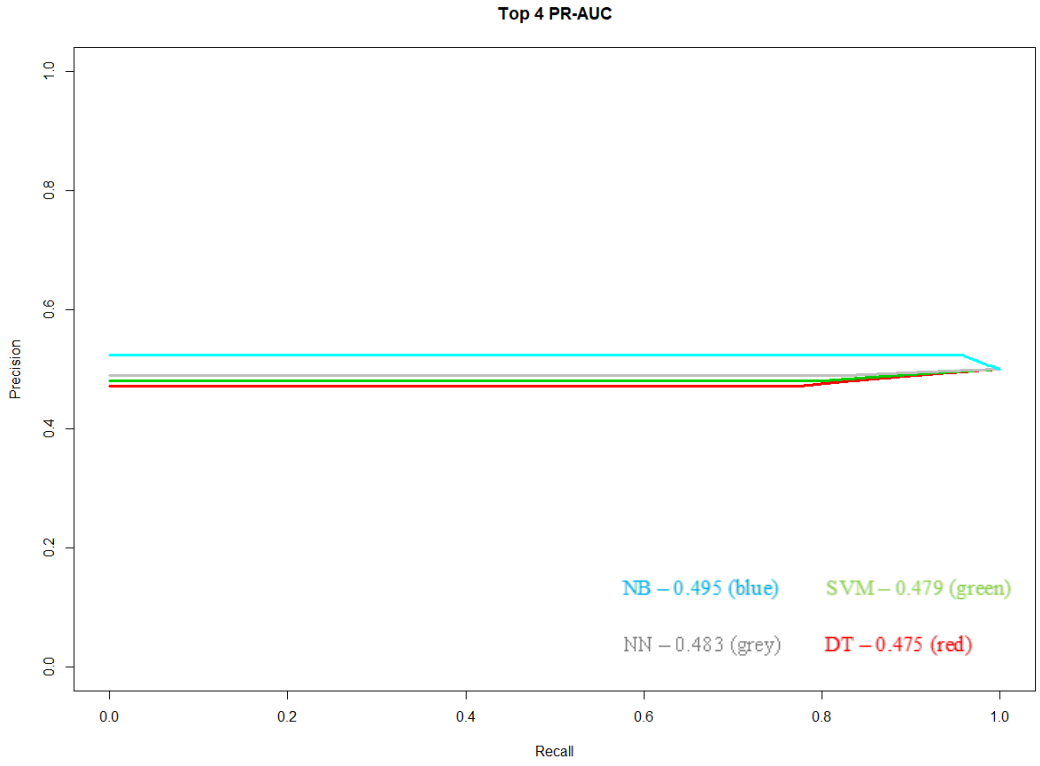


Figure 2.10 Top 4 PR-AUC (Binary)

2.2.2 Benchmarking

While this project employing the latest available NHANES dataset (2013 to 2016) is unique in its study, there are existing research using past NHANES datasets for the prediction of T2D. There is only 1 existing paper that may be used for benchmarking given the context of its study.

Semerdjian and Frank (2017) applied an ensemble classifier to a binary classification prediction of T2D using the 1999-2004 NHANES survey data. The target variable in the study is measured in the same method to the target variable *DIQ010A* in this project. The models trained with 16 features in the study are logistic regression, k-NN, random forest, gradient boosting, SVM, NB (tried but not reported) and an ensemble classifier. The models are evaluated based on ROC-AUC and recall rate. The top performing model, gradient boosting classifier scores a ROC-AUC of 0.84. RF is their next best performing model and k-NN is the worst performer. The recall rate for their best classifier is 35% at a decision boundary of 0.5.

The results of Semerdjian and Frank (2017) study is similar to the results reported in Table 2.3 where the same 2 models, RF and XGB trees perform the best. Table 2.3 reports that RF classifier in this project performs the best based on ROC-AUC, scoring 0.705. The best recall in this project is reported for XGB trees at 82.7%. Given the dataset in this project include over 50 features, performance of the models in this project is comparative.

Another study by Nelson et al. (2002) using the NHANES survey data collected between 1988 and 1994, researched on diet and exercise practices of American adults with T2D by applying multivariate logistic regression. The study reports on the statistical significance of the nutritional and exercise factors and does not provide a similar basis for benchmarking. The conclusion drawn in this study is that more than 30% participants with T2D reported no regular physical activity, and another 38% reported insufficient levels of physical activity. The conclusion supports the results found in this project where daily physical activity levels are significant in the T2D prediction.

2.3 Features importance

One of the project's objectives is to discern if a specific lifestyle has an impact on T2D prediction. For example, work activity level, recreational activity level, smoking and alcohol habits. This may allow insight into better understanding one's lifestyle, determining one's risk of developing T2D. The results for above 11 models aim to provide validation towards the importance of these factors that the population, doctors and healthcare professionals may find useful in reducing the T2D risk.

The *varImp()* evaluation function is employed for this task. This function is not applied to the logistic regression and decision tree models. This is because of the inherent characteristics of these 2 models that already allow for the easy identification of significant features.

The interpretation of logistic regression results[11] is to compare the p-values at a significance level (denoted as α), usually at 0.05. A significance level of 0.05 indicates a 5% risk of concluding that an association exists when there is no actual association. If the p-value is less than or equal to α, it can be concluded that there is a statistically significant association between that pair of dependent and independent variables.

The importance of a feature in decision tree is computed as the total reduction of the Gini index. Gini index, or mean decrease in impurity (MDI), calculates each feature importance as the sum over the number of splits (across all tress) that include the feature, proportionally to the number of samples it splits. The output of tree model clearly displays key indicators[12].

Results of the models that employs *varImp()* evaluation function are calculated in 2 ways. How each feature contributes to the model is estimated by:

1. **XGB, GBM, RF and NN**: Model-based calculation approach using the model information where the advantage lies in it being more closely tied to the model performance
2. **Logitboost, Adaboost, SVM, k-NN and NB**: Models that do not have a specific way to estimate importance use calculations from ROC-AUC analysis conducted on each feature

All measures of importance are scaled to have a maximum value of 100. Table 2.4 below summarises the results of the top 20 important features resulted in each model (excluding DT

[11] See Appendix E for the results of the stepwise logistic regression. The results are also summarised in Table 2.4.
[12] See Appendix F for the output of DT visualized which is also summarised in Table 2.4

and LogR). There are 5 features that consistently express importance across all methods of estimation:

i. *RIDAGEYR*, age of the participant
ii. *BPQ090D2*, where the participant has been told to take prescription for cholesterol
iii. *PAQ710*, measuring sedentary lifestyle by average hours spent sitting and watching TV or videos in a day over the last 30 days
iv. *LBXTC*, total cholesterol level

There are 3 features that express importance across most models (bar 1 method of estimation):

i. *WHQ150*, record of the participant's age at their heaviest weight
ii. *BPXSYave*, systolic blood pressure reading
iii. *PAD680*, measuring sedentary lifestyle by average minutes in a day where a participant is spent sitting (such as sitting at school, at home, getting to and from places, or with friends including time spent sitting at a desk, traveling in a car or bus, reading, playing cards, watching television, or using a computer; do not include time spent sleeping)

These results are also consistent to the results of feature selection phase using NN and k-NN[13]. The significance of age, cholesterol levels and blood pressure readings in T2D prediction is considerable and often highlighted in existing literature (Heredia-Langner et al., 2013; Pimental et al., 2018).

Given that research also shows that the risk of developing T2D is related to weight and BMI, it is sensible that a sedentary lifestyle is a contributing factor. There are also features that accounts for physical activity documented in Table 2.4. For example, *PAD675* measuring minutes spent doing moderate-intensity sports, fitness or recreational activities daily, is consistently significant in the results of boosted trees.

[13] See Appendix C for results of feature selection phase and Appendix G for the final results.

Table 2.4 Significant Features – Binary Classification

Variables	XGB	RF	GBM	DT	NN	Logitboost, Adaboost, SVM, kNN, NB (ROC)	LogR
ALQ120QU	13		13				
ALQ130			20				
BMDAVSAD	10	7	8			9	***
BMXARMC	19	18					*
BMXARML		13					
BMXBMIAA					14		
BMXHT		16					
BMXLEG		15					*
BMXWAIST	11	11	9			7	***
BMXWT		14					*
BPQ020	17					8	
BPQ080						16	
BPQ090D	2	6	2	3	4	2	***
BPXDIave	12	9	12				
BPXPULS							***
BPXSYave	5	4	7		13	3	*
DMDEDUC2						17	.
INDFMIN2	18		17	5		14	***
LBDHDD	9	10	14			13	*
LBDLDL	20	19				20	
LBXTC	4	3	3	2	10	11	***
LBXTR	8	20	15			19	***
MCQ092					12		**
MCQ160B							**
MCQ160M					18		**
MCQ160N							.
MCQ203							*
MCQ300C	6	12	5	4		10	***
MCQ365A			18		15	18	***
MCQ365B			16		17	12	**
MCQ365C						15	**
MCQ365D	15		6		5	6	**
MCQ370A					3		
MCQ370B							.
MCQ370C					11		
MCQ370D					8		
PAD615				7	9		

PAD630							**
PAD645							.
PAD675	7	17	10				***
PAD680	16	8	19	8	2		*
PAQ710	3	2	4	6	19	5	***
RIDAGEYR	1	1	1	1	1	1	***
SMQ020							*
SMQ040A					7		
SMQ910					20		
WHQ070					16		
WHQ150	14	5	11		6	4	***

Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The results of top 2 ranking models – XGB trees and Random forest – are illustrated in Figures 2.11 and 2.12 respectively. The importance of these features is also calculated using the model-based approach which lends support to validation of the features. Results from both models show that medical history (*MCQ160*) and lifestyle factors do indeed play a part in T2D diagnosis risk:

i. the activity levels derived from work or leisure (*PAD/PAQ*)
ii. doctors' prior advice for dietary changes and exercise (*MCQ365*)
iii. the participant's efforts in lifestyle changes (*MCQ370*)
iv. the frequency of smoking (*SMQ020*) and alcohol intake (*ALQ120, ALQ141*)

The results are subject to the sensitivity analysis and robustness tests discussed below, where features are substituted with variations that provide different measures to capture similar meaningfulness. This is to ensure that the results presented in Table 2.4 are meaningful and statistically consistent in its association to the T2D diagnosis.

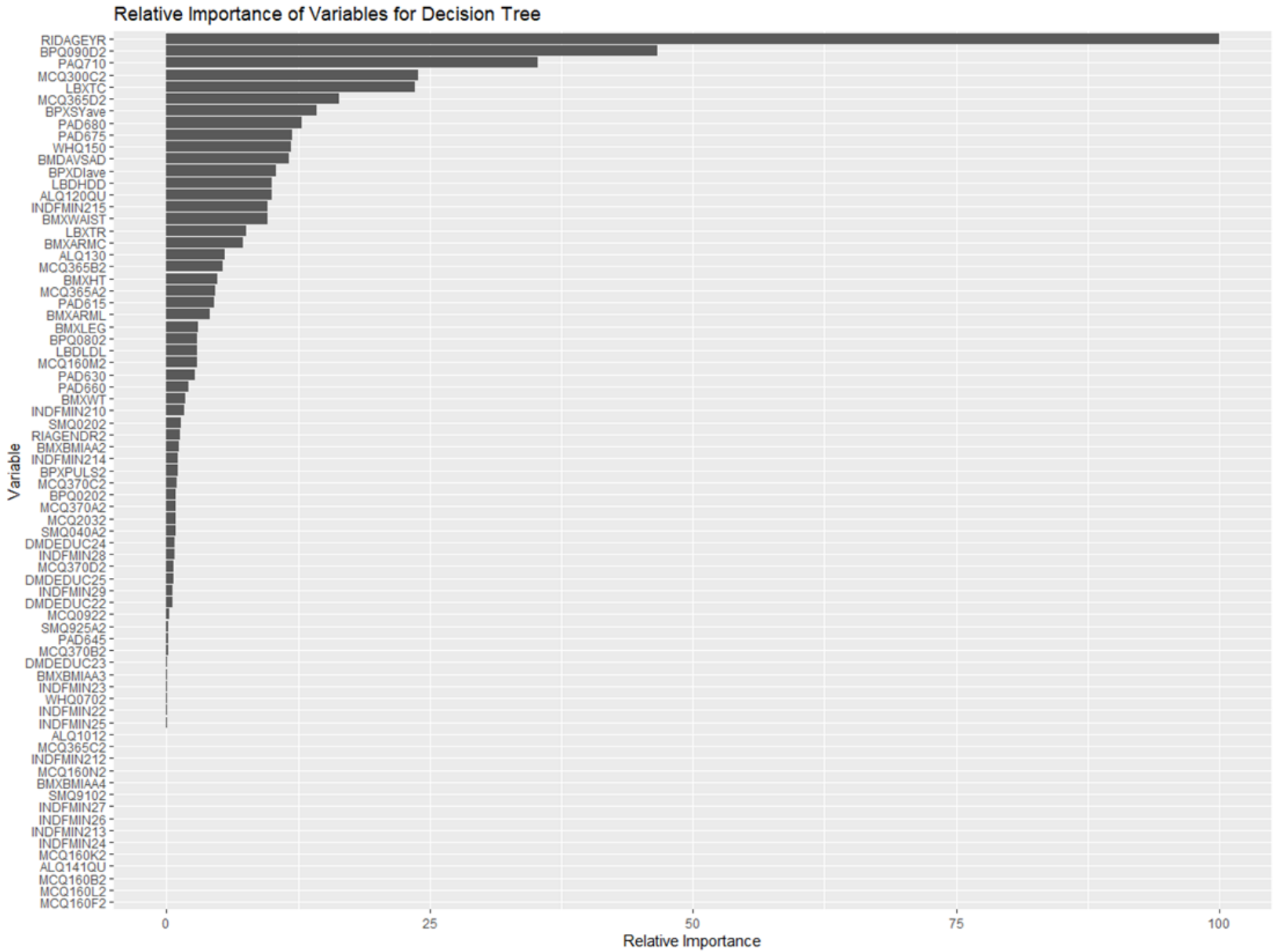


Figure 2.11 Significant Features of XGB Trees (Binary)

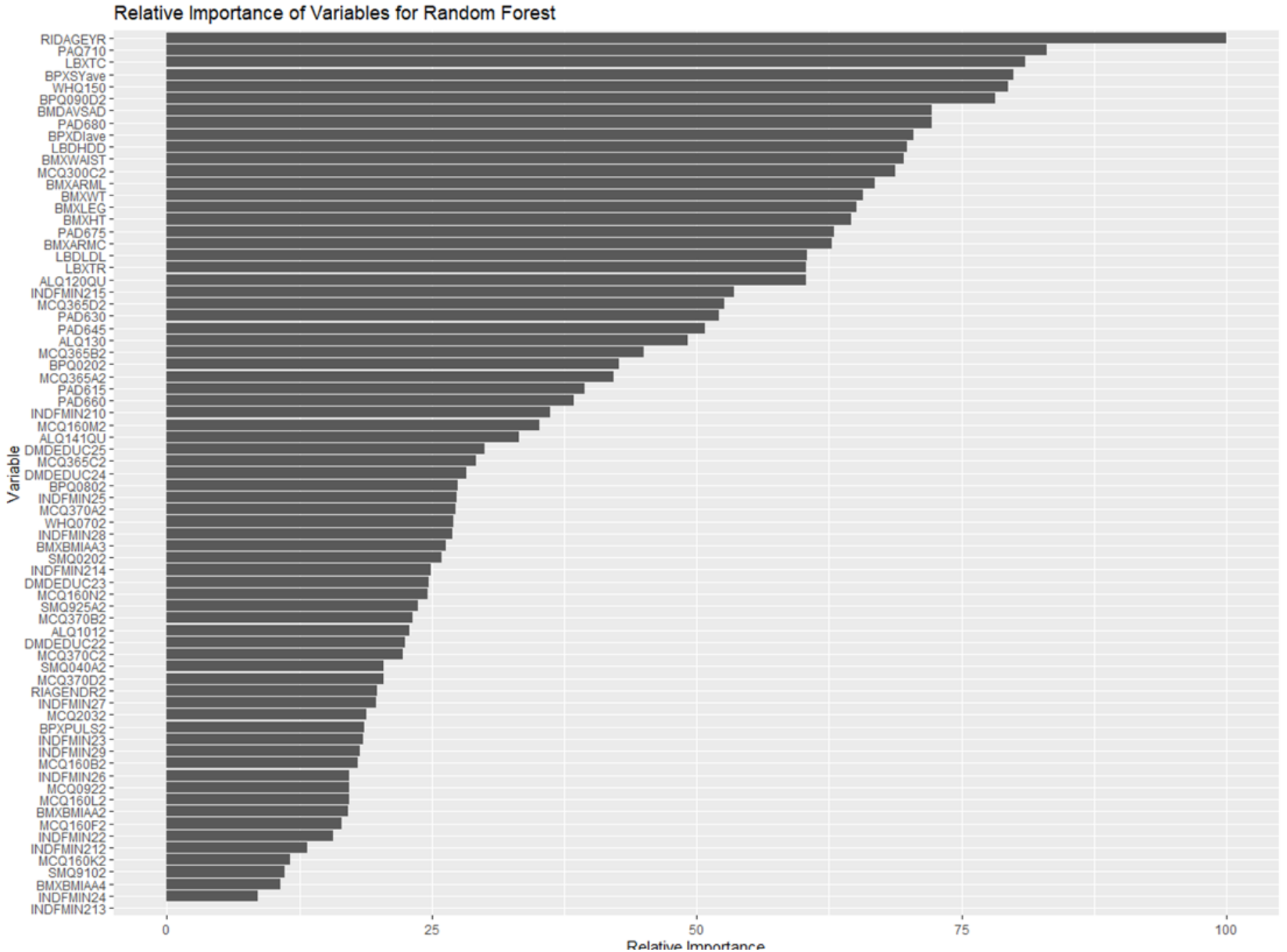


Figure 2.12 Significant Features of RF (Binary)

2.4 Sensitivity analysis and robustness tests

Additional tests are carried out to examine the robustness of the main results. These tests include training the top performing models evaluated above (RF, XGB trees and Logitboost) with different subsets of features. The project tests the robustness of the models with the following methods:

i. alternative model specifications, specifically hyperparameter tuning of the models which have already been presented above
ii. include laboratory data features to the main dataset
iii. features substitution/exclusion

2.4.1 Laboratory data

Since the NHANES data provides laboratory test results on the participants, the data can be leveraged to improve the performance on the classification. The NHANES dataset includes features derived from laboratory analysis of biological specimens (biospecimens). These biospecimens, inclusive of blood and urine, provide even more detailed information about participants' health and nutritional status. It may suggest improved predictive accuracy of T2D diagnosis, but these laboratory tests are not common and not readily available to most patients. The first 2 rounds of feature selection[14] show that some of these features may be significant in the T2D prediction. While some are consistent with existing literature findings like cholesterol and gamma glutamyl transferase, some are interesting to note like calcium, potassium and phosphorus. However, given the motivation for this project is to suggest lifestyle and easy-to-collect factors that can indicate T2D before having to run the actual T2D tests, the main dataset excludes the laboratory test features.

a. Cholesterol (*LBXTC*, *LBXTR*, *LBDHDD*, *LBDLDL*)

These variables are included in the main dataset because of its well-known impact in a T2D diagnosis and in a patient's health. Cholesterol readings are considered common in determining one's overall health. Table 2.4 shows that the 4 cholesterol features are consistently significant across all models. The robustness tests include running a subset of the main dataset that

[14] See Figure 1.1 and Appendix C.

excludes these 4 features. This is to check the importance of other features in the dataset given that results of the cholesterol features in Table 2.4 show corroboration with existing literature.

Table 2.5 compares if the updated models perform better in terms of recall, ROC-AUC and confusion matrix values to the original models trained using the full dataset. The RF model is consistent in its performance for recall metric. It may allow for comparison of important features derived from this model. It is interesting to note that the boosted logistic regression model improves in its predictions which may imply that cholesterol features may experience multi-collinearity with some of the dataset. However, ROC-AUC of the models are not as good which likely points to the overall significance of cholesterol in the T2D prediction.

Table 2.5 Model Comparison – Exclude Cholesterol (Binary)

Model	Recall (%)	ROC-AUC	TP	FP	FN	TN
XGB trees	67.3	0.643	35	17	67	283
	82.7	0.650	43	9	86	264
RF	80.9	0.673	41	11	66	284
	80.8	0.705	42	10	53	297
Logitboost	93.9	0.604	46	3	157	128
	80.0	0.609	36	9	95	201

original scores in grey

By eliminating cholesterol features, the features deemed important by these 3 models still substantiates the original list in Table 2.4. The list in Table 2.6 includes 3 features that previously only showed significance in the NN and LogR results:

i. *MCQ370C*, where it records if the participant is reducing salt in diet
ii. *PAD630*, measuring minutes in a day where the participant does moderate-intensity activities at work
iii. *SMQ020*, where it records if the participant smoked at least 100 cigarettes in life

The results continue to support the evidence that participants at risk of T2D may improve upon their lifestyle with a healthier diet, moderate-intensity movement in a day be it work or recreational and regulate their smoking habit.

Table 2.6 Significant Features – Exclude Cholesterol (Binary)

Variables	RF		XGB		Logitboost (ROC)	
ALQ120QU	16		11	13		
ALQ130	20		16			
BMDAVSAD	6	7	14	10	9	9
BMXARMC	17	18	15	19		
BMXARML	13	13	19			
BMXHT	15	16	18			
BMXLEG	11	15	20		17	
BMXWAIST	8	11	6	11	7	7
BMXWT	12	14			19	
BPQ020				17	8	8
BPQ080					14	16
BPQ090D	4	6	2	2	2	2
BPXDIave	10	9	5	12		
BPXSYave	3	4	9	5	3	3
DMDEDUC2					15	17
INDFMIN2	18		13	18	12	14
MCQ300C	9	12	4	6	10	10
MCQ365A					16	18
MCQ365B			17		11	12
MCQ365C					13	15
MCQ365D			7	15	6	6
MCQ370C					18	
PAD630	19					
PAD675	14	17	8	7		
PAD680	7	8	12	16		
PAQ710	2	2	3	3	5	5
RIDAGEYR	1	1	1	1	1	1
SMQ020					20	
WHQ150	5	5	10	14	4	4
LBDHDD		10		9		13
LBDLDL		19		20		20
LBXTC		3		4		11
LBXTR		20		8		19

original scores in grey

Another robustness test is to replace the continuous variables of *LBXTC*, *LBDHDD*, *LBXTR* and *LBDLDL* with their categorical variations which aim to better reflect if the participant has high cholesterol, high levels of bad cholesterol or good cholesterol:

1. *LBXTCA*, where 1 indicates that the participant total cholesterol level (>=200) is at risk, 2 otherwise
2. *LBDHDDA*, where 1 indicates that the participant HDD cholesterol level (<60) is at risk, 2 otherwise
3. *LBXTRA*, where 1 indicates that the participant Triglyceride level (>=150) is at risk, 2 otherwise
4. *LBDLDLA*, where 1 indicates that the participant LDL cholesterol level (>=100) is at risk, 2 otherwise

Table 2.7 shows that the models overall performance is not as strong as that of the main models. The cholesterol features are fairly significant in the prediction and the results in Table 2.8 also show that the models do not pick up on that as well as the continuous variables. In comparison to all 4 original variables showing significance across the 3 models, 2 out of 4 categorical variables are only picked up in the XGB trees model.

Table 2.7 Model Comparison – Categorical Cholesterol (Binary)

Model	Recall (%)	ROC-AUC	TP	FP	FN	TN
XGB trees	76.9	0.664	40	12	68	282
	82.7	0.650	43	9	86	264
RF	78.9	0.679	41	11	63	287
	80.8	0.705	42	10	53	297
Logitboost	73.1	0.660	38	14	62	288
	80.0	0.609	36	9	95	201

original scores in grey

Table 2.8 Significant Features – Categorical Cholesterol (Binary)

Variables	RF		XGB		Logitboost (ROC)	
ALQ120QU	17		15	13		
ALQ130	20		12			
BMDAVSAD	8	7		10	7	9
BMXARMC	16	18	13	19		
BMXARML	12	13	19			
BMXHT	13	16				
BMXLEG	10	15			19	
BMXWAIST	7	11	6	11	5	7
BMXWT	11	14				
BPQ020				17	8	8
BPQ080					9	16
BPQ090D	5	6	2	2	2	2
BPXDIave	9	9	7	12		
BPXSYave	3	4	8	5	4	3
DMDEDUC2					13	17
INDFMIN2			16	18	15	14
LBDHDDA		10		9		13
LBDLDLA		19	20	20		20
LBXTCA		3	17	4		11
LBXTRA		20		8		19
MCQ300C	15	12	9	6	12	10
MCQ365A					16	18
MCQ365B			14		11	12
MCQ365C					14	15
MCQ365D	19		11	15	10	6
MCQ370C					20	
PAD630	18					
PAD645			18			
PAD675	14	17	10	7		
PAD680	6	8	4	16	18	
PAQ710	4	2	3	3	6	5
RIDAGEYR	1	1	1	1	1	1
SMQ020					17	
WHQ150	2	5	5	14	3	4

original scores in grey

b. High sensitivity C-reactive protein (*HSCRP*)

While existing literature recognises its impact in T2D diagnosis (Ford, 1999, Pradhan et al., 2001), the project is unable to include this feature because this variable is only collected in the 2015-2016 cycle.

c. Gamma glutamyl transferase (*LBXSGTSI*)

This variable significance in predicting T2D is documented in existing literature and widely accepted (André et al., 2006, Lee et al., 2003, Sabanayagam et al., 2009, Wen et al., 2010).

d. Calcium (*LBXSCA*)
e. Potassium (*LBXSKSI*)
f. Phosphorous (*LBXSPH*)

The robustness tests include running an extended subset of the main dataset by including these 4 features – *LBXSGTSI, LBXSCA*, *LBXSKSI* and *LBXSPH*. Table 2.9 shows that the updated models perform better in terms of ROC-AUC to the original models trained using the full dataset.

Table 2.9 Model Comparison – Laboratory Data (Binary)

Model	Recall (%)	ROC-AUC	TP	FP	FN	TN
XGB trees	67.3	0.661	35	17	58	292
	82.7	0.650	43	9	86	264
RF	78.9	0.710	41	11	49	301
	80.8	0.705	42	10	53	297
Logitboost	78.6	0.639	33	9	77	225
	80.0	0.609	36	9	95	201

original scores in grey

Table 2.10 notes that these 4 laboratory data features are significant, consistent to the results of feature selection. 2 of the variables are picked up across all 3 models, including Gamma glutamyl transferase supporting results in existing literature.

Table 2.10 Significant Features – Laboratory Data (Binary)

Variables	RF		XGB		Logitboost (ROC)	
ALQ120QU			8	13		
BMDAVSAD	8	7	11	10	6	9
BMXARMC		18		19		
BMXARML	14	13	10			
BMXHT	18	16				
BMXLEG	10	15				
BMXWAIST	12	11	7	11	7	7
BMXWT	15	14				
BPQ020				17	11	8
BPQ080					12	16
BPQ090D	3	6	2	2	2	2
BPXDIave	13	9	9	12		
BPXSYave	6	4	5	5	5	3
DMDEDUC2					16	17
INDFMIN2				18	14	14
LBDHDD	11	10	16	9	15	13
LBDLDL	19	19		20		20
LBXSCA	16		18			
LBXSGTSI	9		6		9	
LBXSKSI	17		19		19	
LBXSPH	20					
LBXTC	5	3	4	4	8	11
LBXTR		20	13	8	20	19
MCQ300C		12	14	6	18	10
MCQ365A						18
MCQ365B					13	12
MCQ365C					17	15
MCQ365D				15	10	6
PAD645			20			
PAD675		17	12	7		
PAD680	7	8	17	16		
PAQ710	2	2	3	3	4	5
RIDAGEYR	1	1	1	1	1	1
WHQ150	4	5	15	14	3	4

original scores in grey

2.4.2 Physical activity

This robustness test is to replace some features that study a participant's **daily** physical activity captured in minutes: *PAD615, PAD630, PAD645, PAD660* and *PAD675*. These features are replaced with alternatives that the NHANES questionnaire consists of by capturing the participant's **weekly** records for the same activities: *PAQ605, PAQ620, PAQ635, PAQ650* and *PAQ665*. The definitions of these variables are listed in Appendix A.

The robustness tests undertaken for these variables include the NN model as the significance of feature *PAD615* (measuring vigorous work activity) is captured in the model. The results in Table 2.11 show that the models perform much more poorly in terms of recall. This is supported by the results in Table 2.12 where the corresponding features measuring weekly records are not significant. This perhaps signal the importance of daily movement and consistency in activity levels.

Table 2.11 Model Comparison – Physical Activity (Binary)

Model	Recall (%)	ROC-AUC	TP	FP	FN	TN
XGB trees	55.8	0.717	29	23	29	321
	82.7	0.650	43	9	86	264
RF	73.1	0.707	38	14	45	305
	80.8	0.705	42	10	53	297
Logitboost	54.5	0.714	24	20	33	292
	80.0	0.609	36	9	95	201
NN	55.8	0.650	29	23	49	301
	50.0	0.627	26	26	52	298

original scores in grey

Table 2.12 Significant Features – Physical Activity (Binary)

Variables	XGB		RF		NN		Logitboost (ROC)	
ALQ120QU	8	13	19					
BMDAVSAD	14	10	10	7	13		7	9
BMXARMC	16	19	20	18				
BMXARML	15		14	13	20			
BMXBMIAA						14		
BMXHT	18		16	16				
BMXLEG	13		15	15				
BMXWAIST	11	11	9	11	8		6	7
BMXWT	20		13	14				
BPQ0202		17			19		8	8
BPQ080							13	16
BPQ090D	3	2	4	6	2	4	2	2
BPXDIave	6	12	8	9	10			
BPXSYave	17	5	6	4	5	13	3	3
DMDEDUC2					16		16	17
INDFMIN2		18					14	14
LBDHDD	10	9	12	10			15	13
LBDLDL	12	20	18	19			19	20
LBXTC	5	4	7	3	7	10	11	11
LBXTR	19	8	17	20			18	19
MCQ092					4	12		
MCQ160B					12			
MCQ160K					6			
MCQ160M						18		
MCQ160N					14			
MCQ203					9			
MCQ300C	7	6	11	12			9	10
MCQ365A						15	20	18
MCQ365B						17	10	12
MCQ365C							17	15
MCQ365D		15				5	12	6
MCQ370A					18	3		
MCQ370C						11		
MCQ370D						8		
PAD615						9		
PAQ665					17			
PAD675		7		17				
PAD680	9	16	5	8		2		

PAQ710	2	3	2	2		19	5	5
RIAGENDR					15			
RIDAGEYR	1	1	1	1	1	1	1	1
SMQ040A					11	7		
SMQ910						20		
WHQ070						16		
WHQ150	4	14	3	5	3	6	4	4

original scores in grey

2.4.3 Smoking

The first 2 rounds of feature selection show that a participant use of smokeless tobacco may impact one's T2D diagnosis. *SMQ910* is included in the main dataset where it captures whether a participant has used smokeless tobacco before.

This robustness test replaces *SMQ910* with *SMQ915A* to test if recent use of smokeless tobacco in the last 30 days will be a better predictor. The robustness tests undertaken for these variables include the NN model as the significance of feature *SMQ910* is captured in the model. The results in Table 2.13 show that the models perform more poorly in terms of recall. This is supported by the results in Table 2.14 where *SMQ915A* is not significant in the prediction.

Table 2.13 Model Comparison – Smoking (Binary)

Model	Recall (%)	ROC-AUC	TP	FP	FN	TN
XGB trees	55.8	0.741	29	23	24	326
	82.7	0.650	43	9	86	264
RF	61.5	0.733	32	20	29	321
	80.8	0.705	42	10	53	297
Logitboost	71.4	0.611	30	12	67	239
	80.0	0.609	36	9	95	201
NN	63.5	0.646	33	19	60	290
	50.0	0.627	26	26	52	298

original scores in grey

Table 2.14 Significant Features – Smoking (Binary)

Variables	XGB		RF		NN		Logitboost (ROC)	
ALQ120QU	14	13	19		13			
BMDAVSAD	5	10	6	7			4	9
BMXARMC		19	17	18				
BMXARML			18	13				
BMXBMIAA					5	14		
BMXHT			15	16				
BMXLEG	19		11	15			18	
BMXWAIST	11	11	10	11			5	7
BMXWT			14	14				
BPQ020		17			14		12	8
BPQ080							8	16
BPQ090D	2	2	5	6		4	3	2
BPXDIave	9	12	9	9				
BPXSYave	12	5	7	4		13	7	3
DMDEDUC2					12		11	17
INDFMIN2		18			8/10/15/18			14
LBDHDD	8	9	8	10			6	13
LBDLDL	13	20	16	19	6		17	20
LBXTC	3	4	4	3		10	14	11
LBXTR	17	8		20				19
MCQ092						12		
MCQ160M	15				7	18		
MCQ160N					20			
MCQ203					9			
MCQ300C	6	6	12	12	11		10	10
MCQ365A						15		18
MCQ365B	20				2	17	9	12
MCQ365C							16	15
MCQ365D		15			3	5	13	6
MCQ370A						3		
MCQ370B					17			
MCQ370C	18					11	15	
MCQ370D						8		
PAD615						9		
PAD675	16	7	20	17			20	
PAD680	10	16	3	8		2		
PAQ710	7	3	13	2		19	19	5

RIAGENDR					19			
RIDAGEYR	1	1	1	1	1	1	1	1
SMQ020					16			
SMQ040A					4	7		
SMQ910						20		
SMQ915A								
WHQ070						16		
WHQ150	4	14	2	5		6	2	4

original scores in grey

2.4.4 Blood pressure

This robustness test is to replace the continuous variables of blood pressure readings with their categorical variations which aim to better reflect if the participant's systolic or diastolic blood pressure level is at risk:

1. *BPXSYaveC*, where 1 indicates that the participant average Systolic BP level (>=120) is at risk, 2 otherwise
2. *BPXDIaveC*, where 1 indicates that the participant average Diastolic BP level (>=80) is at risk, 2 otherwise

The results in Table 2.15 show that these models do not perform as well. This is supported by the results in Table 2.16 where only *BPXSYaveC* is consistently significant in the Logitboost model. Results of the boosted trees models where the original features show high significance continue to support the finding that models perform better with continuous values. The importance of features are sensitive to its nature.

Table 2.15 Model Comparison – Blood Pressure (Binary)

Model	Recall (%)	ROC-AUC	TP	FP	FN	TN
XGB trees	65.4	0.660	34	18	56	294
	82.7	0.650	43	9	86	264
RF	78.9	0.693	41	11	56	294
	80.8	0.705	42	10	53	297
Logitboost	77.5	0.608	31	9	84	212
	80.0	0.609	36	9	95	201

original scores in grey

Table 2.16 Significant Features – Blood Pressure (Binary)

Variables	RF		XGB		Logitboost (ROC)	
ALQ120QU	19		8	13		
ALQ130	20					
BMDAVSAD	6	7	5	10	5	9
BMXARMC	15	18		19		
BMXARML	7	13	12			
BMXHT	12	16	18			
BMXLEG	10	15	19		20	
BMXWAIST	11	11	9	11	6	7
BMXWT	13	14	17			
BPQ020				17	7	8
BPQ080					8	16
BPQ090D	5	6	3	2	2	2
BPXDIaveC		9		12		
BPXSYaveC		4		5	13	3
DMDEDUC2					14	17
INDFMIN2				18	16	14
LBDHDD	9	10	15	9	15	13
LBDLDL	16	19		20	19	20
LBXTC	4	3	4	4	10	11
LBXTR	14	20	6	8	18	19
MCQ300C	18	12	11	6	12	10
MCQ365A						18
MCQ365B					9	12
MCQ365C			16		17	15
MCQ365D			7	15	11	6
PAD630			13			
PAD675	17	17	20	7		
PAD680	8	8	10	16		
PAQ710	2	2	2	3	4	5
RIDAGEYR	1	1	1	1	1	1
WHQ150	3	5	14	14	3	4

original scores in grey

2.4.5 BMI

The results discussed in Chapter 2.3 show that height and weight are significant in the T2D prediction (RF model). Yet, the BMI variable using the new cut-offs for Asian Americans included in the main dataset seemingly does not influence the prediction. It is only significant in the NN model results. While this can be due to multi-collinearity, the models that overcome this issue also do not capture the significance of BMI feature. Thus, a robustness test is employed to process a subset of features that excludes *BMXHT* and *BMXWT*.

Table 2.17 shows that removing these 2 features improves the performance of models (recall) where multi-collinearity is considered – Logitboost and NN. This is reflected in Table 2.18 where the NN model captures different features that are significant. As the importance of features are calculated using ROC-AUC for the Logitboost model, the results remain constant. Results of the boosted trees models are also fairly consistent.

Table 2.17 Model Comparison – Exclude Height and Weight (Binary)

Model	Recall (%)	ROC-AUC	TP	FP	FN	TN
XGB trees	65.4	0.652	34	18	60	290
	82.7	0.650	43	9	86	264
RF	78.9	0.679	41	11	63	287
	80.8	0.705	42	10	53	297
Logitboost	82.9	0.627	34	7	93	195
	80.0	0.609	36	9	95	201
NN	51.9	0.624	27	25	50	300
	50.0	0.627	26	26	52	298

original scores in grey

Table 2.18 Significant Features – Exclude Height and Weight (Binary)

Variables	XGB		RF		NN		Logitboost (ROC)	
ALQ101					8			
ALQ120QU	20	13	18					
BMDAVSAD	13	10	7	7			9	9
BMXARMC	15	19	17	18				
BMXARML	16		15	13				
BMXBMIAA						14		
BMXHT				16				
BMXLEG			10	15				
BMXWAIST	5	11	12	11			7	7
BMXWT				14				
BPQ020		17					8	8
BPQ080							16	16
BPQ090D	2	2	5	6	7	4	2	2
BPXDIave	11	12	11	9				
BPXSYave	3	5	4	4		13	3	3
DMDEDUC2					5		17	17
INDFMIN2	19	18			11		14	14
LBDHDD	17	9	8	10			13	13
LBDLDL		20	19	19	13		20	20
LBXTC	6	4	3	3	15	10	11	11
LBXTR	14	8	16	20			19	19
MCQ092					12	12		
MCQ160B					10			
MCQ160K					18			
MCQ160M						18		
MCQ203					16			
MCQ300C	9	6	13	12	14		10	10
MCQ365A					9	15	18	18
MCQ365B	18				17	17	12	12
MCQ365C					19		15	15
MCQ365D	7	15			1	5	6	6
MCQ370A						3		
MCQ370C						11		
MCQ370D					2	8		
PAD615						9		
PAD630			20					
PAD675	10	7	14	17				

PAD680	12	16	9	8		2		
PAQ710	4	3	2	2		19	5	5
RIDAGEYR	1	1	1	1	6	1	1	1
SMQ020					20			
SMQ040A					3	7		
SMQ910					4	20		
WHQ070						16		
WHQ150	8	14	6	5		6	4	4

original scores in grey

Further, the robustness tests also consider an alternative measure for BMI by using the original cut-offs, *BMXBMIO*. This test is processed 2 times – with and without *BMXHT* and *BMXWT*. Table 2.19 notes that only the NN model performance improves (recall and ROC-AUC). It is the only model that captured the significance of a BMI feature; however, this result is not as sensitive because the feature is not significant when the height and weight features are removed.

Table 2.19 Model Comparison – *BMXBMIO* (Binary)

Model	***BMXWT BMXHT***	**Recall (%)**	**ROC-AUC**	**TP**	**FP**	**FN**	**TN**
XGB trees	with	73.1	0.696	38	14	49	301
	without	69.2	0.704	36	16	51	299
		82.7	0.650	43	9	86	264
RF	with	75.0	0.698	39	13	50	300
	without	76.9	0.691	40	12	55	295
		80.8	0.705	42	10	53	297
Logitboost	with	72.9	0.647	35	13	67	250
	without	70.7	0.668	29	12	61	236
		80.0	0.609	36	9	95	201
NN	with	53.8	0.637	28	24	45	305
	without	55.8	0.647	29	23	43	307
		50.0	0.627	26	26	52	298

original scores in grey

Table 2.20 notes that the new BMI feature is still only significant in the NN model. It is interesting to note that the original results of NN show that significance of *BMXBMIAA* is allotted to the overweight category. The results of NN using *BMXBMIO* feature captures the significance of the normal weight category. This suggests that the new cut-off levels for Asian Americans are more accurate in one's health diagnosis.

Table 2.20 Significant Features – *BMXBMIO* (Binary)

Variables	XGB			RF			NN			Logitboost (ROC)		
	with	without		with	without		with	without		with	without	
ALQ120QU	18	10	13		19							
BMDAVSAD	8	18	10	7	8	7				6	6	9
BMXARMC	19		19	18	17	18						
BMXARML	9	14		12	13	13						
BMXBMIAA									14			
BMXBMIO								14				
BMXHT	17			15		16						
BMXLEG	11	20		14	12	15		20		18	18	
BMXWAIST	14	6	11	10	10	11	19	10		5	5	7
BMXWT	20			13		14						
BPQ020			17				13	18		7	7	8
BPQ080		12					2			11	11	16
BPQ090D	2	1	2	5	5	6	3	4	4	2	2	2
BPXDIave	10	13	12	11	11	9						
BPXSYave	7	9	5	3	4	4			13	4	4	3
DMDEDUC2										15	15	17
INDFMIN2			18					5		17	17	14
LBDHDD	12	16	9	9	7	10				14	14	13
LBDLDL			20	20	18	19		3				20
LBXTC	3	5	4	2	2	3	18	13	10	8	8	11
LBXTR	6	7	8	17	16	20		6		20	20	19
MCQ092							11		12			

MCQ160B							10					
MCQ160F							17					
MCQ160L							12					
MCQ160M		19					14		18			
MCQ160N								11				
MCQ300C	13	11	6	19	15	12	15	12		13	13	10
MCQ365A							1		15	19	19	18
MCQ365B							8	8	17	12	12	12
MCQ365C							4	15		16	16	15
MCQ365D		8	15				6	2	5	10	10	6
MCQ370A								7	3			
MCQ370C							20	16	11			
MCQ370D								9	8			
PAD615									9			
PAD630					20							
PAD675	16	17	7	16	14	17						
PAD680		15	16	8	9	8			2			
PAQ710	4	4	3	6	6	2			19	9	9	5
RIAGENDR							16					
RIDAGEYR	1	2	1	1	1	1	2	1	1	1	1	1
SMQ040A							5		7			
SMQ910							7		20			
WHQ070									16			
WHQ150	5	3	14	4	3	5			6	3	3	4

original scores in grey

Lastly, the robustness tests include replacing *BMXBMIAA* with the continuous variable, *BMXBMI* that records participants' BMI value as is. This test is processed 2 times – with and without *BMXHT* and *BMXWT*.

Results summarised in Table 2.21 support the above findings. The NN model performs better with the inclusion of the height and weight variable (recall and ROC-AUC). The overall performance improvement evaluated by ROC-AUC is witnessed in the Logitboost and XGB trees models. The performance of RF model is most consistent using continuous variables and, in this case, even performs as well as the original model (recall) when the height and weight features are removed. This is supported by the results in Table 2.22 where *BMXBMI* is only significant in the RF model.

Table 2.21 Model Comparison – *BMXBMI* (Binary)

Model	***BMXWT BMXHT***	**Recall (%)**	**ROC-AUC**	**TP**	**FP**	**FN**	**TN**
XGB trees	with	69.2	0.676	36	16	81	269
	without	67.3	0.704	35	17	50	300
		82.7	0.650	43	9	86	264
RF	with	80.1	0.696	42	10	57	293
	without	80.8	0.691	42	10	60	290
		80.8	0.705	42	10	53	297
Logitboost	with	73.9	0.639	34	12	86	210
	without	69.0	0.645	29	13	56	252
		80.0	0.609	36	9	95	201
NN	with	69.2	0.660	36	16	59	291
	without	50.0	0.654	26	26	55	295
		50.0	0.627	26	26	52	298

original scores in grey

Table 2.22 Significant Features – *BMXBMI* (Binary)

Variables	XGB			RF			NN			Logitboost (ROC)		
	with	without		with	without		with	without		with	without	
ALQ120QU	14	5	13	20	19							
BMDAVSAD	8	6	10	9	10	7	8			11	11	9
BMXARMC		7	19	17	12	18						
BMXARML	10	17		13	13	13						
BMXBMIAA									14			
BMXBMI				18	16							
BMXHT				11		16						
BMXLEG	18	18		16	14	15				18	18	
BMXWAIST	17	20	11	14	11	11				6	6	7
BMXWT				12		14						
BPQ0202			17	3						5	5	8
BPQ080							4	10		9	9	16
BPQ090D	3	2	2		3	6		2	4	2	2	2
BPXDIave	13	13	12	10	9	9						
BPXPULS				5			10					
BPXSYave	6	12	5		2	4			13	3	3	3
DMDEDUC2										13	13	17
INDFMIN2			18				11/15/1718	20		14	14	14
LBDHDD	15	10	9	7	7	10				17	17	13
LBDLDL			20		20	19		8				20
LBXTC	2	4	4	6	6	3	3	11	10	10	10	11
LBXTR	12	8	8	15	15	20				20	20	19

MCQ0922								9	12				
MCQ160B								19					
MCQ160M								16	18				
MCQ160N								18					
MCQ300C	9	15	6		18	12	14			15	16	10	
MCQ365A		16					9	17	15	19	19	18	
MCQ365B	16						2	15	17	12	12	12	
MCQ365C							19			16	15	15	
MCQ365D	20	14	15				7	7	5	7	7	6	
MCQ370A							16		3				
MCQ370B							12						
MCQ370C							1	3	11				
MCQ370D							6	4	8				
PAD615									9				
PAD630	19						20						
PAD675	7	19	7	19	17	17							
PAD680	5	9	16	8	8	8			2				
PAQ710	4	3	3	2	4	2		6	19	8	8	5	
RIDAGEYR	1	1	1	1	1	1	5	1	1	1	1	1	
SMQ040A								5	7				
SMQ9102									20				
SMQ925A								14					
WHQ070									16				
WHQ150	11	11	14	4	5	5	13	12	6	4	4	4	

original scores in grey

2.5 Findings

The classifier that performs most consistently across the sensitivity analysis and robustness tests is the RF model. Boosted trees models overcome the inherent multi-collinearity within the dataset. Healthcare professionals and insurance providers may consider the features that signal importance to the T2D prediction in RF model for any pre-diabetic diagnosis strategies.

The features that are common across all or most of the tests conducted include having prescribed cholesterol medication, age, waist circumference, family history and weight history. Their commonality enhances their influence in practical implications, giving healthcare professionals greater confidence to consider these variables with higher scrutiny.

To summarise some of the findings in the above discussion:

i. Models perform better with continuous values and variable importance is sensitive to the nature of the features
ii. Daily movement and consistency in activity levels, in comparison to weekly aggregates, likely contribute more to reduce risk of T2D
iii. Body measurements, in particular waist circumference and sagittal abdominal diameter, are more significant than BMI in the T2D prediction. This supports the research findings in obesity amongst Asians that BMI, even with the new cut-off points, do not necessarily convey their health status (WHO, 2004). People of Asian descent have different body composition. They tend to have less muscle and store more fat, especially around their waist. While they are smaller build and may not appear overweight, dangerous visceral fats can still be present especially around the sagittal abdominal area (Pou et al., 2009).

3.0 Multi-class classification

The NHANES dataset is comprehensive in terms of assessing a participant's health status using multiple methods such as laboratory tests and a questionnaire on medical conditions (Omogbai, 2016). The questionnaire includes a past/present diagnosis question on whether the participant has been told by a doctor if he/she has diabetes. Additionally, FPG level is measured as part of the laboratory test component. The dataset thus allows for the multi-class feature, *DIQ010B*, to be engineered. The prediction model can further analyse which factors play more significant roles at different stages of the T2D diagnosis.

As it is rational to expect that most participants do not have T2D, the inherent class imbalance is pre-processed using *SmoteClassif()* function from *UBL* package. Table 3.1 shows the proportion of the classes in *DIQ010B* before and after dataset balancing.

Table 3.1 Dataset Class Proportion – *DIQ010B*

	Before SMOTE	**%**	**After SMOTE**	**%**
1 – diagnosed diabetes participant answered '1' (yes) or '3' (borderline) to *DIQ010* question of whether a doctor has advised that he/she has diabetes, as well as a recorded FPG reading >= 126	110	11.71	235	25.03
2 – undiagnosed diabetes participant answered '2' (no) to *DIQ010* question of whether a doctor has advised that he/she has diabetes, as well as a recorded FPG reading >= 126	13	1.38	234	24.92
3 – prediabetes FPG reading between 100 – 125	151	16.08	235	25.03
4 – no diabetes FPG reading <= 100	665	70.08	235	25.03

Following the results and discussion of the binary classification problem, 8 of the 11 models are undertaken at this stage. The classifiers employed to predict T2D in the multi-class classification problem are:

1. Decision Tree (DT)
2. Random Forest (RF)
3. Extreme Gradient Boosting Trees (XGB)
4. Stochastic Gradient Boosted Trees (GBM)
5. Neural Network
6. Support Vector Machines (SVM)
7. Naïve Bayes (NB)
8. k-Nearest Neighbours (k-NN)

The following sections will document the modelling approach, evaluate performance of the models, detail the sensitivity analysis and robustness tests undertaken, and conclude on the weighted importance of features selected by these models.

3.1 Classification models

Model performances are evaluated using a combination of metrics derived from the confusion matrix, ROC-AUC and PR-AUC. This is summarised in Tables 3.3 and 3.4. The models reported below are trained using the balanced dataset. The models have also been trained using the imbalanced dataset, but the models are unable to make effective predictions, especially for the smaller classes '2' and '3'[15]. These 2 minority classes are also the individuals with highest T2D risk. Thus, the model performances are discussed below by drawing on recall and F1 scores calculated for these 2 minority classes.

3.1.1 Classification trees

There are 4 types of classification trees – Decision trees, Random forest, XGB trees and GBM trees. The boosted trees, in particular Random forest and XGB trees report consistently good

[15] See Appendix H.

results above for the binary classification. The nature of these models also overcome the risk of skewed or misleading results due to inherent multi-collinearity within the dataset.

Random forest and GBM trees perform the best out of the 4 types, and out of the 8 models trained. For minority class '3', GBM trees achieve the highest recall at 80.4%, followed by Random forest at 66.1%. Random forest model also achieved the highest F1 score for class '3' at 0.457. The recall and F1 scores for class '2' achieved by these 2 models are also ranked amongst the top 2 scores. The biggest ROC-AUC amongst the classification trees is illustrated by GBM trees at 0.672 which is also the second highest curve[16].

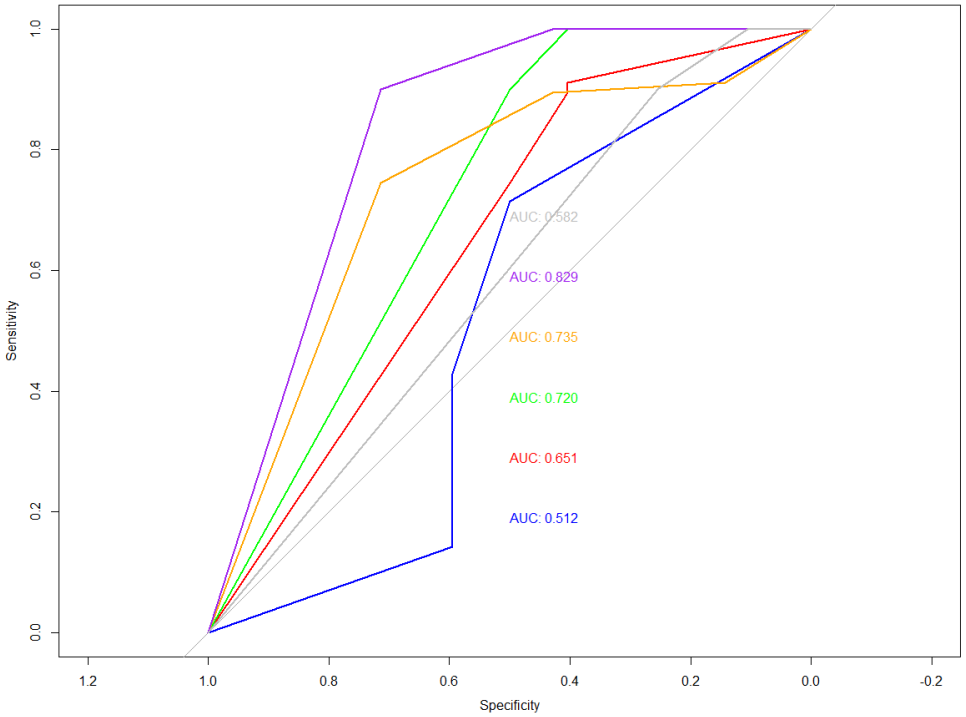


Figure 3.1 ROC-AUC of GBM Trees (Multi-class)

[16] See Appendix I.

3.1.2 Neural network

A feed forward neural network is employed in the initial modelling phase as part of feature selection. It results in a 119-1-4 network with 128 weights for *DIQ010B* model. This model is also undertaken in the modelling phase to allow comparison of features. The resulting feed forward neural network in this phase is a 73-9-4 network with 706 weights [17]. The model performance has an average showing in comparison to the others.

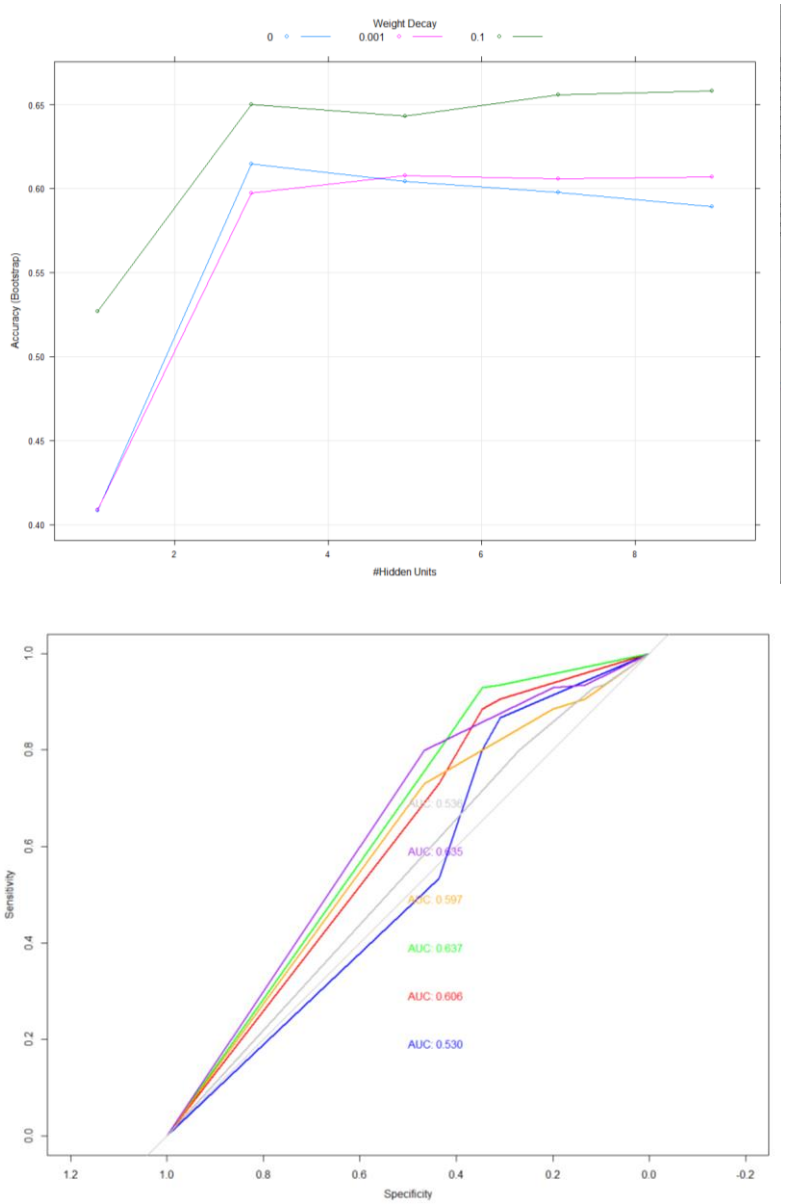


Figure 3.2 Neural Network and the ROC-AUC Graph (Multi-class)

[17] See Appendix I.

3.1.3 k-Nearest Neighbours

The k-NN algorithm is a robust and versatile classifier. It is usually applied as a benchmark for more complex classifiers such as NN and SVM. Despite its simplicity, k-NN can often outperform the more powerful classifiers which is true in this case. The k-NN classifier algorithm employed in the feature selection phase is also utilised for the modelling phase. The resulting k-NN classifier, using Euclidean distance shows an above average performance in comparison to the other models. The k-NN model scores the highest recall and F1 scores for class '2' at 42.9% and 0.357, respectively.

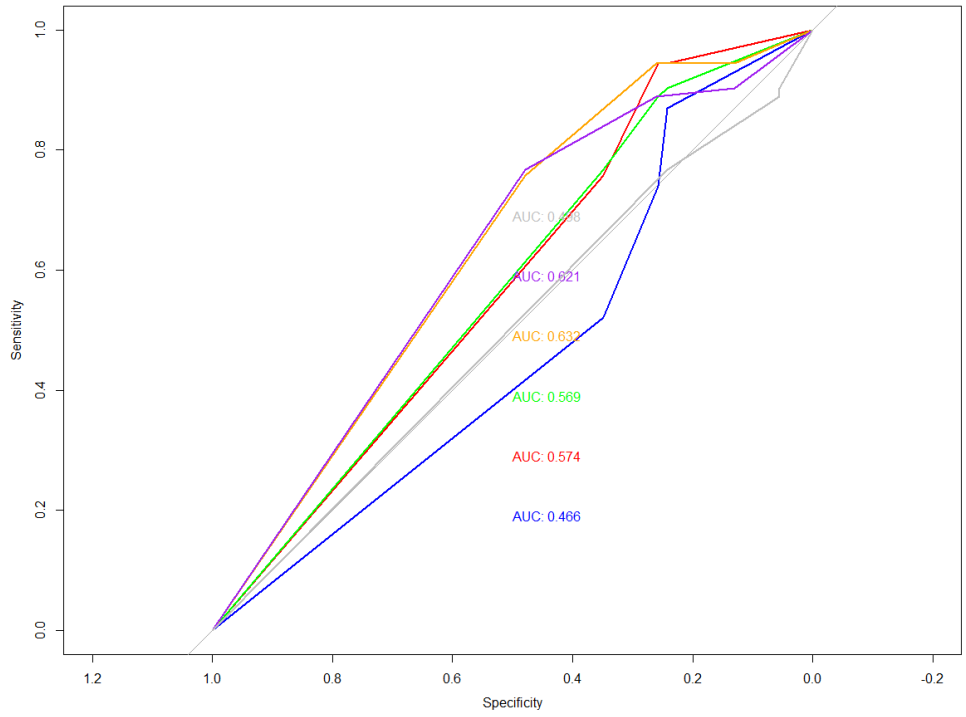


Figure 3.3 ROC-AUC of k-NN (Multi-class)

3.1.4 Support Vector Machines

While SVMs are inherently binary classifiers, the model can also be used for multi-class classification problems. Knerr et al. (1990) proposes implement the "1-against-1" approach for multi- class classification. The SVM models trained in the project include the linear, radial and polynomial kernels. The 3 models share relatively similar performances noted in Table 3.2. The SVM model is also employed for benchmarking purposes using different classification approaches. This is discussed in Chapter 3.2.2.

Table 3.2 SVM Models Performance

		Linear	Linear Grid Tuning	Radial	Polynomial
Class '2'	Recall	0	0.143	0.143	0.286
	F1	0	0.071	0.607	0.182
Class '3'	Recall	0.393	0.286	0.154	0.339
	F1	0.224	0.230	0.332	0.244
ROC		0.620	0.624	0.592	0.601

The SVM models show better scores for class '3' predictions in general. The best ROC-AUC of 0.624 is illustrated in Figure 3.4 for the SVM model with linear kernel using the default grid search routine specified. This is also the better SVM model for binary classification.

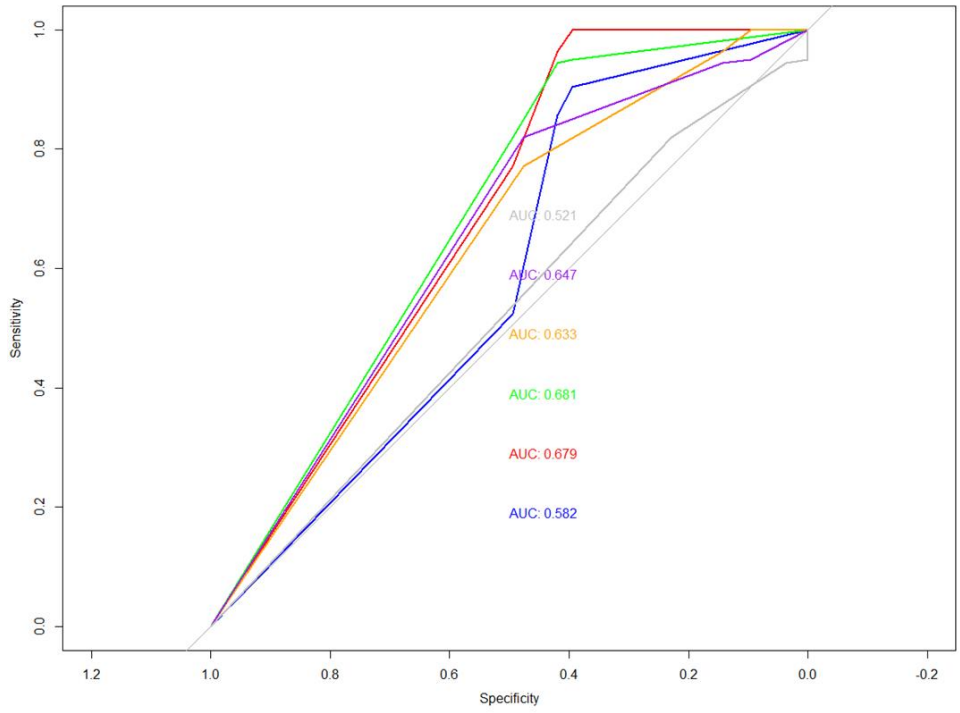


Figure 3.4 ROC-AUC of SVM (Multi-class)

3.1.5 Naïve Bayes

NB is another popular algorithm employed for multi-class prediction. The NB model can predict the probability of multiple classes of the target variable, in this case *DIQ010B*.

While the NB models trained[18] present the highest ROC-AUC of 0.753, the recall and F1 scores are very low between 0 – 3%. The better NB model with the highest ROC-AUC is trained using the *naivebayes* package. However, the NB model is not able to fully predict all classes well which is noted in Table 3.3. The unsteady performance of NB model across all classes is also illustrated in Figure 3.5.

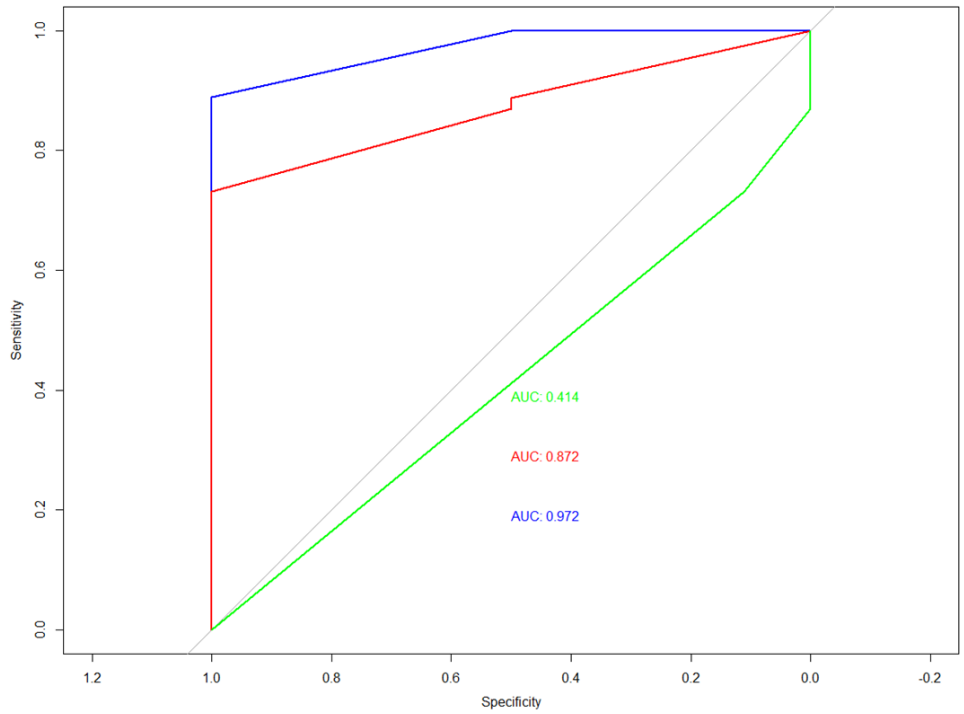


Figure 3.5 ROC-AUC of NB (Multi-class)

[18] See Appendix I for the tuning parameters of the best resulting NB model in terms of recall. The models are trained using *'nb'* and *'naïve_bayes'* methods within 2 R packages, *klaR* and *naivebayes*.

3.2 Validation criteria

The medical domain commonly evaluates models using test data as well as cross-validation. The 8 models are subject to the same evaluation process outlined in Chapter 2.2. The above discussion has already outlined 3 evaluation metrics the project uses for the multi-class prediction: Recall, F1 score and ROC-AUC[19].

The predictive models are evaluated on both training and test sets as part of validation. The test set prediction informs of how well the model has generalized to new unseen data. It is very likely that the models overfit on training data. While it is natural for model accuracy to drop for test sets, the results for these 8 models all significantly dropped in comparison to the models of binary classification. This is reasonable as there are more classes and imbalance in the dataset. The model accuracy scores for test set are between 40% to 65% (with the exception of DT at 15.9%).

3.2.1 Evaluation of models

Tables 3.3 and 3.4 summarises the evaluation metrics calculated for each model. The top 3 models that perform well in terms of recall for class '2' are k-NN, GBM trees and RF. The GBM trees has the best performance in terms of F1 Score and ROC-AUC (excluding NB).

Table 3.3 Summary of Model Evaluation Ranked Based on Recall for Class '2'

Rank	Model	Recall (%)	F1 Score	ROC-AUC	PR-AUC
1	k-NN	**42.9**	0.200	0.560	0.438
2	GBM	28.6	**0.286**	**0.672**	0.355
3	RF	28.6	0.250	0.639	**0.444**
4	XGB	28.6	0.235	0.609	0.389
5	NN	14.3	0.091	0.590	0.425
6	SVM 'Linear Grid'	14.3	0.071	0.624	**0.444**
7	DT	14.3	0.041	0.564	0.339
8	NB	0	0	**0.753**	**0.570**

[19] The formulae for evaluation metrics are set out in Table 2.2 and Chapter 2.2.

The GBM trees and RF models are consistent top performers for the metrics measuring class '3'. Overall, the top predictive models for the multi-class classification are GBM trees, RF and k-NN which will be employed for sensitivity analysis and robustness tests.

Table 3. 4 Summary of Model Evaluation Ranked Based on Recall for Class '3'

Rank	Model	Recall (%)	F1 Score
1	GBM	**80.4**	0.251
2	RF	66.1	**0.457**
3	XGB	58.9	0.252
4	DT	44.6	0.154
5	NN	41.1	0.225
6	k-NN	35.7	0.245
7	SVM 'Linear Grid'	28.6	0.230
8	NB	1.8	0.031

Figure 3.6 illustrates the top 4 ROC-AUC where NB, RF, GBM trees and SVM are ranked as such. Even though NB model seemingly performs well in Figure 3.6, the overall evaluation performance is low and unstable for prediction.

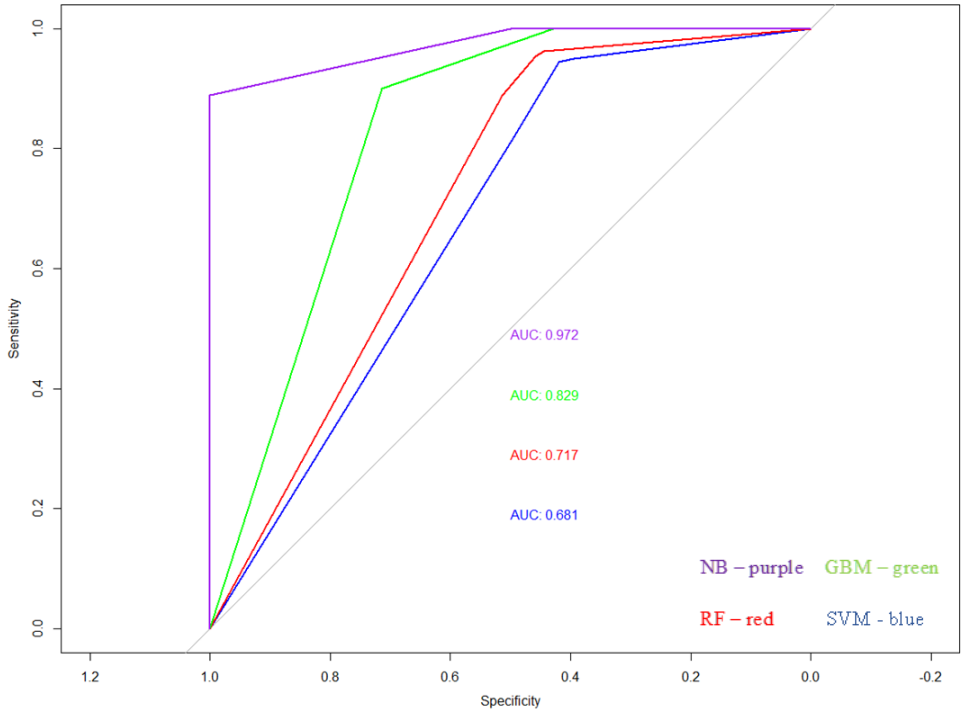


Figure 3.6 Top 4 ROC-AUC (Multi-class)

Figure 3.7 illustrates the top 4 PR-AUC out of the 8 models. 7 models, excluding NB, present very similar results for the PR-AUC within a close range of 0.30 to 0.45 (very small difference of 0.15).

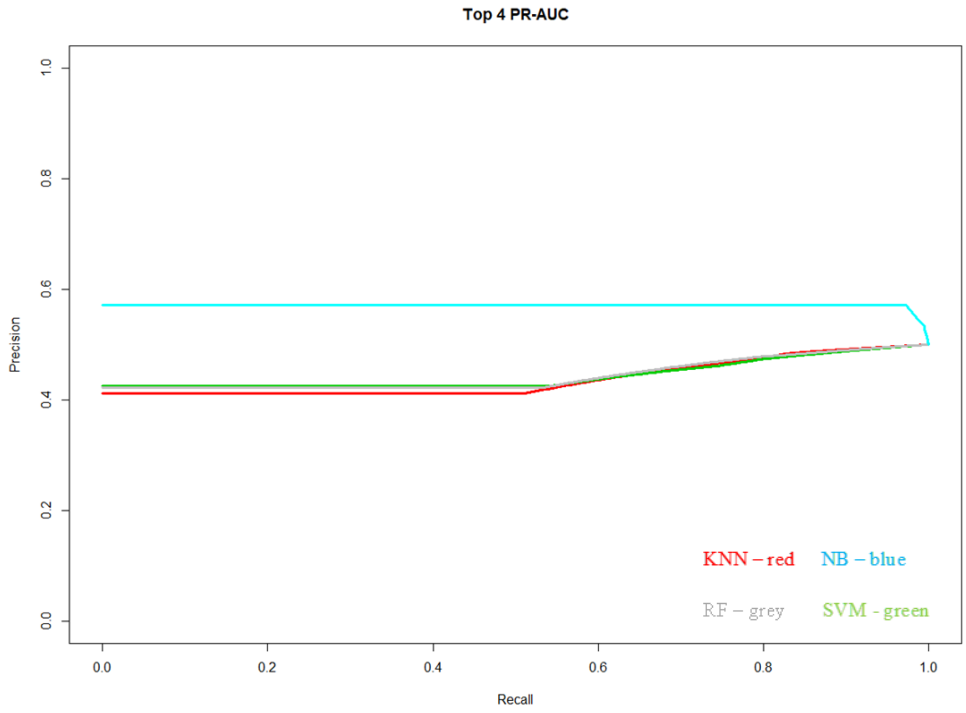


Figure 3.7 Top 4 PR-AUC (Multi-class)

3.2.2 Benchmarking

There are no existing studies that engage the same features used in this project for a multi-class prediction of T2D. The target variable *DIQ010B* is formulated based on 2 studies – Yu et al. (2010) and Omogbai (2016). Both studies employ the same 1999-2004 NHANES survey data and the SVM model.

Yu et al. (2010) apply 14 features to the SVM model, following the same basis for the multi-class target variable. The SVM model is trained using 2 classification schemes. The first scheme pits classes 1 and 2 versus classes 3 and 4, with the SVM radial model scoring the highest ROC-AUC of 0.835. The second scheme pits classes 2 and 3 against class 4 scoring a ROC-AUC of 0.731 with the SVM linear model.

Omogbai (2016) employs the same classification scheme basis and trained various SVM models using different subsets of features. The best SVM linear model scores a ROC-AUC of 0.848.

While the results reported in Table 3.3 are based on models trained using a different classification basis, the 2 classification schemes by Yu et al. (2010) and Omogbai (2016) are undertaken for the SVM model to provide a basis for benchmarking. The performance of the SVM models are documented in Table 3.5. As the 2 studies and this project analyse different cycles of NHANES survey and features, it is inconsistent to make direct comparisons. Further, this project focuses on the Asian American subgroup whereas these 2 studies study the entire sample of American adults.

Table 3. 5 SVM Models Performance

Classification Scheme	Model	Recall (%)	ROC-AUC
1	SVM 'Linear Grid'	**57.7**	0.681
	SVM Radial	34.6	**0.678**
	SVM Polynomial	50.0	0.668
2	SVM 'Linear Grid'	41.3	**0.581**
	SVM Radial	**55.6**	0.580
	SVM Polynomial	39.7	0.561

3.3 Features importance

The *varImp()* evaluation function is employed for this task. This function is not applied to the decision tree model. The way results of the models that employs *varImp()* evaluation function are calculated have been discussed in Chapter 2.3. For multi-class classification, some of the models – RF, NN and SVM/NB/k-NN which employs ROC-AUC estimates – assess the features importance separately for each class. All measures of importance are scaled to have a maximum value of 100.

Table 3.6 below summarises the results. For models with separate class calculations, the features are ranked by their average scores. There are 6 features that consistently express importance across all methods of estimation:

i. *ALQ120QU*, measures the number of days participant has had any type of alcoholic beverage in the past 12 months
ii. *BMXLEG*, measures upper leg length
iii. *BPXSYave*, systolic blood pressure reading
iv. *LBDLDL*, LDL cholesterol level (bad cholesterol)
v. *LBXTR*, Triglyceride level (bad cholesterol)
vi. *RIDAGEYR*, age of the participant

There are also 6 features that express importance across most models (bar 1 method of estimation):

i. *BMXHT*, height
ii. *BPQ090D*, where the participant has been told by a doctor/healthcare professional to take prescription for cholesterol
iii. *BPXDIave*, diastolic blood pressure reading
iv. *LBXTC*, total cholesterol
v. *PAD680*, measuring sedentary lifestyle by average minutes in a day where a participant is spent sitting (such as sitting at school, at home, getting to and from places, or with friends including time spent sitting at a desk, traveling in a car or bus, reading, playing cards, watching television, or using a computer; do not include time spent sleeping)
vi. *PAQ710*, measuring average hours in a day the participant spend sitting and watching TV or videos for a month

Table 3.6 Significant Features – Multi-class Classification

Variables	GBM	RF	XGB	DT	NN	SVM, k-NN, NB (ROC)
ALQ120QU	5	6	5	9	10	4
BMDAVSAD	16	11	17			16
BMXARMC		19	20			
BMXARML	11	16	18			
BMXHT	19	14	15		9	
BMXLEG	15	10	14		13	20
BMXWAIST	10	7	10			15
BMXWT		17				
BPQ020						11
BPQ080						18
BPQ090D	12		11	3	17	5
BPXPULS					16	
BPXDIave	6	8	6	5		13
BPXSYave	2	3	2	3	7	7
LBDHDD	17	15	16		14	
LBDLDL	3	2	3	2	8	3
LBXTC	8	5	7	4		1
LBXTR	1	1	1	1	6	6
MCQ160B					19	
MCQ160M					15	
MCQ160N					11	
MCQ300C	13		12			12
MCQ365B					12	10
MCQ365C					2	17
MCQ365D					3	9
MCQ370B						14
PAD630	9	18	8			
PAD675	18	20	19	7		
PAD680	7	9	9	6	4	
PAQ710	14	12	13		5	19
RIDAGEYR	4	4	4	10	1	2
SMQ020					20	
SMQ040A					18	
SMQ925A				8		
WHQ150	20	13				8

Figure 3.8 illustrates the results of GBM trees. The plots for RF and k-NN models are illustrated in Appendix J. The plots show that there are also other significant features within the dataset and that the importance of each feature vary across classes. The top 20 features are summarised in Table 3.6.

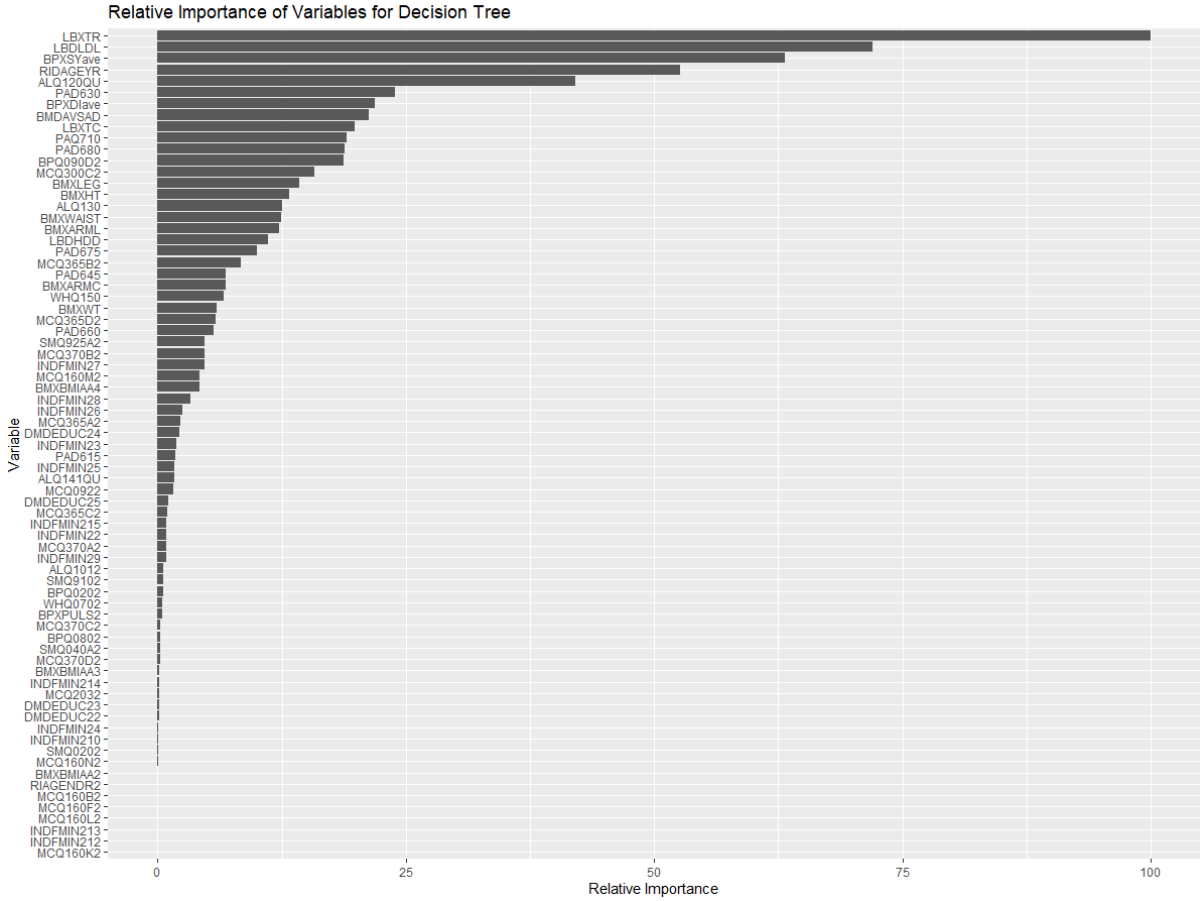


Figure 3.8 Significant Features of GBM (Multi-class)

Table 3.7 presents the results for models – RF and SVM/NB/k-NN (ROC) – that provide for the separate ranking of features within the 4 classes of target variable *DIQ010B*. The results for NN, although separated by class, calculated same importance values across the classes.

The consumption of alcohol, though ranked highly across most classes, its significance dropped considerably for class 3 which measures participants who are pre-diabetic. The results seem to suggest that alcohol consumption is a better predictor for those who are already diabetic, i.e. classes 1 and 2. Cholesterol level is more a significant predictor for pre-diabetes, in comparison to the others. Timely intervention and appropriate advice or treatment to lower cholesterol may be able to aid in reducing or negating the risk of pre-diabetes advancing to full-blown T2D.

It is interesting to note that there are some areas that only come up more significant in either of the estimation methods. For ROC-AUC estimation method, features regarding medical history are highly valued. For RF model-based estimation method, features of physical activity levels are more highly valued.

Overall, these results are consistent to the findings for the binary prediction, in particular the detrimental impact of a sedentary lifestyle. The predictive values of age, bad cholesterol levels and blood pressure readings for T2D diagnosis are also consistently strong in this multi-class prediction.

Further, the results also support the results of sensitivity analysis for binary classification where the frequency of alcohol intake in lifestyle choices is a contributing factor. This is also reinforced by the significance of *MCQ160N* which records whether the participant has had a gout diagnosis. Gout is a common form of inflammatory arthritis caused by excess uric acid in the bloodstream. Gout is common amongst those with high alcohol consumption which interferes with the removal of uric acid from the body.

There are also features that accounts for physical activity documented in Tables 3.5 and 3.6. For example, *PAD630* and *PAD 675* measuring a lifestyle that involves moderate-intensity work, sports, fitness or recreational activities daily, is consistently significant in the results of boosted trees.

These results will be further validated by applying sensitivity analysis and robustness tests discussed below.

Table 3.7 Significant Features Across Classes

	RF 1	ROC 1	RF 2	ROC 2	RF 3	ROC 3	RF 4	ROC 4
ALQ120QU	5	4	3	5	17	3	10	16
BMDAVSAD	13	20	18	9	18	19	5	5
BMXARMC	19		16		16		11	
BMXARML	12		11		15		19	
BMXHT	15		12		4		18	
BMXLEG	9	14	9	20	10	16	12	14
BMXWAIST	8	19	15	8	11	20	6	4
BMXWT	11		19		14		15	
BPQ020		13		11		11		7
BPQ080		18		16		18		12
BPQ090D		6		4		12		2
BPXDIave	7	10	7	15	5	7	8	18
BPXSYave	6	15	1	10	9	4	4	6
LBDHDD	18		14		7		16	
LBDLDL	3	2	4	3	1	2	2	17
LBXTC	2	1	5	2	3	1	14	11
LBXTR	4	5	2	6	2	5	1	19
MCQ300C		9		17		15		10
MCQ365B		8		13		14		8
MCQ365C		17		19		10		15
MCQ365D		12		14		9		9
MCQ370B		7		12		6		20
PAD630	16		10		20		17	
PAD675	20		17		19		20	
PAD680	17		6		8		9	
PAQ710	10	16	8	18	12	17	13	13
RIDAGEYR	1	3	13	1	6	8	3	1
WHQ150	14	11	20	7	13	13	7	3

3.4 Sensitivity analysis and robustness tests

Additional tests are carried out to examine the robustness of the main results. The rationale for running the following tests has been set out in Chapter 2.4. This project tests the robustness of the results derived from top performing models evaluated above (RF, GBM trees and k-NN) with the following methods:

i. alternative model specifications, specifically hyperparameter tuning of the models which have already been presented above
ii. include laboratory data features to the main dataset
iii. features substitution/exclusion
iv. organise the dataset to a 3-class classification problem instead of 4 classes by excluding patients who have already been diagnosed (exclusion of class 1)
v. organise the dataset to a binary classification problem (1 versus rest)

3.4.1 Laboratory data

As the NHANES dataset provides data collected from laboratory tests offering more precise insight into a participant's health status, the data can be leveraged to possibly improve the performance on the classification. Results of feature selection show that there are important laboratory features which may aid in the T2D prediction. As the motivation for this project is to suggest lifestyle and easy-to-collect factors that can indicate T2D before having to run the actual T2D tests, the main dataset excludes the laboratory test features.

a. Cholesterol (*LBXTC*, *LBXTR*, *LBDHDD*, *LBDLDL*)

Tables 3.6 and 3.7 show that the cholesterol features are consistently significant. The robustness tests include running a subset of the main dataset that excludes these 4 features.

Table 3.8 compares if the updated models perform better to the original models trained using the full dataset. The k-NN model is most consistent in its performance whereas the predictive ability for RF model on these 2 classes drops (recall). This may be due to cholesterol features ranking highly for RF model noted in Tables 3.5 and 3.6 and their importance in the T2D prediction.

Table 3.8 Model Comparison – Exclude Cholesterol (Multi-class)

Model	ROC-AUC	Class '2'		Class '3'	
		Recall (%)	F1 Score	Recall (%)	F1 Score
GBM trees	0.601	42.6	0.11	50.0	0.24
	0.672	28.6	0.29	80.4	0.25
RF	0.640	28.6	0.13	39.3	0.26
	0.639	28.6	0.25	66.1	0.46
k-NN	0.574	42.9	0.13	39.3	0.24
	0.560	42.9	0.20	35.7	0.25

original scores in grey

By eliminating cholesterol features, the features deemed important by these 3 models summarised in Table 3.8[20] still substantiates the original list in Table 3.6. It is also worthwhile to note that a new feature signals its importance in Table 3.9 across RF and k-NN models:

i. *PAD660*, measures minutes in a day that a participant spend doing vigorous-intensity sports, fitness or recreational activities

The results continue to support the evidence that participants at risk of T2D may improve upon their lifestyle with some form of daily physical activity.

[20] Appendix K documents the class-based results by RF and k-NN.

Table 3.9 Significant Features – Exclude Cholesterol (Multi-class)

Variables	GBM		RF		k-NN (ROC)	
ALQ120QU	3	5	4	6	2	4
BMDAVSAD	9	16	6	11	13	16
BMXARMC	14		15	19		
BMXARML	13	11	13	16		
BMXHT	11	19	11	14		
BMXLEG	8	15	10	10	18	20
BMXWAIST	7	10	5	7	12	15
BMXWT			14	17		
BPQ020					8	11
BPQ080					15	18
BPQ090D	15	12	3		3	5
BPXDIave	5	6	8	8	10	13
BPXSYave	1	2	1	3	4	7
DMDEDUC2					17	
INDFMIN2					19	
LBDHDD		17		15		
LBDLDL		3		2		3
LBXTC		8		5		1
LBXTR		1		1		6
MCQ300C	6	13			9	12
MCQ365B					7	10
MCQ365C					14	17
MCQ365D					6	9
MCQ370B	19		18		11	14
PAD630	12	9	16	18		
PAD645	20					
PAD660			19		20	
PAD675	16	18	17	20		
PAD680	4	7	7	9		
PAQ710	10	14	9	12	16	19
RIDAGEYR	2	4	2	4	1	2
WHQ150	18	20	12	13	5	8

original scores in grey

Another robustness test is to replace the continuous variables of *LBXTC*, *LBDHDD*, *LBXTR* and *LBDLDL* with their categorical variations – *LBXTCA*, *LBDHDDA*, *LBXTRA* and *LBDLDLA*.

Table 3.10 shows that the new models do not perform very well in terms of recall after removing the 4 key significant features. Similar to the binary classification results reported in Chapter 2, the models do not pick up on the categorical variables as well as the continuous variables. Table 3.11 reveals that the categorical cholesterol features drop out completely.

Table 3.10 Model Comparison – Categorical Cholesterol (Multi-class)

Model	ROC-AUC	Class '2'		Class '3'	
		Recall (%)	F1 Score	Recall (%)	F1 Score
GBM trees	0.623	28.6	0.11	44.6	0.23
	0.672	28.6	0.29	80.4	0.25
RF	0.659	28.6	0.15	35.7	0.26
	0.639	28.6	0.25	66.1	0.46
k-NN	0.582	28.6	0.09	17.9	0.13
	0.560	42.9	0.20	35.7	0.25

original scores in grey

Table 3.11 Significant Features – Categorical Cholesterol (Multi-class)

Variables	GBM		RF		k-NN (ROC)	
ALQ120QU	3	5	3	6	2	4
ALQ130	20					
BMDAVSAD	8	16	6	11	12	16
BMXARMC	16		16	19		
BMXARML	10	11	13	16		
BMXHT	13	19	10	14		
BMXLEG	15	15	9	10	19	20
BMXWAIST	5	10	5	7	9	15
BMXWT			12	17		
BPQ020					14	11
BPQ080					15	18
BPQ090D	9	12	4		3	5
BPXDIave	12	6	8	8	10	13
BPXSYave	1	2	2	3	5	7
DMDEDUC2					17	
INDFMIN2					18	
LBDHDD		17		15		
LBDLDL		3		2		3
LBXTC		8		5		1
LBXTR		1		1		6
MCQ300C	6	13	7		4	12
MCQ365B					11	10
MCQ365C					16	17
MCQ365D					6	9
MCQ370B	18				8	14
MCQ370C					13	
PAD615			20			
PAD630	7	9	18	18		
PAD645	19					
PAD675	17	18	17	20		
PAD680	4	7	14	9		
PAQ710	11	14	11	12		19
RIDAGEYR	2	4	1	4	1	2
SMQ925A			19			
WHQ150	14	20	15	13	7	8

original scores in grey

b. Gamma glutamyl transferase (*LBXSGTSI*)
c. Calcium (*LBXSCA*)
d. Potassium (*LBXSKSI*)
e. Phosphorous (*LBXSPH*)

The robustness tests include running an extended subset of the main dataset by including these 4 features – *LBXSGTSI, LBXSCA, LBXSKSI* and *LBXSPH*. As the motivation for this project is to suggest lifestyle and easy-to-collect factors that can indicate T2D before having to run the actual T2D tests, the main dataset excludes the laboratory test features. The results in Table 3.12 note that overall performance for the models, evaluated by ROC-AUC, significantly improves with the inclusion of laboratory test data. Table 3.12 also shows that the inclusion of laboratory data improves the models' predictive ability for class 3 (pre-diabetes). This is a good sign because early diagnosis of prediabetes and timely intervention (such as lifestyle changes) can be reversed and do not necessarily have to progress into T2D.

Table 3.12 Model Comparison – Laboratory Data (Multi-class)

Model	ROC-AUC	Class '2'		Class '3'	
		Recall (%)	F1 Score	Recall (%)	F1 Score
GBM trees	0.706	14.3	15.4	82.1	0.29
	0.672	28.6	0.29	80.4	0.25
RF	0.685	14.3	0.17	67.9	0.39
	0.639	28.6	0.25	66.1	0.46
k-NN	0.585	28.6	0.11	35.7	0.24
	0.560	42.9	0.20	35.7	0.25

original scores in grey

Table 3.13 indicates the importance of laboratory test features in a T2D prediction, especially the presence/absence of potassium, *LBXSKSI* and phosphorous, *LBXSPH*. These 2 features are significant across all 3 models.

Table 3.13 Significant Features – Laboratory Data (Multi-class)

Variables	GBM		RF		k-NN (ROC)	
ALQ120QU	5	5	11	6	4	4
BMDAVSAD	15	16		11	20	16
BMXARMC				19		
BMXARML	20	11	16	16		
BMXHT	18	19	18	14		
BMXLEG	13	15	12	10		20
BMXWAIST		10	9	7	19	15
BMXWT				17		
BPQ020					15	11
BPQ080					12	18
BPQ090D	14	12			5	5
BPXDIave	6	6	7	8	9	13
BPXSYave	4	2	4	3	6	7
LBDHDD	12	17	15	15		
LBDLDL	2	3	1	2	3	3
LBXSCA			17			
LBXSGTSI			6			
LBXSKSI	17		10		14	
LBXSPH	16		14		17	
LBXTC	7	8	5	5	2	1
LBXTR	1	1	3	1	7	6
MCQ300C	8	13			10	12
MCQ365B					13	10
MCQ365C						17
MCQ365D					11	9
MCQ370B					18	14
PAD630	9	9	19	18		
PAD675		18		20		
PAD680	10	7	8	9		
PAQ710	11	14	13	12	16	19
RIDAGEYR	3	4	2	4	1	2
SMQ925A			20			
WHQ150	19	20		13	8	8

original scores in grey

3.4.2 Physical activity

This robustness test is to replace some features that study a participant's **daily** physical activity captured in minutes: *PAD615, PAD630, PAD645, PAD660* and *PAD675*. These features are replaced with alternatives that the NHANES questionnaire consists of by capturing the participant's **weekly** records for the same activities: *PAQ605, PAQ620, PAQ635, PAQ650* and *PAQ665*. The definitions of these variables are listed in Appendix A.

Results documented in Table 3.14 indicate an improvement in overall model performance (ROC-AUC) but a slight decline in the predictive accuracy for the minority classes (recall). *PAD630* and *PAD675* are significant in the main results. *PAD645* and *PAD660* signal their importance in the results of some robustness tests. Table 3.15 shows that none of the corresponding weekly features is significant. The results in Table 3.15 is more conclusive in its signal than that of the binary classification in Table 2.12; insinuating that daily movement and consistency in activity levels are critical for a lifestyle change in reducing or negating T2D risk.

Table 3.14 Model Comparison – Physical Activity (Multi-class)

Model	ROC-AUC	Class '2'		Class '3'	
		Recall (%)	F1 Score	Recall (%)	F1 Score
GBM trees	0.714	14.3	0.22	78.6	0.28
	0.672	28.6	0.29	80.4	0.25
RF	0.695	28.6	0.36	60.7	0.46
	0.639	28.6	0.25	66.1	0.46
k-NN	0.575	28.6	0.10	46.4	0.30
	0.560	42.9	0.20	35.7	0.25

original scores in grey

Table 3.15 Significant Features – Physical Activity (Multi-class)

Variables	GBM		RF		k-NN (ROC)	
ALQ120QU	6	5	8	6	5	4
BMDAVSAD	17	16	10	11	19	16
BMXARMC			19	19		
BMXARML	9	11	12	16		
BMXHT	19	19	17	14		
BMXLEG	15	15	15	10	20	20
BMXWAIST	11	10	7	7	14	15
BMXWT	18		13	17		
BPQ020					13	11
BPQ080					15	18
BPQ090D	14	12	6		4	5
BPXDIave	5	6	9	8	10	13
BPXSYave	4	2	4	3	7	7
LBDHDD	16	17	18	15		
LBDLDL	2	3	2	2	3	3
LBXTC	8	8	5	5	1	1
LBXTR	1	1	1	1	6	6
MCQ300C	10	13			11	12
MCQ365B					9	10
MCQ365C					16	17
MCQ365D	20				12	9
MCQ370B					17	14
PAD630		9		18		
PAD675		18		20		
PAD680	7	7	14	9		
PAQ710	12	14	11	12	18	19
RIDAGEYR	3	4	3	4	2	2
SMQ925A			20			
WHQ150	13	20	16	13	8	8

original scores in grey

3.4.3 Smoking

This robustness test replaces *SMQ910* with *SMQ915A* to test if recent use of smokeless tobacco in the last 30 days will be a better predictor. The model performance documented in Table 3.16 is fairly similar to that of the main models. This is supported by the results in Table 3.17 where neither feature is significant. This is not to say smoking as a lifestyle choice is insignificant to the T2D prediction. Other test results recorded suggest otherwise.

Table 3.16 Model Comparison – Smoking (Multi-class)

Model	ROC-AUC	Class '2'		Class '3'	
		Recall (%)	F1 Score	Recall (%)	F1 Score
GBM trees	0.636	14.3	0.12	76.8	0.25
	0.672	28.6	0.29	80.4	0.25
RF	0.652	28.6	0.27	62.5	0.43
	0.639	28.6	0.25	66.1	0.46
k-NN	0.581	28.6	0.11	42.9	0.27
	0.560	42.9	0.20	35.7	0.25

original scores in grey

Table 3.17 Significant Features – Smoking (Multi-class)

Variables	GBM		RF		k-NN (ROC)	
ALQ120QU	5	5	8	6	3	4
ALQ130	19					
BMDAVSAD	13	16	6	11	19	16
BMXARMC			18	19		
BMXARML	15	11	16	16		
BMXHT	17	19	14	14		
BMXLEG	16	15	15	10		20
BMXWAIST	11	10	7	7	20	15
BMXWT			10	17		
BPQ020					15	11
BPQ080					17	18
BPQ090D	20	12			5	5
BPXDIave	7	6	13	8	18	13
BPXSYave	2	2	4	3	9	7
LBDHDD	18	17	11	15		
LBDLDL	4	3	2	2	4	3
LBXTC	6	8	5	5	2	1
LBXTR	1	1	1	1	6	6
MCQ300C	9	13	17		7	12
MCQ365B					11	10
MCQ365C					16	17
MCQ365D					12	9
MCQ370B					13	14
MCQ370C					10	
PAD630	12	9	19	18		
PAD675	8	18	20	20		
PAD680	14	7	12	9		
PAQ710	10	14	9	12	14	19
RIDAGEYR	3	4	3	4	1	2
WHQ150		20		13	8	8

original scores in grey

3.4.4 Blood pressure

This robustness test is to replace the continuous variables of blood pressure readings with their categorical variations – *BPXSYaveC* and *BPXDIaveC*.

The model performance documented in Table 3.18 is fairly similar to that of the main models. Both systolic and diastolic blood pressure readings are deemed important to the T2D prediction in the main results. The results in Table 3.19 note that only *BPXSYaveC* is still significant across the 3 models. The results support the finding that models perform better with continuous values and the importance of features are sensitive to its nature.

Table 3.18 Model Comparison – Blood Pressure (Multi-class)

Model	ROC-AUC	Class '2'		Class '3'	
		Recall (%)	F1 Score	Recall (%)	F1 Score
GBM trees	0.701	28.6	0.22	69.6	0.26
	0.672	28.6	0.29	80.4	0.25
RF	0.677	28.6	0.29	58.9	0.45
	0.639	28.6	0.25	66.1	0.46
k-NN	0.616	14.3	0.07	35.7	0.26
	0.560	42.9	0.20	35.7	0.25

original scores in grey

Table 3.19 Significant Features – Blood Pressure (Multi-class)

Variables	GBM		RF		k-NN (ROC)	
ALQ120QU	4	5	6	6	5	4
BMDAVSAD	14	16		11	17	16
BMXARMC			16	19		
BMXARML	16	11	13	16		
BMXHT	15	19	10	14		
BMXLEG	9	15	8	10	18	20
BMXWAIST	11	10	7	7	16	15
BMXWT			12	17		
BPQ020					14	11
BPQ080					12	18
BPQ090D	8	12	4		3	5
BPXDIaveC		6		8		13
BPXSYaveC	7	2	14	3	11	7
LBDHDD	18	17	15	15		
LBDLDL	2	3	3	2	4	3
LBXTC	5	8	5	5	2	1
LBXTR	1	1	1	1	6	6
MCQ300C	10	13			7	12
MCQ365B					13	10
MCQ365C						17
MCQ365D					9	9
MCQ370B			19		10	14
PAD630	13	9	17	18		
PAD660	17				19	
PAD675	19	18	18	20		
PAD680	6	7	11	9		
PAQ710	12	14	9	12	15	19
RIDAGEYR	3	4	2	4	1	2
SMQ925A			20			
WHQ150	20	20		13	8	8

original scores in grey

3.4.5 BMI

The results discussed in Chapter 2.3 show that height and weight are significant in the T2D prediction. Height is significant across most of the models whereas the importance of weight in a T2D prediction is only captured in the RF model results. To test if the lack of significance in the prediction for BMI feature in the main dataset may be due to multi-collinearity, a robustness test is employed to process a subset of features that excludes *BMXHT* and *BMXWT*.

While the exclusion of these 2 features improves the model performance in binary classification, Table 3.20 shows that this does not hold true for the multi-class models. This is further validated by the results documented in Table 3.21 where the significance of *BMXBMIAA* is still not captured.

Table 3.20 Model Comparison – Exclude Height and Weight (Multi-class)

Model	ROC-AUC	Class '2'		Class '3'	
		Recall (%)	F1 Score	Recall (%)	F1 Score
GBM trees	0.643	28.6	0.22	69.6	0.26
	0.672	28.6	0.29	80.4	0.25
RF	0.672	28.6	0.31	60.7	0.43
	0.639	28.6	0.25	66.1	0.46
k-NN	0.560	42.9	0.18	37.5	0.26
	0.560	42.9	0.20	35.7	0.25

original scores in grey

Table 3.21 Significant Features – Exclude Height and Weight (Multi-class)

Variables	GBM		RF		k-NN (ROC)	
ALQ120QU	5	5	9	6	4	4
BMDAVSAD	14	16	10	11	16	16
BMXARMC	19		18	19		
BMXARML	11	11	15	16		
BMXHT		19		14		
BMXLEG	18	15	12	10	20	20
BMXWAIST	16	10	8	7	15	15
BMXWT				17		
BPQ020					11	11
BPQ080					18	18
BPQ090D	9	12	6		5	5
BPXDIave	7	6	7	8	13	13
BPXSYave	1	2	4	3	7	7
LBDHDD	13	17	16	15		
LBDLDL	3	3	1	2	3	3
LBXTC	6	8	5	5	1	1
LBXTR	2	1	2	1	6	6
MCQ300C	15	13			12	12
MCQ365B					10	10
MCQ365C					17	17
MCQ365D					9	9
MCQ370B					14	14
PAD630	12	9	17	18		
PAD675	17	18	19	20		
PAD680	8	7	11	9		
PAQ710	10	14	13	12	19	19
RIDAGEYR	4	4	3	4	2	2
SMQ925A			20			
WHQ150	20	20	14	13	8	8

original scores in grey

Further, the robustness tests also consider an alternative measure for BMI by using the original cut-offs, *BMXBMIO*. This test is processed 2 times – with and without *BMXHT* and *BMXWT*.

The model performance documented in Table 3.22 is fairly similar to that of the main models. The results in Table 3.23 also supports the discussion for Table 3.20 where the predictive ability of categorical BMI features is just not sensitive, with or without the height and weight features. The earlier comparison between continuous and categorical natures of features is also validated by the next robustness test undertaken where the continuous values of BMI are employed in the prediction models.

Table 3.22 Model Comparison – *BMXBMIO* (Multi-class)

Model	***BMXHT BMXWT***	ROC-AUC	Class '2'		Class '3'	
			Recall (%)	F1 Score	Recall (%)	F1 Score
GBM trees	with	0.652	28.6	0.15	69.6	0.24
	without	0.661	28.6	0.24	75.0	0.26
		0.672	28.6	0.29	80.4	0.25
RF	with	0.637	28.6	0.27	55.4	0.40
	without	0.654	28.6	0.24	53.6	0.42
		0.639	28.6	0.25	66.1	0.46
k-NN	with	0.572	14.3	0.06	39.3	0.24
	without	0.581	14.3	0.06	39.3	0.24
		0.560	42.9	0.20	35.7	0.25

original scores in grey

Table 3.23 Significant Features – *BMXBMIO* (Multi-class)

Variables	GBM			RF			k-NN (ROC)		
	with	without		with	without		with	without	
ALQ120QU	6	5	5	7	7	6	5	4	4
BMDAVSAD	16	19	16	9	8	11	19	19	16
BMXARMC		20		18	16	19			
BMXARML	18	12	11	13	17	16			
BMXHT	19		19	12		14			
BMXLEG	17	18	15	17	14	10			20
BMXWAIST	15	13	10	10	9	7	16	16	15
BMXWT				15		17			
BPQ020							12	12	11
BPQ080							17	18	18
BPQ090D	12	9	12	5	4		3	3	5
BPXDIave	7	10	6	8	11	8	10	10	13
BPXSYave	2	2	2	4	5	3	6	6	7
DMDEDUC2							18	17	
LBDHDD	20	16	17	11	12	15			
LBDLDL	4	4	3	2	2	2	4	5	3
LBXTC	5	6	8	6	6	5	2	2	1
LBXTR	1	1	1	1	1	1	7	7	6
MCQ300C	10	14	13				9	9	12
MCQ365B							13	13	10
MCQ365C									17
MCQ365D							11	11	9
MCQ370B							14	14	14
MCQ370C							15	15	
PAD630	9	8	9	19	18	18			
PAD675	14	11	18	20	19	20			
PAD680	8	7	7	14	15	9			
PAQ710	13	15	14	16	13	12			19
RIDAGEYR	3	3	4	3	3	4	1	1	2
SMQ925A								20	
WHQ150	11	17	20		10	13	8	8	8

original scores in grey

Lastly, the robustness tests include replacing *BMXBMIAA* with the continuous variable, *BMXBMI* that records participants' BMI value as is. This test is processed 2 times – with and without *BMXHT* and *BMXWT*.

Results recorded in Table 3.24 detect improvement in the recall metric for GBM trees and k-NN models which includes *BMXBMI*, *BMXHT* and *BMXWT*. There is also improvement to overall model performance for RF and k-NN models (ROC-AUC). Table 3.25 shows that *BMXBMI* is significant across the boosted trees models, even alongside *BMXHT* and *BMXWT* in the RF model.

Table 3.24 Model Comparison – *BMXBMI* (Multi-class)

Model	***BMXHT BMXWT***	ROC-AUC	Class '2'		Class '3'	
			Recall (%)	F1 Score	Recall (%)	F1 Score
GBM trees	with	0.667	28.6	0.21	85.7	0.29
	without	0.654	28.6	0.22	80.4	29.3
		0.672	28.6	0.29	80.4	0.25
RF	with	0.697	28.6	0.33	62.5	0.40
	without	0.670	28.6	0.27	64.3	0.43
		0.639	28.6	0.25	66.1	0.46
k-NN	with	0.578	28.6	0.11	35.7	0.24
	without	0.591	28.6	0.10	30.4	0.21
		0.560	42.9	0.20	35.7	0.25

original scores in grey

Table 3.25 Significant Features – *BMXBMI* (Multi-class)

Variables	GBM			RF			k-NN (ROC)		
	with	without		with	without		with	without	
ALQ120QU	5	4	5	6	6	6	3	3	4
BMDAVSAD	12		16	8	9	11	17	17	16
BMXARMC		20		19	16	19			
BMXARML	20	16	11	15	14	16			
BMXBMI	17	19		18	17				
BMXHT			19	13		14			
BMXLEG	14	10	15	11	12	10			20
BMXWAIST	18	15	10	7	7	7	14	14	15
BMXWT				14		17			
BPQ020							12	12	11
BPQ080							15	15	18
BPQ090D2	9	11	12				4	4	5
BPXDIave	10	8	6	10	8	8	11	11	13
BPXSYave	3	2	2	4	4	3	7	7	7
LBDHDD	19	17	17	16	15	15	19	19	
LBDLDL	2	3	3	1	1	2	5	5	3
LBXTC	6	6	8	5	5	5	1	1	1
LBXTR	1	1	1	2	2	1	6	6	6
MCQ300C	8	7	13		11		8	8	12
MCQ365B							10	10	10
MCQ365C									17
MCQ365D							16	16	9
MCQ370B							13	13	14
PAD630	11	14	9	17	18	18			
PAD675	15	18	18		19	20			
PAD660							20	20	
PAD680	7	12	7	12	13	9			
PAQ710	13	9	14	9	10	12			19
RIDAGEYR	4	5	4	3	3	4	2	2	2
SMQ925A				20	20		18	18	
WHQ150	16	13	20			13	9	9	8

original scores in grey

3.4.6 3-class classification

One of the project objectives is to provide an analytical approach for predicting individuals who are at risk of T2D. The subsequent robustness test follows this notion and excluded the original class '1' where patients have already been diagnosed. The dataset is reorganised to a 3-class classification problem (original classes '2', '3', '4'). GBM trees, RF and k-NN models are trained using the 3-class balanced dataset. Table 3.26 shows that the models perform better for the prediction of prediabetes, especially improving the ROC-AUC of RF.

Table 3.26 Model Comparison – Exclude Diagnosed (Multi-class)

Model	ROC-AUC	Class '1'		Class '2'	
		Recall (%)	F1 Score	Recall (%)	F1 Score
GBM trees	0.684	28.6	0.31	75.0	0.33
	0.672	28.6	0.29	80.4	0.25
RF	0.867	14.3	0.25	60.7	0.41
	0.639	28.6	0.25	66.1	0.46
k-NN	0.542	14.3	0.06	48.2	0.31
	0.560	42.9	0.20	35.7	0.25

original scores in grey

The features deemed important in Table 3.27 substantiates the original list. It also includes 3 features that were not previously significant:

i. *PAD660*, where it records if the participant is reducing salt in diet
ii. *SMQ020*, where it records if the participant smoked at least 100 cigarettes in life
iii. *SMQ925A*, 1where it records if the participant has smoked a cigarette even 1 time in their life

The results continue to support the evidence that participants at risk of T2D may improve upon their lifestyle with a healthier diet, daily physical activity be it work or recreational and moderate their alcohol and smoking habits.

Table 3.27 Significant Features – Exclude Diagnosed (Multi-class)

Variables	GBM		RF		k-NN (ROC)	
ALQ120QU	6	5	4	6	5	4
BMDAVSAD	15	16		11	6	16
BMXARMC			18	19		
BMXARML	12	11	11	16		
BMXHT	10	19	10	14	19	
BMXLEG	11	15	15	10		20
BMXWAIST		10	6	7	10	15
BMXWT	17		12	17		
BPQ020						11
BPQ080						18
BPQ090D		12			7	5
BPXDIave	4	6	7	8		13
BPXSYave	2	2	3	3	1	7
DMDEDUC2				16		
LBDHDD	13	17	14	15	17	
LBDLDL	3	3	2	2	9	3
LBXTC	20	8	9	5	18	1
LBXTR	1	1	1	1	2	6
MCQ300C		13			20	12
MCQ365B						10
MCQ365C						17
MCQ365D					8	9
MCQ370B	14		16		4	14
PAD630	18	9	19	18		
PAD645	16					
PAD660					14	
PAD675	8	18	17	20		
PAD680	5	7	5	9	11	
PAQ710	7	14	13	12		19
RIDAGEYR	19	4	8	4	3	2
SMQ020				15		
SMQ925A	9		20		13	
WHQ150		20		13	12	8

original scores in grey

3.4.7 Binary classification

Another approach towards multi-class classification problems is to apply binary classification algorithms. The heuristic method of dividing the dataset into a binary classification problem is by separating 1 class versus the rest of classes. The dataset is reorganised to slightly resemble the binary classification problem discussed in Chapter 2 by grouping all participants who are diagnosed or at risk of T2D (classes '1', '2' and '3') as class '1' and participants not at risk of T2D as class '2'. The 3 top performing models of the binary classification problem, RF, XGB trees and Logitboost, and the SVM model which is commonly applied for the 1 versus rest classification are trained using the dataset. This dataset does not undergo SMOTE because classes '1', '2' and '3' add up to 40% of the dataset (110, 13 and 151 respectively).

Table 3.28 documents some improvements to the ROC-AUC for 3 of the models. The results in Table 3.28, however, show that the models perform rather poorly in terms of recall, compared to the models trained using the original binary classification dataset. The above results[21] provide evidence that importance of features weigh in differently for the classes, particularly for class '3'. It is not as accurate to combine participants who are already diabetic and participants who are pre-diabetic and at risk. The intervention methods and treatment advice doctors and healthcare professionals offer to either class will after all, be different.

Table 3.28 Model Comparison – Binary (1 vs Rest)

Model	Recall (%)	ROC-AUC	TP	FP	FN	TN
XGB trees	54.6	0.674	59	49	55	239
	82.7	0.650	43	9	86	264
RF	30.6	0.696	33	75	10	284
	80.8	0.705	42	10	53	297
SVM	25.0	0.754	27	81	10	284
	69.2	0.689	36	16	48	302
Logitboost	41.67	0.676	45	63	30	264
	80.0	0.609	36	9	95	201

original scores in grey

The overall features listed as important in Table 3.29 are fairly similar to the original results in Chapter 2, with the exception of some new features such as *BMXBMIAA*, *MCQ370C* and *PAD645*.

[21] See Table 3.6 and Appendices I and J.

Table 3.29 Significant Features – Binary (1 vs Rest)

Variables	XGB		RF		Logitboost and SVM (ROC)	
ALQ120QU	13	13				
BMDAVSAD	8	10	5	7	3	9
BMXARMC	6	19	19	18	14	
BMXARML	11		20	13	18	
BMXBMIAA			16			
BMXHT	16			16		
BMXLEG	4		13	15		
BMXWAIST	7	11	8	11	2	7
BMXWT			12	14	10	
BPQ020	15	17	7		7	8
BPQ080			11		8	16
BPQ090D	10	2	3	6	5	2
BPXDIave	20	12		9		
BPXSYave	14	5	6	4	4	3
DMDEDUC2						17
INDFMIN2		18				14
LBDHDD	12	9		10	15	13
LBDLDL	2	20	2	19		20
LBXTC	5	4	9	3		11
LBXTR	1	8	1	20	9	19
MCQ300C	18	6	15	12		10
MCQ365A						18
MCQ365B			18		11	12
MCQ365C			17		16	15
MCQ365D		15	14		12	6
MCQ370C					20	
PAD645	19					
PAD675		7		17		
PAD680	17	16		8	19	
PAQ710		3		2	17	5
RIAGENDR					13	
RIDAGEYR	3	1	4	1	1	1
WHQ150	9	14	10	5	6	4

original scores of binary classification in grey

3.5 Findings

One of the more positive findings derived from the above results and discussion is that the models show more stability in suggesting features that are important. The list of features across the different tests do not vary as much as those undertaken in Chapter 2 for binary classification. The features that are common across all or most of the tests conducted include frequency of alcohol intake, sagittal abdominal diameter measurement and daily time recorded of sedentary lifestyle choices. Features like age, waist, weight history, prescribed cholesterol medication, diastolic and systolic blood pressure and cholesterol levels are also common across the test results. It supports the findings in binary classification problem reported in Chapter 2. Their commonality enhances their influence in practical implications, giving healthcare professionals and insurance providers greater confidence to consider these variables with higher scrutiny.

The 3 classifiers undertaken for sensitivity analysis and robustness tests show varied performance. The boosted trees do perform better (recall) in expected scenarios for the class '3' prediabetes prediction. While the overall model performance of the models may improve in terms of ROC-AUC, the recall metric for the minority classes do not. It is more critical that the predictive model is able to more accurately predict the risk of T2D for pre-diabetic individuals. Given that one of the project's objectives is to help reduce risk of T2D, being able to detect prediabetes using lifestyle factors can offer timely intervention to avoid prediabetes progressing to T2D. The RF and k-NN models weigh feature importance according to classes[22] and the relevant industries can consider these features for any related strategies involving T2D.

i. *PAD660*, where it records if the participant is reducing salt in diet
ii. *SMQ020*, where it records if the participant smoked at least 100 cigarettes in life
iii. *SMQ925A*, 1where it records if the participant has smoked a cigarette even 1 time in their life

The findings in the above discussion support the discussion in Chapter 2.5 regarding continuous variables, daily activity, and body measurements versus BMI to better represent obesity in Asian Americans.

[22] See Appendix K.

4.0 Conclusion

The motivation of this project is to present statistically significant factors that can be easily collected from individuals without any special medical procedures when assessing their risk of T2D. Comprehensive tests for a proper diagnosis can be recommended for those who show high risk. The models in this project may not be implemented; rather, the results derived from this project offer insight into developing appropriate and effective strategies tailored to individuals of Asian descent. The findings of this project have practical implications for doctors in practice and insurance providers. These key personnel can confidently use this information to their advantage, making more specialised and data-driven decisions.

The following discussion reviews:

i. the strategic applications of the results derived from supervised machine learning classifiers
ii. how the results can practically impact healthcare advice for preventing T2D
iii. how the results contribute to existing literature using lifestyle and other non-invasive features
iv. limitations of the analysis and discusses the insights for future research

4.1 Implications

Many of the non-invasive features may be correlated such as body measurements and physical activity levels of an individual. Should the strategy developed by healthcare professionals and insurance providers to aid in the prediction involve using machine learning techniques, boosted trees classifiers offer the most stable performance and overcome the inherent multi-collinearity issues. The findings help to discern more effective health and lifestyle factors that affect Asian Americans at risk of T2D, which is unique to this project.

4.1.1 Prevention of T2D

Results of the multi-class prediction observe that the models perform relatively well in indicating features that may help individuals with prediabetes to prevent the condition from progressing towards T2D. The models respond with different features for individuals who are already diagnosed with T2D or have high plasma glucose levels that corresponds to a T2D

diagnosis. The findings help highlight key health indicators and lifestyle choices that are more vital to observe for either type of diagnosis.

Specifically, restricting one's alcohol consumption is more relevant for those who are already diabetic. Factors that form a healthy lifestyle, reduce visceral fat stored around the waist and lower cholesterol levels, are more relevant to help lower the risk of pre-diabetic Asian Americans. These factors comprise of daily exercise to circumvent a sedentary lifestyle and reducing salt intake. Having to sit for long periods of time during a day may be inevitable for individuals with such occupational demands. The results imply that the integration of moderate-intensity levels of movement daily can be enough to reduce the risk of T2D. This suggests that preventive lifestyle changes include being mindful of not letting the sedentary lifestyle take over and taking extra care every day to undo some of the effects.

Doctors and healthcare professionals can adapt the findings of this project to prescribe more effective and personalised advice and treatment for patients who are of Asian descent. Asian individuals who have different body composition, require different means of assessing their health status. Body measurements, specifically waist circumference and sagittal abdominal diameter, are more significant than BMI in the T2D prediction.

4.1.2 Business strategy for insurance providers

Insurance providers assess an individual's risk and determine whether to take them on as clients and what types of coverage and premiums to charge. The findings highlight some questions regarding an individual's medical history and physical activity levels are more valuable than others for predicting T2D risk. Specifically, questions regarding alcohol habits, family history, weight history, exercise and nutritional advice received from doctors, and daily activity levels in particular the minutes/hours spent sitting and minutes/hours spent on moderate intensity work or recreational activities. The above findings include the observation that predictive models respond better to continuous values. This suggest an insight into designing questionnaires that yield more effective collection of data and thus, more accurate predictions.

4.2 Limitations and future research

This project is subject to 3 limitations.

First, the modelling and results are subject to analytic limitations of the dataset. The features are derived from data collected in the survey questions. Participants in the survey are selected using a cluster sampling methodology and are notified in advance of their testing dates. The data is also collected over a 2-year cycle. This do not allow research to take into account any unique circumstances that may impact the data and does not allow for research to track changes over time. The accuracy of the results is also reliant on participants' answers to the survey questions. The NHANES survey takes this into account for factors that can be observed. For example, the survey requests participants to report their own height and weight. The test section of the survey also measures the participants' height and weight by having the survey officials take these measurements. It has been found that participants do lie in the questionnaire.

Second, only 1 summary statistic (average) of relevant feature is included in the model building. This is due to 2 reasons – reducing multicollinearity issues and number of similar features ranked important. Future research can consider employing various (if not all) summary statistics for each feature in feature selection to build a more robust model.

Third, there are trade-offs depending on which metric the model performance is evaluated on. Type 2 errors (i.e. less false negative predictions where patients are predicted to have T2D but do not) are to be minimised in the medical context. The advantage of NHANES survey data is it being population-based, and thus not subject to recall bias.

Future research can consider using features derived from laboratory test results to further improve the performance on the classification.

References

André, Philippe, Balkau, B., Born, C., Charles, M-A., Eschwège, E. and DESIR Study Group. (2006). *Three-year increase of gamma-glutamyl transferase level and development of type 2 diabetes in middle-aged men and women: the DESIR cohort*. Diabetologia, Vol. 49, Issue 11: 2599-2603.

Firouzi, Shelby A., Tucker, L. A., LeCheminant, J. D., and Bailey, B. W. (2018). *Sagittal abdominal Diameter, waist circumference, and BMI as predictors of multiple measures of glucose metabolism: An NHANES investigation of US adults*. Journal of Diabetes Research: 1-14.

Ford, Earl S. (1999). *Body mass index, diabetes, and C-reactive protein among US adults*. Diabetes Care, Vol. 22, Issue 12: 1971-1977.

Ghanvatkar, S., and Rajan, V. (2019). *Deep recurrent neural networks for mortality prediction in intensive care using clinical time series at multiple resolutions*.

Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., and Galstyan, A. (2019). *Multitask learning and benchmarking with clinical time series data*. Scientific Data, Issue 6: 1-18.

Heredia-Langner, Alejandro, Jarman, K. H., Amidan, B. G., and Pounds, J. G. (2013). *Genetic algorithms and classification trees in feature discovery: diabetes and the NHANES database*. In Proceedings of the International Conference on Data Mining (DMIN), p. 1. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2013.

Hsu, William C., Araneta, M. R. G., Kanaya, A. M., Chiang, J. L., and Fujimoto, W. (2015). *BMI cut points to identify at-risk Asian Americans for Type 2 Diabetes screening*. Diabetes Care, Vol. 38: 150-158.

Knerr, Stefan, Personnaz, L., and Dreyfus, G. *Single-layer learning revisited: a stepwise procedure for building and training a neural network*. Neurocomputing, pp. 41-50. Springer, Berlin, Heidelberg, 1990.

Lee, D-H., Ha, M-H., Kim, J-H., Christiani, D. C., Gross, M. D., Steffes, M., Blomhoff, R., and Jacobs, D. R. (2003). *Gamma-glutamyl transferase and diabetes—a 4 year follow-up study*. Diabetologia Vol. 46, Issue 3: 359-364.

Lee, Ji Won R., Brancati, F. L., and Yeh, H. (2011) *Trends in the prevalence of Type 2 Diabetes in Asians versus Whites*. Diabetes Care, Vol. 34: 353-357.

McNeely, Marguerite J., and Boyko, E. J. (2004) *Type 2 Diabetes prevalence in Asian Americans*. Diabetes Care, Vol. 27, Issue 1: 66-69.

Nelson, Karin M., Reiber, G., and Boyko, E.J. (2002). *Diet and exercise among adults with type 2 diabetes: findings from the third national health and nutrition examination survey (NHANES III)*. Diabetes Care, Vol. 25, Issue 10: 1722-1728.

Nguyen, Binh P., Pham, H. N., Tran, H., Nghiem, N., Nguyen Q. H., Do, T. T. T., Tran, C. T., and Simpson, C. R. (2019). *Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records*. Computer Methods and Programs in Biomedicine, Vol. 182: 1-9

Nguyen, Tam H., Nguyen T., Fischer, T., Ha, W., and Tran, T. V. (2015) *Type 2 Diabetes among Asian Americans: Prevalence and prevention*. World Journal of Diabetes, Vol. 6, Issue 4: 543-547.

Omogbai, Aileme. (2016). *Application of multiview techniques to NHANES dataset*. ArXiv abs/1608.04783.

Pimentel, Angela, Carreiro, A. V., Ribeiro, R. T. and Gamboa, H. (2018). *Screening diabetes mellitus 2 based on electronic health records using temporal feature*s. Health Informatics Journal, Vol. 24, Issue 2: 194-205.

Pou, Karla M., Massaro, J.M., Hoffmann, U., Lieb, K., Vasan, R.S., O'Donnell, C.J., and Fox, C.S. (2009). *Patterns of abdominal fat distribution: the Framingham Heart Study*. Diabetes Care, Vol. 32, Issue 3: 481-485.

Pradhan, Aruna D., Manson, J.E., Rifai, N., Buring, J.E., and Ridker, P.M. (2001). *C-reactive protein, interleukin 6, and risk of developing Type 2 Diabetes Mellitus*. Jama, Vol. 286, Issue 3: 327-334.

Sabanayagam, Charumathi, Shankar, A., Li, J., Pollard, C., and Ducatman, A. (2009). *Serum gamma-glutamyl transferase level and diabetes mellitus among US adults*. European Journal of Epidemiology, Vol. 24, Issue 7: 369-373.

Semerdjian, John, and Frank, S. (2017). *An ensemble classifier for predicting the onset of Type II Diabetes*. Cornell University.

Sniderman, Allan D., Bhopal R., Prabhakaran, D., Sarrafzadegan, N., and Tchernof, A. (2007). *Why might South Asians be so susceptible to central obesity and its atherogenic consequences? The adipose tissue overflow hypothesis*. International Journal of Epidemiology, Vol. 36, Issue 1: 220-225.

Wang, Elsie J., Wong, E. C., Dixit, A. A., Fortmann, S. P., Linde, R. B., and Palaniappan, L. P. (2011). *Type 2 diabetes: Identifying high risk Asian American subgroups in a clinical population*. Diabetes Research and Clinical Practice, Volume 93, 248-254.

Wen, Jiangping, Liang, Y., Wang, F., Sun, L., Guo, Y., Duan, X., Liu, X., Wong, T.Y., Lu, X., and Wang, N. (2010). *C-reactive protein, gamma-glutamyl transferase and Type 2 Diabetes in a Chinese population*. Clinica Chimica Acta, Vol. 411, Issue 3-4: 198-203.

WHO expert consultation. (2004). *Appropriate bodymass index for Asian populations and its implications for policy and intervention strategies*. The Lancet, Vol. 363, Issue 9403: 157-163.

University of Chicago Medical Center. (2016). *Asian Americans are at high risk for diabetes but rarely get screened*. ScienceDaily.

Yim, Jeong Yoon, Kim, D., Lim, S. H., Park, M. J., Choi, S. H., Lee, C. H., Kim, S. S., and Cho, S. (2010). *Sagittal abdominal diameter is a strong anthropometric measure of visceral adipose tissue in the Asian general population*. Diabetes Care, Vol. 33, Issue 12: 2665-2670.

Yu, Wei, Liu, T., Valdez, R., Gwinn, M., and Khoury, M.J. (2010). *Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes*. BMC Medical Informatics and Decision Making, Vol. 10, Issue 1: 16.

Appendix A – Variable list

Variable Name	Description
SEQN	Participant Unique Identifier/Merge Key
	Outcome Variables
DIQ010A	1 indicates that the participant has diagnosed/undiagnosed diabetes, 2 otherwise
DIQ010B	1 indicates that the participant has diagnosed diabetes, 2 indicates the participant has undiagnosed diabetes, 3 indicates that the participant has prediabetes, 4 indicates that the participant has no diabetes
	Questionnaire
	Alcohol Use
ALQ101	1 indicates that the participant has had at least 12 alcohol drinks per year, 2 otherwise
ALQ120QU	Number of days participant has had any type of alcoholic beverage in the past 12 months
ALQ130	Average number of alcoholic drinks per day in the past 12 months
ALQ141QU	Number of days participant had 4/5 drinks in the past year (12 months)
	Blood Pressure & Cholesterol
BPQ020	1 indicates that a doctor or health professional has told participant he/she has high blood pressure, 2 otherwise
BPQ080	1 indicates that a doctor or health professional has told participant he/she has high cholesterol level, 2 otherwise
BPQ090D	1 indicates that a doctor or health professional has told participant he/she have to take prescription for cholesterol, 2 otherwise
	Diabetes
DIQ050	1 indicates that the participant is taking insulin now
DIQ160	1 indicates that a doctor or health professional has told participant he/she has prediabetes 2 otherwise
DIQ170	1 indicates that a doctor or health professional has told participant he/she has health risk for diabetes, 2 otherwise
DIQ172	1 indicates that the participant feels he/she has health risk for diabetes, 2 otherwise
DIQ180	1 indicates that the participant has had a blood test for high blood sugar or diabetes within the past three years, 2 otherwise
	Medical Conditions
MCQ092	1 indicates that the participant has ever received a blood transfusion, 2 otherwise
MCQ160B	1 indicates that a doctor or health professional has told participant he/she had congestive heart failure, 2 otherwise
MCQ160F	1 indicates that a doctor or health professional has told participant he/she had a stroke, 2 otherwise
MCQ160K	1 indicates that a doctor or health professional has told participant he/she had chronic bronchitis, 2 otherwise

MCQ160L	1 indicates that a doctor or health professional has told participant he/she has any liver condition, 2 otherwise
MCQ160M	1 indicates that a doctor or health professional has told participant he/she had thyroid problem, 2 otherwise
MCQ160N	1 indicates that a doctor or health professional has told participant he/she has gout, 2 otherwise
MCQ203	1 indicates that a doctor or health professional has told the participant he/she had yellow skin, yellow eyes or jaundice (exclude infant jaundice), 2 otherwise
MCQ300c	Family history of diabetes
MCQ365a	1 indicates that a doctor or health professional has told participant he/she has to lose weight, 2 otherwise
MCQ365b	1 indicates that a doctor or health professional has told participant he/she has to exercise, 2 otherwise
MCQ365c	1 indicates that a doctor or health professional has told participant he/she has to reduce salt in diet, 2 otherwise
MCQ365d	1 indicates that a doctor or health professional has told participant he/she has to reduce fat/calories, 2 otherwise
MCQ370a	1 indicates that the participant is now controlling or losing weight, 2 otherwise
MCQ370b	1 indicates that the participant is now increasing exercise, 2 otherwise
MCQ370c	1 indicates that the participant is now reducing salt in diet, 2 otherwise
MCQ370d	1 indicates that the participant is now reducing fat in diet, 2 otherwise

Physical Activity

PAQ605	1 indicates that the participant work (such as paid or unpaid work, household chores, and yard work) involve vigorous-intensity activity that causes large increases in breathing or heart rate like carrying or lifting heavy loads, digging or construction work for at least 10 minutes continuously per day in a typical week, 2 otherwise
PAD615	How much time in minutes do the participant spend doing vigorous-intensity activities at work on a typical day?
PAQ620	1 indicates that the participant work (such as paid or unpaid work, household chores, and yard work) involve moderate-intensity activity that causes small increases in breathing or heart rate such as brisk walking or carrying light loads for at least 10 minutes continuously per day in a typical week, 2 otherwise
PAD630	How much time in minutes do the participant spend doing moderate-intensity activities at work on a typical day?
PAQ635	1 indicates that the participant walks or uses a bicycle for at least 10 minutes continuously to get to and from places in a typical week, 2 otherwise
PAD645	How much time in minutes do the participant spend walking or bicycling for travel on a typical day?
PAQ650	1 indicates that the participant does any vigorous-intensity sports, fitness, or recreational activities that cause large increases in breathing or heart rate like running or basketball for at least 10 minutes continuously in a typical week, 2 otherwise
PAD660	How much time in minutes do the participant spend doing vigorous-intensity sports, fitness or recreational activities on a typical day?
PAQ665	1 indicates that the participant does any moderate-intensity sports, fitness, or recreational activities that cause a small increase in breathing or heart rate such as brisk walking, bicycling, swimming, or volleyball for at least 10 minutes continuously in a typical week, 2 otherwise

PAD675	How much time in minutes do the participant spend doing moderate-intensity sports, fitness or recreational activities on a typical day?
PAD680	How much time in minutes do the participant spend sitting (such as sitting at school, at home, getting to and from places, or with friends including time spent sitting at a desk, traveling in a car or bus, reading, playing cards, watching television, or using a computer; do not include time spent sleeping) on a typical day?
PAQ710	How much time in hours do the participant spend sitting and watching TV or videos over the past 30 days on a typical day?
	Weight History
WHQ060	1 indicates that the change between the participant's current weight and his/hers weight a year ago is intentional, 2 otherwise
WHQ070	1 indicates that the participant has tried to lose weight during the past 12 months, 2 otherwise
WHQ150	How old was the participant at their heaviest weight?
	Smoking - Cigarette Use
SMQ020	1 indicates that the participant smoked at least 100 cigarettes in life, 2 otherwise
SMQ040A	1 indicates that the participant smoke cigarettes now, 2 otherwise
SMQ910	1 indicates that the participant has used smokeless tobacco (include chewing tobacco, snuff, dip, snus and dissolvable tobacco) before, 2 otherwise
SMQ915	How many days has the participant used smokeless tobacco over the past 30 days?
SMQ915A	1 indicates that the participant has used smokeless tobacco in the last 30 days, 2 otherwise
SMQ925A	1 indicates that the participant has smoked a cigarette even 1 time in their life, 2 otherwise
	Laboratory
	Cholesterol - Total
LBXTC	Total Cholesterol (mg/dL)
LBXTCA	1 indicates that the participant total cholesterol level (>=200) is at risk, 2 otherwise
	Cholesterol - High-Density Lipoprotein (HDL)
LBDHDD	Direct HDL-Cholesterol (mg/dL)
LBDHDDA	1 indicates that the participant HDD cholesterol level (<60) is at risk, 2 otherwise
	Cholesterol - Low - Density Lipoprotein (LDL) & Triglycerides
LBXTR	Triglyceride (mg/dL)
LBXTRA	1 indicates that the participant Triglyceride level (>=150) is at risk, 2 otherwise
LBDLDL	LDL-cholesterol (mg/dL)
LBDLDLA	1 indicates that the participant LDL cholesterol level (>=100) is at risk, 2 otherwise
	Complete Blood Count with 5-Part Differential - Whole Blood
LBXWBCSI	White blood cell count (1000 cells/uL)
LBXPLTSI	Platelet count (1000 cells/uL)
LBXRBCSI	Red blood cell count (million cells/uL)
LBXLYPCT	Lymphocyte percent (%)

LBXMOPCT	Monocyte percent (%)
LBXNEPCT	Segmented neutrophils percent (%)
LBXEOPCT	Eosinophils percent (%)
LBXBAPCT	Basophils percent (%)
LBDLYMNO	Lymphocyte number (1000 cells/uL)
LBDMONO	Monocyte number (1000 cells/uL)
LBDNENO	Segmented neutrophils num (1000 cell/uL)
LBDEONO	Eosinophils number (1000 cells/uL)
LBDBANO	Basophils number (1000 cells/uL)
LBXHGB	Hemoglobin (g/dL)
LBXHCT	Hematocrit (%)
LBXMCVSI	Mean cell volume (fL)
LBXMCHSI	Mean cell hemoglobin (pg)
LBXMC	Mean Cell Hemoglobin Concentration (g/dL)
LBXRDW	Red cell distribution width (%)
LBXMPSI	Mean platelet volume (fL)

Glycohemoglobin - HbA1c

LBXGH	Glycohemoglobin (%)

Plasma Fasting Glucose

LBXGLU	Fasting Glucose (mg/dL)

Insulin

LBXIN	Insulin (uU/mL)

Standard Biochemistry Profile

LBXSAL	Albumin, refrigerated serum (g/dL)
LBXSAPSI	Alkaline Phosphatase (ALP) (IU/L)
LBXSASSI	Aspartate Aminotransferase (AST) (IU/L)
LBXSATSI	Alanine Aminotransferase (ALT) (IU/L)
LBXSBU	Blood Urea Nitrogen (mg/dL)
LBXSC3SI	Bicarbonate (mmol/L)
LBXSCA	Total Calcium (mg/dL)
LBXSCH	Cholesterol, refrigerated serum (mg/dL)
LBXSCK	Creatine Phosphokinase (CPK) (IU/L)
LBXSCLSI	Chloride (mmol/L)
LBXSCR	Creatinine, refrigerated serum (mg/dL)
LBXSGB	Globulin (g/dL)
LBXSGL	Glucose, refrigerated serum (mg/dL)
LBXSGTSI	Gamma Glutamyl Transferase (GGT) (U/L)
LBXSIR	Iron, refrigerated serum (ug/dL)
LBXSKSI	Potassium (mmol/L)
LBXSLDSI	Lactate Dehydrogenase (LDH) (U/L)
LBXSNASI	Sodium (mmol/L)
LBXSOSSI	Osmolality (mmol/Kg)
LBXSPH	Phosphorus (mg/dL)
LBXSTB	Total Bilirubin (mg/dL)

LBXSTP	Total Protein (g/dL)
LBXSTR	Triglycerides, refrig serum (mg/dL)
LBXSUA	Uric acid (mg/dL)

Examination

Blood Pressure

BPXSY1	Systolic: Blood pressure (first reading) mm Hg
BPXSY2	Systolic: Blood pressure (second reading) mm Hg
BPXSY3	Systolic: Blood pressure (third reading) mm Hg
BPXSYmax	Maximum reading out of the multiple individual readings for the participant
BPXSYmin	Minimum reading out of the multiple individual readings for the participant
BPXSYave	Average reading out of the multiple individual readings for the participant
BPXSYaveC	1 indicates that the participant average Systolic BP level (>=120) is at risk, 2 otherwise
BPXDI1	Diastolic: Blood pressure (first reading) mm Hg
BPXDI2	Diastolic: Blood pressure (second reading) mm Hg
BPXDI3	Diastolic: Blood pressure (third reading) mm Hg
BPXDImax	Maximum reading out of the multiple individual readings for the participant
BPXDImin	Minimum reading out of the multiple individual readings for the participant
BPXDIave	Average reading out of the multiple individual readings for the participant
BPXDIaveC	1 indicates that the participant average Diastolic BP level (>=80) is at risk, 2 otherwise
BPXPULS	1 indicates that the participant has a regular pulse, 2 otherwise

Body Measures

BMXWT	Weight (kg)
BMXHT	Standing Height (cm)
BMXBMI	Body Mass Index (kg/m**2)
BMXBMIAA	Obese: BMXBMI >= 27 Overweight: 23 ≤ BMXBMI < 27 Normal: 18.5 ≤ BMXBMI < 23 Underweight: BMXBMI < 18.5
BMXBMIO	Obese: BMXBMI >= 30 Overweight: 25 ≤ BMXBMI < 30 Normal: 18.5 ≤ BMXBMI < 25 Underweight: BMXBMI < 18.5
BMXLEG	Upper Leg Length (cm)
BMXARML	Upper Arm Length (cm)
BMXARMC	Arm Circumference (cm)
BMXWAIST	Waist Circumference (cm)
BMXSAD1	Sagittal Abdominal Diameter 1st (cm)
BMDAVSAD	Average Sagittal Abdominal Diameter (cm)
BMXSADmax	Maximum reading out of the multiple individual readings for the participant
BMXSADmin	Minimum reading out of the multiple individual readings for the participant

Demographics

RIAGENDR	Gender of the participant.

RIDAGEYR	Age in years at screening
DMDEDUC2	Education level - Adults 20+
INDFMIN2	Annual family income

*Variables in light grey are removed through starter selection process and feature selection process after cleaning.

*Variables in light blue are used in sensitivity analysis for validation.

Appendix B – Features engineered

Variable Name	Cleaning	Feature Engineering	New Variable Name
ALQ141Q *ALQ141U*	*ALQ141U* represents the units of measure to be combined with the numerical value recorded in *ALQ141Q* for meaningfulness. *ALQ141U* consists of 3 options – week, month and year. Rows where *ALQ141U* indicates week will be multiplied by 52. Rows where *ALQ141U* indicates month will be multiplied by 12. The values in *ALQ141Q* will then reflect 1 standardised unit. It will be renamed under a new variable.		**ALQ141QU**, where the recorded values indicate the number of days a participant had 4/5 drinks during the year (i.e. days per year).
DIQ010	This is the outcome categorical variable where participants are asked if they were told by doctor or health professional that they have diabetes. As there are 3 recorded values for this original variable where 1 indicates 'yes', 2 indicates 'no', and 3 indicates 'borderline', the value '3' will be recoded to '1' as the participant is considered to have diabetes. If the participant's reading for the Fasting Plasma Glucose (FPG) test (*LBXGLU*) is available, it will also be used to help in cleaning this variable. If the participant is not told by a doctor (coded '2' in *DIQ010*) but has a FPG reading of 126 mg/dL and above, the new variable will be recoded to 1.		**DIQ010A**, where 1 indicates that the participant has diabetes (coded '1' or '3' in *DIQ010* or *LBXGLU* >= 126), and 2 indicates the participant has no diabetes (coded '2' in *DIQ010* or *LBXGLU* < 126)
		Another possible outcome variable is feature engineered to reflect more accurate diagnosis of a participant diabetes status. The participant's reading FPG test (*LBXGLU*) is also used in this step. There are 4 categories in this new variable.	**DIQ010B**, where: 1 – the participant has diagnosed diabetes (coded '1' or '3' in *DIQ010* and *LBXGLU* >= 126) 2 – the participant has undiagnosed diabetes (coded '2' in *DIQ010* and *LBXGLU* >= 126) 3 – the participant has prediabetes (100 < *LBXGLU* < 125) 4 – the participant has no diabetes (*LBXGLU* <= 100)

SMQ040	*SMQ040* consists of 3 options – everyday (1), some days (2) and not at all (3) – to the question if participants smoke now. To better reflect if the participant is a smoker or not, rows with value '2'are recoded to '1' and rows with value '3' are recoded to '2'. This is consistent to the other categorical variables where 1 indicates yes, and 2 indicates no. NA values are treated with median.		**SMOQ040A**, where 1 indicates the participant is a smoker, and 2 otherwise.
SMQ915	NA values are treated with median.	*SMQ915* records continuous values to the question of the number of days a participant used smokeless tobacco during the last 30 days. To better reflect if the participant used smokeless tobacco or not, rows with values 1 to 30 are recoded to '1' and rows with value '0' are recoded to '2'. This is consistent to the other categorical variables where 1 indicates yes, and 2 indicates no.	**SMQ915A**, where 1 indicates the participant use smokeless tobacco, and 2 otherwise.
LBXTC	NA values are treated with median.	A categorical variable is created to reflect if the participant's total cholesterol level is at risk. If *LBXTC* >= 200, recode to 1 (at risk). If *LBXTC* < 200, recode to 2 (normal/not at risk).	**LBXTCA**, where 1 indicates the participant is at risk, 2 otherwise.
LBXTR	NA values are treated with median.	A categorical variable is created to reflect if the participant's triglyceride level is at risk. If *LBXTR* >= 150, recode to 1 (at risk). If *LBXTC* < 150, recode to 2 (normal/not at risk).	**LBXTRA**, where 1 indicates the participant is at risk, 2 otherwise.
LBDLDL	NA values are treated with median.	A categorical variable is created to reflect if the participant's LDL (bad) cholesterol level is at risk. If *LBDLDL* >= 100, recode to 1 (at risk). If *LBDLDL* < 100, recode to 2 (normal/not at risk).	**LBDLDLA**, where 1 indicates the participant is at risk, 2 otherwise.
LBDHDD	NA values are treated with median.	A categorical variable is created to reflect if the participant's HDD (good) cholesterol level is at risk. If *LBDHDD* < 60, recode to 1 (at	**LBDHDDA**, where 1 indicates the participant is at risk, 2 otherwise.

	risk). If *LBDHDD* >= 60, recode to 2 (good).

BPXSY1 *BPXSY2* *BPXSY3* *BPXSY4*		A continuous variable is created to reflect the highest (maximum) systolic blood pressure reading the participant has recorded if taken the test more than once. Most participants are required to take the test 3 times.	**BPXSYmax**
		A continuous variable is created to reflect the lowest (minimum) systolic blood pressure reading the participant has recorded if taken the test more than once. Most participants are required to take the test 3 times.	**BPXSYmin**
		A continuous variable is created to reflect the average of multiple systolic blood pressure readings the participant has recorded. Most participants are required to take the test 3 times.	**BPXSYave**
		A categorical variable is created to reflect if the participant's <u>average</u> systolic blood pressure level is at risk. If *BPXSYave* >= 120, recode to 1 (high/at risk). If *BPXSYave* < 120, recode to 2 (normal/not at risk).	**BPXSYaveC**, where 1 indicates the participant is at risk, 2 otherwise.
	NA values are treated with median after the above steps are completed.		
BPXDI1 *BPXDI2* *BPXDI3* *BPXDI4*	Convert rows with 0 values in *BPXDI1*, *BPXDI2*, *BPXDI3* to NA values.		
		A continuous variable is created to reflect the highest (maximum) diastolic blood pressure reading the participant has recorded if taken the test more than once. Most participants are required to take the test 3 times.	**BPXDImax**
		A continuous variable is created to reflect the lowest (minimum) diastolic blood pressure reading the participant has recorded if taken the test more than once.	**BPXDImin**

Most participants are required to take the test 3 times.

BMXSAD1 *BMXSAD2* *BMXSAD3* *BMXSAD4*		A continuous variable is created to reflect the highest (maximum) sagittal abdominal diameter reading the participant has recorded if taken the test more than once. Most participants are required to take the test 2 times.	**BMXSADmax**
		A continuous variable is created to reflect the lowest (minimum) sagittal abdominal diameter reading the participant has recorded if taken the test more than once. Most participants are required to take the test 2 times.	**BMXSADmin**
	NA values are treated with median after the above steps are completed.		

Appendix C – Feature selection

using pre-processed dataset

k-NN

ROC curve variable importance

only 20 most important variables shown (out of 100)

	Importance
RIDAGEYR	0.7938
BPQ090D	0.7474
WHQ150	0.6874
BPXSYmax	0.6770
BMXWAIST	0.6749
LBXSOSSI	0.6725
BPXSY1	0.6715
BPQ080	0.6707
MCQ365B	0.6702
BPXSYave	0.6687
BMXSAD1	0.6682
BMXSADmin	0.6681
BMXSADmax	0.6668
BMDAVSAD	0.6668
BPQ020	0.6668
MCQ365D	0.6644
BPXSY2	0.6631
MCQ300C	0.6592
BPXSYmin	0.6541
BPXSY3	0.6501

ROC curve variable importance

variables are sorted by maximum importance across the classes

only 20 most important variables shown (out of 100)

	X1	X2	X3	X4
RIDAGEYR	0.7545	0.8189	0.6444	0.8189
LBXTC	0.7794	0.7794	0.7794	0.6778
BPQ090D	0.7312	0.7686	0.6350	0.7686
LBDLDL	0.7636	0.7636	0.7636	0.6269
LBXSCH	0.7552	0.7552	0.7552	0.6742
WHQ150	0.6458	0.7172	0.6458	0.7172
BPXSY3	0.6364	0.6540	0.7066	0.6540
BPXDI3	0.7059	0.7059	0.7059	0.6149
BMXWAIST	0.5914	0.7049	0.5706	0.7049
MCQ300C	0.7035	0.6611	0.6116	0.7035
BPXSYaveC	0.6479	0.6479	0.6999	0.6384
BMXSAD1	0.5721	0.6981	0.5357	0.6981
BMXSADmax	0.5733	0.6962	0.5378	0.6962
BMXSADmin	0.5726	0.6961	0.5428	0.6961
BPXSYmin	0.6203	0.6562	0.6961	0.6562
BMDAVSAD	0.5718	0.6959	0.5357	0.6959
BPQ020	0.6234	0.6931	0.6234	0.6931
BPXSYave	0.6119	0.6754	0.6885	0.6754
LBXSOSSI	0.6506	0.6881	0.6021	0.6881
BPXSYmax	0.5884	0.6868	0.6691	0.6868

RFE

Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

Variables	Accuracy	Kappa	AccuracySD	KappaSD	Selected
4	0.8712	0.2894	0.01606	0.11539	
8	0.8872	0.3580	0.02096	0.11555	
16	0.8872	0.3632	0.01545	0.07953	
100	0.8893	0.2621	0.02031	0.16824	*

The top 5 variables (out of 100):

BPQ090D, RIDAGEYR, LBXSCH, LBXTC, LBXSOSSI

```
 [1] "BPQ090D"    "RIDAGEYR"   "LBXSCH"     "LBXTC"      "LBXSOSSI"   "MCQ300C"    "MCQ365D"    "LBXSCLSI"   "LBDLDL"     "BPQ020"     "BPQ080"     "MCQ365B"
[13] "BPXSY1"     "WHQ150"     "LBXTR"      "LBXSNASI"   "LBXSTR"     "LBXSCR"     "LBXSGTSI"   "BPXSY2"     "BMXWAIST"   "BPXSYmax"   "BMXSAD1"    "LBXSPH"
[25] "BMXSADmax"  "BPXSYave"   "BMDAVSAD"   "BPXDI3"     "LBXTCA"     "BMXSADmin"  "BMXLEG"     "MCQ365A"    "LBXRBCSI"   "BPXDIave"   "BPXDImax"   "LBXSCA"
[37] "BPXSYmin"   "LBDLDLA"    "MCQ365C"    "BPXDImin"   "BMXWT"      "LBXSBU"     "LBXSKSI"    "LBDHDD"     "LBXSAPSI"   "BMXBMI"     "BPXSY3"     "BMXARMC"
[49] "LBXSUA"     "BPXDI1"     "LBXSAL"     "LBXSASSI"   "LBXTRA"     "BMXBMIO"    "LBXSATSI"   "BMXARML"    "BMXHT"      "BPXSYaveC"  "MCQ370C"    "LBXSCK"
[61] "INDHHIN2"   "MCQ160N"    "MCQ370D"    "BPXDI2"     "LBXSTB"     "RIAGENDR"   "LBXSTP"     "BMXBMIAA"   "DMDEDUC2"   "MCQ370A"    "LBXSLDSI"   "LBDHDDA"
[73] "PAD675"     "PAD660"     "BPXDIaveC"  "PAQ605"     "PAD680"     "PAQ650"     "LBXSIR"     "SEQN"       "WHQ070"     "LBXWBCSI"   "ALQ130"     "LBXSC3SI"
[85] "ALQ141QU"   "LBXSGB"     "LBXPLTSI"   "MCQ370B"    "SMQ910"     "MCQ160L"    "SMQ915"     "SMQ915A"    "PAQ665"     "PAD615"     "WHQ060"     "PAQ620"
[97] "SMQ040A"    "PAD645"     "PAD630"     "PAQ635"
```

Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

Variables	Accuracy	Kappa	AccuracySD	KappaSD	Selected
4	0.7254	0.3093	0.03335	0.08528	
8	0.7338	0.3143	0.01556	0.03940	
16	0.7454	0.3455	0.02884	0.09452	*
100	0.7412	0.2581	0.03578	0.12073	

The top 5 variables (out of 16):

LBXTR, LBDLDL, BPQ090D, RIDAGEYR, LBXTRA

```
 [1] "LBXTR"    "LBDLDL"   "BPQ090D"  "RIDAGEYR" "LBXTRA"   "LBXTC"    "LBXSOSSI" "LBXSCH"   "LBDLDLA"  "LBXSCR"   "MCQ300C"  "MCQ365D"  "BPQ020"
[14] "LBXSGTSI" "BMXSAD1"  "LBXRBCSI"
```

Neural Networks

nnet variable importance

only 20 most important variables shown (out of 119)

	Overall
RIDAGEYR	100.00
LBXSOSSI	97.80
PAQ650	82.09
WHQ070	81.74
MCQ300C	77.53
LBDLDLA2	75.95
BMXBMIO3	75.10
INDHHIN23	70.77
INDHHIN214	70.62
MCQ370A2	69.10
DMDEDUC22	67.15
LBXSCLSI	67.03
BMXBMIO2	66.34
BPXDIaveC2	63.76
INDHHIN212	63.18
LBXSNASI	62.23
LBXTCA2	61.81
BPQ090D2	61.64
RIAGENDR	61.21
BPQ080	60.47

nnet variable importance

variables are sorted by maximum importance across the classes

only 20 most important variables shown (out of 119)

	Overall	1	2	3	4
LBXSOSSI	100.00	100.00	100.00	100.00	100.00
LBXSNASI	71.58	71.58	71.58	71.58	71.58
LBXSLDSI	31.58	31.58	31.58	31.58	31.58
BPXSY3	30.93	30.93	30.93	30.93	30.93
LBDLDL	30.83	30.83	30.83	30.83	30.83
RIDAGEYR	27.20	27.20	27.20	27.20	27.20
BPXDI2	26.04	26.04	26.04	26.04	26.04
LBXSCLSI	25.85	25.85	25.85	25.85	25.85
BMXWT	25.03	25.03	25.03	25.03	25.03
PAD645	23.14	23.14	23.14	23.14	23.14
LBXSASSI	22.85	22.85	22.85	22.85	22.85
LBXSATSI	19.37	19.37	19.37	19.37	19.37
BPXSYmax	18.26	18.26	18.26	18.26	18.26
BMXARML	17.21	17.21	17.21	17.21	17.21
LBXSCH	16.78	16.78	16.78	16.78	16.78
LBXTR	16.77	16.77	16.77	16.77	16.77
LBXSBU	16.73	16.73	16.73	16.73	16.73
BPXDI3	15.90	15.90	15.90	15.90	15.90
LBXSKSI	15.11	15.11	15.11	15.11	15.11
BPXSYmin	15.04	15.04	15.04	15.04	15.04

Appendix D – Hyperparameters tuning

Logitboost (packages: caTools and caret)

```
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 1148, 1148, 1148
Resampling results across tuning parameters:

  nIter  Accuracy   Kappa
  21     0.8664344  0.7328688
  37     0.8768873  0.7537747
  60     0.9257420  0.8514241
  74     0.9267855  0.8535862

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was nIter = 74.
```

Random forest (packages: randomForest and caret)

```
Random Forest

1722 samples
  56 predictor
   2 classes: '1', '2'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 1377, 1378, 1378, 1377, 1378
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
  1     0.9210246  0.8420543
  2     0.9494759  0.8989533
  3     0.9506387  0.9012787
  4     0.9506404  0.9012823

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 4
```

Decision trees (packages: rpart, rattle and caret)

```
Call:
rpart(formula = DIQ010A ~ ., data = balancedtrain_A1, method = "class",
    model = TRUE, parms = list(split = "information"), control = rpart.control(
        minbucket = 7, usesurrogate = 0, maxsurrogate = 0))
  n= 1722

          CP nsplit rel error    xerror       xstd
1 0.49012776      0 1.0000000 1.0418118 0.02407706
2 0.05981417      1 0.5098722 0.5493612 0.02151256
3 0.03368177      3 0.3902439 0.4297329 0.01979564
4 0.02555168      4 0.3565621 0.4123113 0.01949749
5 0.01742160      5 0.3310105 0.3960511 0.01920677
6 0.01451800      6 0.3135889 0.3658537 0.01863299
7 0.01000000      8 0.2845528 0.3368177 0.01803644
```

XGB trees (package: caret)

eXtreme Gradient Boosting

1722 samples
56 predictor
2 classes: '1', '2'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1550, 1550, 1550, 1549, 1550, 1550, ...
Resampling results across tuning parameters:

eta	max_depth	colsample_bytree	subsample	nrounds	Accuracy	Kappa
0.3	1	0.6	0.50	50	0.8931409	0.7862678
0.3	1	0.6	0.50	100	0.9122933	0.8245853
0.3	1	0.6	0.50	150	0.9163564	0.8327093
0.3	1	0.6	0.75	50	0.8896659	0.7793280
0.3	1	0.6	0.75	100	0.9059148	0.8118312
0.3	1	0.6	0.75	150	0.9186920	0.8373846
0.3	1	0.6	1.00	50	0.8902507	0.7804999
0.3	1	0.6	1.00	100	0.9099879	0.8199782
0.3	1	0.6	1.00	150	0.9181073	0.8362178
0.3	1	0.8	0.50	50	0.8925729	0.7851387
0.3	1	0.8	0.50	100	0.9088218	0.8176383
0.3	1	0.8	0.50	150	0.9146290	0.8292513
0.3	1	0.8	0.75	50	0.8954833	0.7909625
0.3	1	0.8	0.75	100	0.9099879	0.8199782
0.3	1	0.8	0.75	150	0.9123068	0.8246154
0.3	1	0.8	1.00	50	0.8966427	0.7932793
0.3	1	0.8	1.00	100	0.9129016	0.8258031
0.3	1	0.8	1.00	150	0.9198481	0.8397004
0.3	2	0.6	0.50	50	0.9262367	0.8524701
0.3	2	0.6	0.50	100	0.9442432	0.8884865
0.3	2	0.6	0.50	150	0.9535421	0.9070821
0.3	2	0.6	0.75	50	0.9279944	0.8559837
0.3	2	0.6	0.75	100	0.9401936	0.8803812
0.3	2	0.6	0.75	150	0.9459974	0.8919904
0.3	2	0.6	1.00	50	0.9337982	0.8675873
0.3	2	0.6	1.00	100	0.9488809	0.8977606
0.3	2	0.6	1.00	150	0.9564458	0.9128884
0.3	2	0.8	0.50	50	0.9227618	0.8455170
0.3	2	0.8	0.50	100	0.9436786	0.8873563
0.3	2	0.8	0.50	150	0.9442600	0.8885151
0.3	2	0.8	0.75	50	0.9297318	0.8594627
0.3	2	0.8	0.75	100	0.9541235	0.9082443
0.3	2	0.8	0.75	150	0.9558644	0.9117256
0.3	2	0.8	1.00	50	0.9326422	0.8652755
0.3	2	0.8	1.00	100	0.9471535	0.8943060
0.3	2	0.8	1.00	150	0.9564491	0.9128938
0.3	3	0.6	0.50	50	0.9384460	0.8768878
0.3	3	0.6	0.50	100	0.9529742	0.9059458
0.3	3	0.6	0.50	150	0.9564558	0.9129062
0.3	3	0.6	0.75	50	0.9512266	0.9024502
0.3	3	0.6	0.75	100	0.9605189	0.9210345
0.3	3	0.6	0.75	150	0.9576153	0.9152263
0.3	3	0.6	1.00	50	0.9547217	0.9094364
0.3	3	0.6	1.00	100	0.9616850	0.9233669
0.3	3	0.6	1.00	150	0.9564558	0.9129073
0.3	3	0.8	0.50	50	0.9430938	0.8861841
0.3	3	0.8	0.50	100	0.9459874	0.8919731
0.3	3	0.8	0.50	150	0.9500605	0.9001189
0.3	3	0.8	0.75	50	0.9477450	0.8954862
0.3	3	0.8	0.75	100	0.9570339	0.9140645
0.3	3	0.8	0.75	150	0.9576153	0.9152264
0.3	3	0.8	1.00	50	0.9570339	0.9140645
0.3	3	0.8	1.00	100	0.9582135	0.9164218
0.3	3	0.8	1.00	150	0.9576287	0.9152496
0.4	1	0.6	0.50	50	0.8948851	0.7897780
0.4	1	0.6	0.50	100	0.9111339	0.8222631
0.4	1	0.6	0.50	150	0.9250773	0.8501487
0.4	1	0.6	0.75	50	0.8983835	0.7967591
0.4	1	0.6	0.75	100	0.9105626	0.8211270
0.4	1	0.6	0.75	150	0.9227584	0.8455177
0.4	1	0.6	1.00	50	0.8966461	0.7932943
0.4	1	0.6	1.00	100	0.9163698	0.8327410
0.4	1	0.6	1.00	150	0.9204295	0.8408585
0.4	1	0.8	0.50	50	0.8983768	0.7967398
0.4	1	0.8	0.50	100	0.9094166	0.8188207
0.4	1	0.8	0.50	150	0.9221838	0.8443607
0.4	1	0.8	0.75	50	0.8977887	0.7955759
0.4	1	0.8	0.75	100	0.9140509	0.8280965
0.4	1	0.8	0.75	150	0.9227517	0.8455018
0.4	1	0.8	1.00	50	0.9007158	0.8014272
0.4	1	0.8	1.00	100	0.9169478	0.8338965
0.4	1	0.8	1.00	150	0.9186853	0.8373709
0.4	2	0.6	0.50	50	0.9291605	0.8583143
0.4	2	0.6	0.50	100	0.9489011	0.8977964
0.4	2	0.6	0.50	150	0.9465788	0.8931492
0.4	2	0.6	0.75	50	0.9274197	0.8548358
0.4	2	0.6	0.75	100	0.9488943	0.8977844
0.4	2	0.6	0.75	150	0.9535455	0.9070861
0.4	2	0.6	1.00	50	0.9285623	0.8571267
0.4	2	0.6	1.00	100	0.9506419	0.9012845
0.4	2	0.6	1.00	150	0.9570272	0.9140510
0.4	2	0.8	0.50	50	0.9337915	0.8675781
0.4	2	0.8	0.50	100	0.9442499	0.8884953
0.4	2	0.8	0.50	150	0.9489011	0.8977971
0.4	2	0.8	0.75	50	0.9338016	0.8675967
0.4	2	0.8	0.75	100	0.9483163	0.8966273
0.4	2	0.8	0.75	150	0.9523928	0.9047793
0.4	2	0.8	1.00	50	0.9378747	0.8757439
0.4	2	0.8	1.00	100	0.9558711	0.9117363
0.4	2	0.8	1.00	150	0.9558644	0.9117239
0.4	3	0.6	0.50	50	0.9471636	0.8943266
0.4	3	0.6	0.50	100	0.9558744	0.9117436
0.4	3	0.6	0.50	150	0.9564525	0.9129008
0.4	3	0.6	0.75	50	0.9523827	0.9047596
0.4	3	0.6	0.75	100	0.9576153	0.9152276
0.4	3	0.6	0.75	150	0.9581967	0.9163897
0.4	3	0.6	1.00	50	0.9541303	0.9082533
0.4	3	0.6	1.00	100	0.9552964	0.9105860
0.4	3	0.6	1.00	150	0.9552931	0.9105808
0.4	3	0.8	0.50	50	0.9413429	0.8826804
0.4	3	0.8	0.50	100	0.9489044	0.8978037
0.4	3	0.8	0.50	150	0.9518047	0.9036037
0.4	3	0.8	0.75	50	0.9494791	0.8989527
0.4	3	0.8	0.75	100	0.9552863	0.9105689
0.4	3	0.8	0.75	150	0.9587680	0.9175316
0.4	3	0.8	1.00	50	0.9570372	0.9140676
0.4	3	0.8	1.00	100	0.9570406	0.9140745
0.4	3	0.8	1.00	150	0.9582000	0.9163947

Tuning parameter 'gamma' was held constant at a value of 0
Tuning parameter 'min_child_weight' was held constant at a value of 1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were nrounds = 100, max_depth = 3, eta = 0.3, gamma = 0, colsample_bytree = 0.6, min_child_weight = 1 and subsample = 1.

GBM trees (packages: gbm and caret)

```
Stochastic Gradient Boosting

1722 samples
  56 predictor
   2 classes: '1', '2'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1550, 1550, 1549, 1550, 1549, 1550, ...
Resampling results across tuning parameters:

  interaction.depth  n.trees  Accuracy   Kappa
  1                   50      0.8594603  0.7189173
  1                  100      0.8809417  0.7618794
  1                  150      0.8983667  0.7967339
  2                   50      0.8832773  0.7665482
  2                  100      0.9140577  0.8281142
  2                  150      0.9245060  0.8490110
  3                   50      0.8972073  0.7944142
  3                  100      0.9308778  0.8617545
  3                  150      0.9419075  0.8838148

Tuning parameter 'shrinkage' was held constant at a value of 0.1
Tuning parameter 'n.minobsinnode' was held constant at a value of 10
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were n.trees = 150, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10.
```

Adaboost trees (packages: fastAdaboost and caret)

```
Boosted Classification Trees

1722 samples
  56 predictor
   2 classes: '1', '2'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1550, 1550, 1550, 1549, 1550, 1549, ...
Resampling results across tuning parameters:

  maxdepth  iter  Accuracy   Kappa
  1          50   0.8368228  0.6736557
  1         100   0.8548528  0.7097184
  1         150   0.8612313  0.7224727
  2          50   0.8774970  0.7550005
  2         100   0.8920083  0.7840250
  2         150   0.9042109  0.8084207
  3          50   0.9018820  0.8037589
  3         100   0.9181476  0.8362912
  3         150   0.9274365  0.8548696

Tuning parameter 'nu' was held constant at a value of 0.1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were iter = 150, maxdepth = 3 and nu = 0.1.
```

Neural network (packages: neuralnet and caret)

```
Neural Network

1722 samples
  56 predictor
   2 classes: '1', '2'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 1722, 1722, 1722, 1722, 1722, 1722, ...
Resampling results across tuning parameters:

  size  decay  Accuracy   Kappa
  1     0.000  0.8579735  0.7152656
  1     0.001  0.8576459  0.7149672
  1     0.100  0.8711181  0.7420818
  3     0.000  0.8701670  0.7398946
  3     0.001  0.8788661  0.7574784
  3     0.100  0.8781720  0.7561019
  5     0.000  0.8767498  0.7534439
  5     0.001  0.8746798  0.7490980
  5     0.100  0.8826873  0.7652034
  7     0.000  0.8757952  0.7512707
  7     0.001  0.8779316  0.7556517
  7     0.100  0.8953968  0.7905798
  9     0.000  0.8794184  0.7585223
  9     0.001  0.8808556  0.7614580
  9     0.100  0.8981876  0.7962341

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were size = 9 and decay = 0.1.

nnets[["finalModel"]]
a 73-9-1 network with 676 weights
output(s): .outcome
options were - entropy fitting  decay=0.1
```

SVM (package: caret)

```
Support Vector Machines with Linear Kernel

1722 samples
56 predictor
2 classes: '1', '2'

Pre-processing: centered (73), scaled (73)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 1550, 1550, 1550, 1549, 1550, 1550, ...
Resampling results across tuning parameters:

  C     Accuracy   Kappa
0.00         NaN        NaN
0.01   0.8853878  0.7707703
0.05   0.8768708  0.7537373
0.10   0.8758984  0.7517894
0.25   0.8747379  0.7494740
0.50   0.8728055  0.7456106
0.75   0.8731953  0.7463891
1.00   0.8724212  0.7448404
1.25   0.8718399  0.7436776
1.50   0.8728088  0.7456157
1.75   0.8724224  0.7448422
2.00   0.8730038  0.7460050
5.00   0.8749384  0.7498738

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was C = 0.01.
```

k-NN (packages: kknn and caret)

```
k-Nearest Neighbors

1722 samples
  56 predictor
   2 classes: '1', '2'

Pre-processing: scaled (73)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 1550, 1549, 1550, 1549, 1550, 1550, ...
Resampling results across tuning parameters:

  kmax  Accuracy   Kappa
  5     0.9245013  0.8489996
  7     0.9233452  0.8466843
  9     0.9233452  0.8466843

Tuning parameter 'distance' was held constant at a value of 2
Tuning parameter 'kernel' was held constant at a value of optimal
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were kmax = 5, distance = 2 and kernel = optimal.
```

Naïve Bayes (packages: klaR and caret)

```
Naive Bayes

1722 samples
  56 predictor
   2 classes: '1', '2'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1550, 1549, 1550, 1550, 1550, 1550, ...
Resampling results across tuning parameters:

  usekernel  fL  adjust  Accuracy   Kappa
  FALSE      0   0             NaN        NaN
  FALSE      0   1             NaN        NaN
  FALSE      0   2             NaN        NaN
  FALSE      0   3             NaN        NaN
  FALSE      0   4             NaN        NaN
  FALSE      0   5             NaN        NaN
  FALSE      1   0             NaN        NaN
  FALSE      1   1             NaN        NaN
  FALSE      1   2             NaN        NaN
  FALSE      1   3             NaN        NaN
  FALSE      1   4             NaN        NaN
  FALSE      1   5             NaN        NaN
  FALSE      2   0             NaN        NaN
  FALSE      2   1             NaN        NaN
  FALSE      2   2             NaN        NaN
  FALSE      2   3             NaN        NaN
  FALSE      2   4             NaN        NaN
  FALSE      2   5             NaN        NaN
  FALSE      3   0             NaN        NaN
  FALSE      3   1             NaN        NaN
  FALSE      3   2             NaN        NaN
  FALSE      3   3             NaN        NaN
  FALSE      3   4             NaN        NaN
  FALSE      3   5             NaN        NaN
  FALSE      4   0             NaN        NaN
  FALSE      4   1             NaN        NaN
  FALSE      4   2             NaN        NaN
  FALSE      4   3             NaN        NaN
  FALSE      4   4             NaN        NaN
  FALSE      4   5             NaN        NaN
  FALSE      5   0             NaN        NaN
  FALSE      5   1             NaN        NaN
  FALSE      5   2             NaN        NaN
  FALSE      5   3             NaN        NaN
  FALSE      5   4             NaN        NaN
  FALSE      5   5             NaN        NaN
   TRUE      0   0             NaN        NaN
   TRUE      0   1       0.8403045  0.6805749
   TRUE      0   2       0.8432249  0.6864026
   TRUE      0   3       0.8449556  0.6898788
   TRUE      0   4       0.8490052  0.6979940
   TRUE      0   5       0.8327329  0.6654477
   TRUE      1   0             NaN        NaN
   TRUE      1   1       0.8403045  0.6805749
   TRUE      1   2       0.8432249  0.6864026
   TRUE      1   3       0.8449556  0.6898788
   TRUE      1   4       0.8490052  0.6979940
   TRUE      1   5       0.8327329  0.6654477
   TRUE      2   0             NaN        NaN
   TRUE      2   1       0.8403045  0.6805749
   TRUE      2   2       0.8432249  0.6864026
   TRUE      2   3       0.8449556  0.6898788
   TRUE      2   4       0.8490052  0.6979940
   TRUE      2   5       0.8327329  0.6654477
   TRUE      3   0             NaN        NaN
   TRUE      3   1       0.8403045  0.6805749
   TRUE      3   2       0.8432249  0.6864026
   TRUE      3   3       0.8449556  0.6898788
   TRUE      3   4       0.8490052  0.6979940
   TRUE      3   5       0.8327329  0.6654477
   TRUE      4   0             NaN        NaN
   TRUE      4   1       0.8403045  0.6805749
   TRUE      4   2       0.8432249  0.6864026
   TRUE      4   3       0.8449556  0.6898788
   TRUE      4   4       0.8490052  0.6979940
   TRUE      4   5       0.8327329  0.6654477
   TRUE      5   0             NaN        NaN
   TRUE      5   1       0.8403045  0.6805749
   TRUE      5   2       0.8432249  0.6864026
   TRUE      5   3       0.8449556  0.6898788
   TRUE      5   4       0.8490052  0.6979940
   TRUE      5   5       0.8327329  0.6654477

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were fL = 0, usekernel = TRUE and adjust = 4.
```

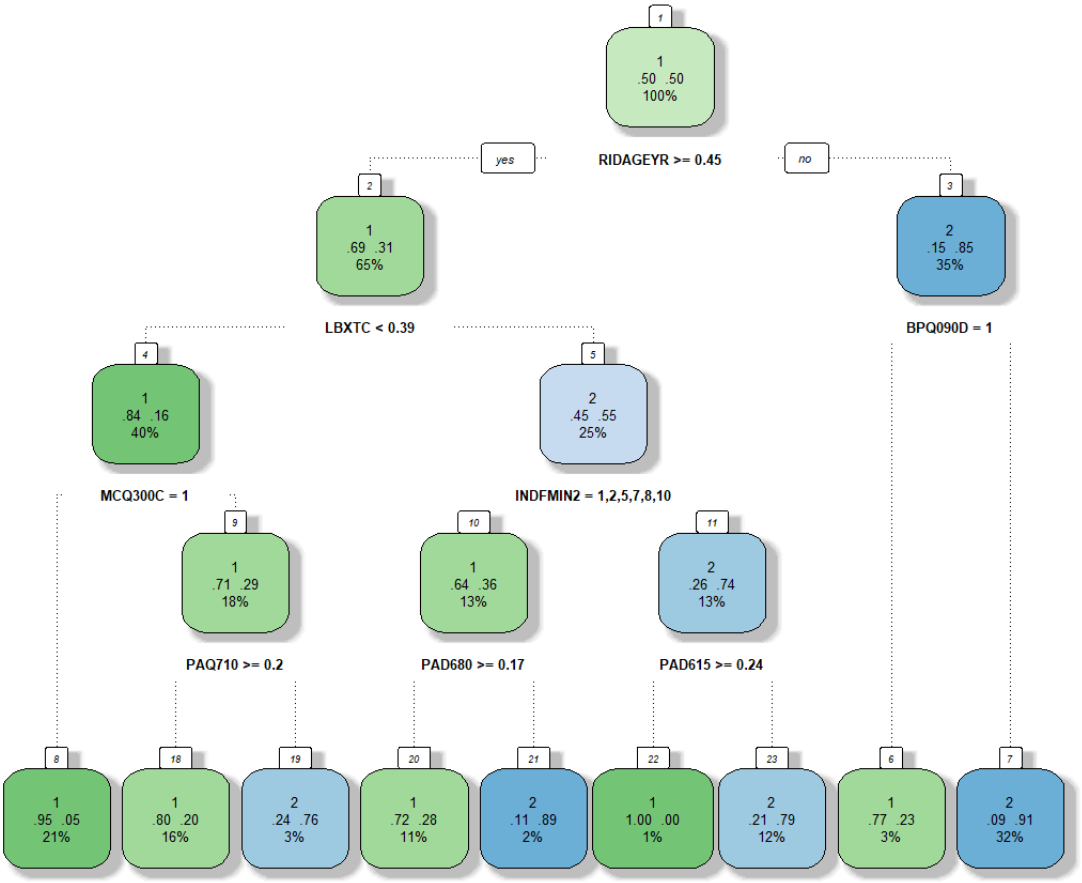
Appendix E – Stepwise LogR

Results of Stepwise (backwards) Logistic Regression

```
summary(diabetesA1_glm2)

Call:
glm(formula = DIQ010A ~ ALQ130 + PAD630 + PAD645 + PAD675 + PAD680 +
    PAQ710 + WHQ150 + LBXTC + LBDHDD + LBXTR + BMXWT + BMXHT +
    BMXLEG + BMXARMC + BMXWAIST + BMDAVSAD + RIDAGEYR + BPXSYave +
    BPQ020 + BPQ090D + MCQ092 + MCQ160B + MCQ160F + MCQ160K +
    MCQ160M + MCQ160N + MCQ203 + MCQ300C + MCQ365A + MCQ365B +
    MCQ365C + MCQ365D + MCQ370B + MCQ370D + SMQ020 + BPXPULS +
    DMDEDUC2 + INDFMIN2 + SMQ925A, family = binomial, data = balancedtrain_A1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8863  -0.3615   0.0001   0.2945   3.5566

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -22.7518   559.4103  -0.041 0.967558
ALQ130        2.2950     1.0936   2.099 0.035853 *
PAD630       -3.8592     1.1293  -3.417 0.000632 ***
PAD645       11.8188     5.9393   1.990 0.046598 *
PAD675        4.3573     1.1852   3.676 0.000237 ***
PAD680        1.5593     0.6001   2.599 0.009362 **
PAQ710       -1.4026     0.3417  -4.105 4.04e-05 ***
WHQ150        2.8527     0.5977   4.773 1.82e-06 ***
LBXTC         4.3494     0.6934   6.272 3.56e-10 ***
LBDHDD        2.3741     0.9894   2.400 0.016417 *
LBXTR        -4.2884     1.2193  -3.517 0.000436 ***
BMXWT        -3.6125     1.6507  -2.188 0.028636 *
BMXHT        -2.2072     1.1875  -1.859 0.063082 .
BMXLEG        2.0446     1.0573   1.934 0.053129 .
BMXARMC       3.3052     1.3539   2.441 0.014634 *
BMXWAIST     -6.4817     1.8684  -3.469 0.000522 ***
BMDAVSAD      5.2514     1.5677   3.350 0.000809 ***
RIDAGEYR     -5.8556     0.6439  -9.095  < 2e-16 ***
BPXSYave     -1.6425     0.6675  -2.461 0.013874 *
BPQ020        0.3060     0.2024   1.512 0.130507
BPQ090D2      1.5290     0.2067   7.398 1.38e-13 ***
MCQ092       -1.1020     0.3667  -3.005 0.002653 **
MCQ160B2      2.3989     0.7705   3.113 0.001849 **
MCQ160F2     15.8516   559.4079   0.028 0.977394
MCQ160K2     -1.2579     0.8043  -1.564 0.117832
MCQ160M2      0.8493     0.2832   2.999 0.002712 **
MCQ160N2      0.6593     0.3818   1.727 0.084243 .
MCQ2032       0.8588     0.4004   2.145 0.031976 *
MCQ300C2      2.1113     0.2038  10.362  < 2e-16 ***
MCQ365A2      0.9811     0.2338   4.196 2.72e-05 ***
MCQ365B2      0.6402     0.2035   3.146 0.001655 **
MCQ365C2     -0.6316     0.2188  -2.887 0.003888 **
MCQ365D2      0.6728     0.2060   3.267 0.001089 **
MCQ370B2     -0.3647     0.1932  -1.887 0.059145 .
MCQ370D2     -0.3797     0.1887  -2.013 0.044146 *
SMQ020        0.4797     0.2043   2.348 0.018880 *
BPXPULS2      3.5703     0.7957   4.487 7.22e-06 ***
DMDEDUC22     0.3059     0.4196   0.729 0.466025
DMDEDUC23     0.6293     0.3387   1.858 0.063183 .
DMDEDUC24     1.4318     0.3491   4.101 4.11e-05 ***
DMDEDUC25     0.9795     0.3196   3.064 0.002182 **
INDFMIN22     0.1967     0.7144   0.275 0.783042
INDFMIN23     2.1461     0.6483   3.310 0.000933 ***
INDFMIN24     1.4866     0.6498   2.288 0.022145 *
INDFMIN25     0.7961     0.6101   1.305 0.191954
INDFMIN26     1.4355     0.6430   2.233 0.025573 *
INDFMIN27     2.0019     0.5702   3.511 0.000447 ***
INDFMIN28     1.5048     0.5679   2.650 0.008052 **
INDFMIN29     2.4588     0.6295   3.906 9.39e-05 ***
INDFMIN210    1.0287     0.5487   1.875 0.060824 .
INDFMIN212    2.1072     0.7825   2.693 0.007087 **
INDFMIN213    3.0138     1.2487   2.414 0.015794 *
INDFMIN214    2.8492     0.6083   4.684 2.82e-06 ***
INDFMIN215    2.5950     0.5327   4.871 1.11e-06 ***
SMQ925A2     -0.3305     0.2110  -1.567 0.117217
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2387.20  on 1721  degrees of freedom
Residual deviance:  913.02  on 1667  degrees of freedom
AIC: 1023
```

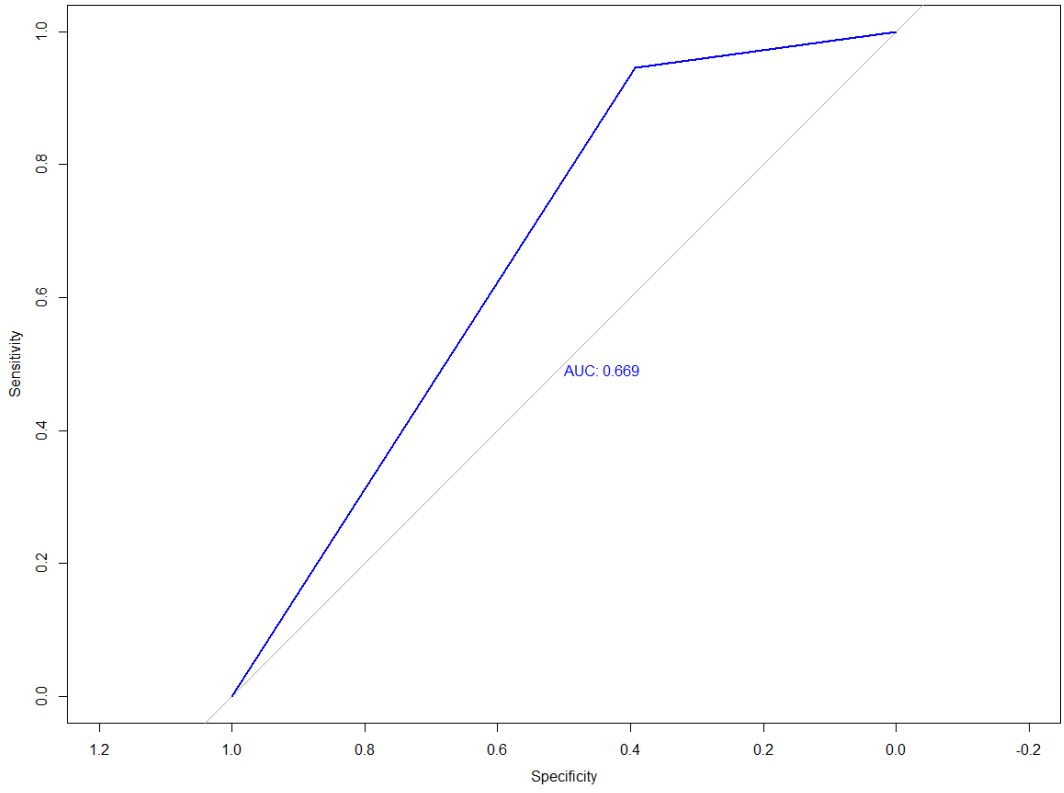
Appendix F – Classification trees

Decision tree

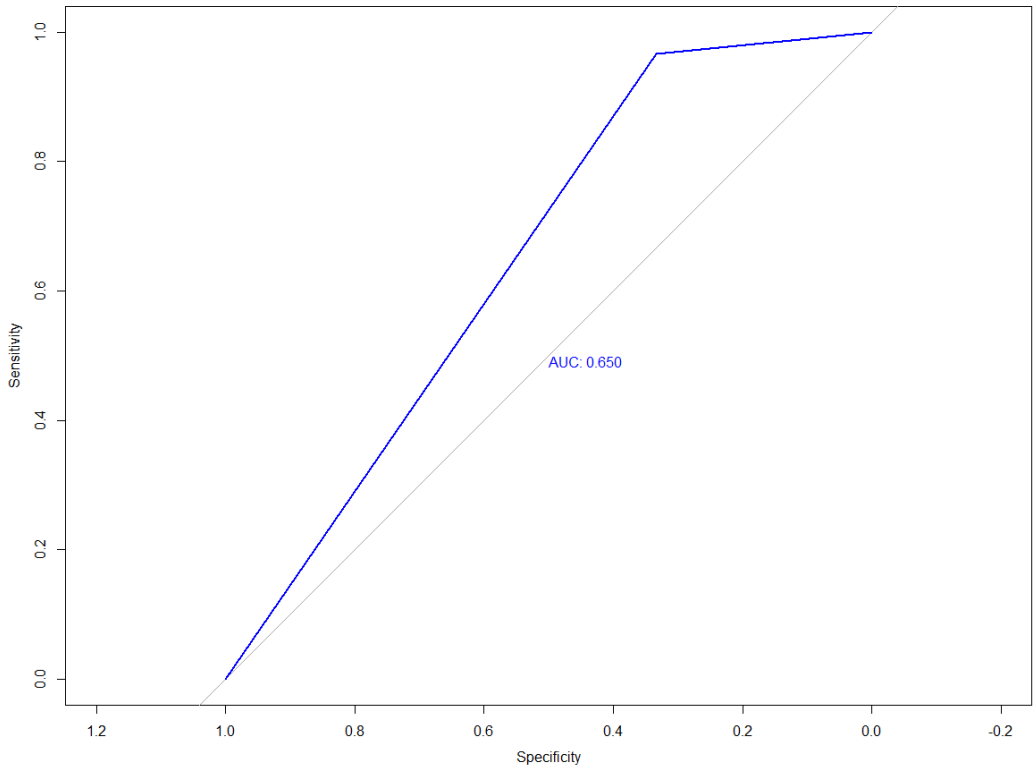


```
Variable importance
RIDAGEYR     LBXTC   BPQ090D   MCQ300C INDFMIN2    PAQ710    PAD615    PAD680
      45        17        11         7         6         5         4         4

Node number 1: 1722 observations,    complexity param=0.4901278
  predicted class=1  expected loss=0.5  P(node) =1
    class counts:   861   861
   probabilities: 0.500 0.500
  left son=2 (1116 obs) right son=3 (606 obs)
  Primary splits:
      RIDAGEYR < 0.450367     to the right, improve=243.7913, (0 missing)
      BPQ090D  splits as  LR, improve=204.9811, (0 missing)
      PAQ710   < 0.2005409    to the right, improve=128.6756, (0 missing)
      BPXSYave < 0.292231     to the right, improve=120.7481, (0 missing)
      WHQ150   < 0.2541836    to the right, improve=116.5843, (0 missing)

Node number 2: 1116 observations,    complexity param=0.05981417
  predicted class=1  expected loss=0.3109319  P(node) =0.6480836
    class counts:   769   347
   probabilities: 0.689 0.311
  left son=4 (683 obs) right son=5 (433 obs)
  Primary splits:
      LBXTC   < 0.3879237    to the left,  improve=92.07186, (0 missing)
      PAQ710  < 0.0004721307 to the right, improve=58.82599, (0 missing)
      MCQ300C splits as  LR, improve=48.14206, (0 missing)
      LBDLDL  < 0.331437     to the left,  improve=48.00254, (0 missing)
      BPQ090D splits as  LR, improve=47.84210, (0 missing)

Node number 3: 606 observations,    complexity param=0.03368177
  predicted class=2  expected loss=0.1518152  P(node) =0.3519164
    class counts:    92   514
   probabilities: 0.152 0.848
  left son=6 (53 obs) right son=7 (553 obs)
  Primary splits:
      BPQ090D  splits as  LR, improve=59.57922, (0 missing)
      MCQ365D  splits as  LR, improve=56.78055, (0 missing)
      MCQ365B  splits as  LR, improve=54.93585, (0 missing)
      BMDAVSAD < 0.3125       to the right, improve=49.97909, (0 missing)
      BMXWT    < 0.3467139    to the right, improve=43.82135, (0 missing)

Node number 4: 683 observations,    complexity param=0.014518
  predicted class=1  expected loss=0.1610542  P(node) =0.3966318
    class counts:   573   110
   probabilities: 0.839 0.161
  left son=8 (365 obs) right son=9 (318 obs)
  Primary splits:
      MCQ300C  splits as  LR, improve=38.47861, (0 missing)
      INDFMIN2 splits as  LLLLLRRLRLRLLR, improve=25.35526, (0 missing)
      PAQ710   < 0.2010917    to the right, improve=24.79317, (0 missing)
      WHQ150   < 0.1807657    to the right, improve=20.41375, (0 missing)
      ALQ120QU < 2.821057e-05 to the right, improve=20.41375, (0 missing)

Node number 5: 433 observations,    complexity param=0.05981417
  predicted class=2  expected loss=0.4526559  P(node) =0.2514518
    class counts:   196   237
   probabilities: 0.453 0.547
  left son=10 (216 obs) right son=11 (217 obs)
  Primary splits:
      INDFMIN2 splits as  LLRRLRLLRLRRRR, improve=32.53747, (0 missing)
      PAD680   < 0.1695843    to the right, improve=27.75970, (0 missing)
      BMDAVSAD < 0.3484649    to the right, improve=25.66385, (0 missing)
      BMXWAIST < 0.3565438    to the right, improve=21.30477, (0 missing)
      PAQ710   < 0.0006690818 to the right, improve=21.29330, (0 missing)

Node number 6: 53 observations
  predicted class=1  expected loss=0.2264151  P(node) =0.03077816
    class counts:    41    12
   probabilities: 0.774 0.226

Node number 7: 553 observations
  predicted class=2  expected loss=0.09222423  P(node) =0.3211382
    class counts:    51   502
   probabilities: 0.092 0.908

Node number 8: 365 observations
  predicted class=1  expected loss=0.04931507  P(node) =0.2119628
    class counts:   347    18
   probabilities: 0.951 0.049

Node number 9: 318 observations,    complexity param=0.014518
  predicted class=1  expected loss=0.2893082  P(node) =0.184669
    class counts:   226    92
   probabilities: 0.711 0.289
  left son=18 (269 obs) right son=19 (49 obs)
  Primary splits:
      PAQ710   < 0.2010917    to the right, improve=27.75536, (0 missing)
      INDFMIN2 splits as  LRRRLRRRRLRLLR, improve=20.37219, (0 missing)
      ALQ120QU < 2.821057e-05 to the right, improve=13.51697, (0 missing)
      BPXPULS  splits as  LR, improve=12.80804, (0 missing)
      RIDAGEYR < 0.7840737    to the right, improve=12.28924, (0 missing)

Node number 10: 216 observations,    complexity param=0.02555168
  predicted class=1  expected loss=0.3564815  P(node) =0.1254355
    class counts:   139    77
   probabilities: 0.644 0.356
  left son=20 (188 obs) right son=21 (28 obs)
  Primary splits:
      PAD680   < 0.1695843    to the right, improve=20.295900, (0 missing)
      BPXSYave < 0.3624874    to the right, improve=12.013470, (0 missing)
      BMDAVSAD < 0.3484649    to the right, improve=10.679870, (0 missing)
      BMXWAIST < 0.3127883    to the right, improve= 9.537253, (0 missing)
      BMXWT    < 0.227859     to the right, improve= 9.511664, (0 missing)

Node number 11: 217 observations,    complexity param=0.0174216
  predicted class=2  expected loss=0.2626728  P(node) =0.1260163
    class counts:    57   160
   probabilities: 0.263 0.737
  left son=22 (15 obs) right son=23 (202 obs)
  Primary splits:
      PAD615 < 0.2428822    to the right, improve=21.69585, (0 missing)
      PAD630 < 0.2074473    to the right, improve=20.43212, (0 missing)
      PAQ710 < 0.006924393  to the right, improve=12.00072, (0 missing)
      LBXTC  < 0.5926724    to the left,  improve=11.45393, (0 missing)
      LBDHDD < 0.2051945    to the left,  improve=10.19699, (0 missing)
```

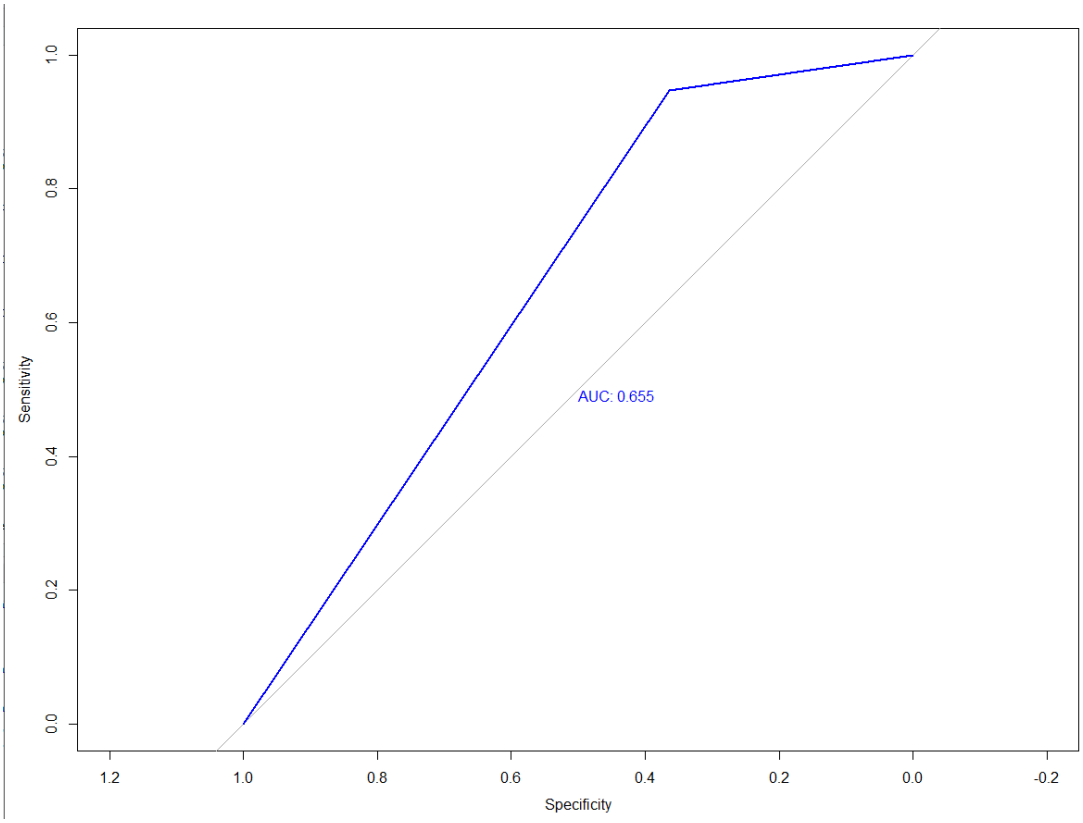
```
Node number 18: 269 observations
  predicted class=1  expected loss=0.204461  P(node) =0.1562137
    class counts:   214    55
   probabilities: 0.796 0.204

Node number 19: 49 observations
  predicted class=2  expected loss=0.244898  P(node) =0.02845528
    class counts:    12    37
   probabilities: 0.245 0.755

Node number 20: 188 observations
  predicted class=1  expected loss=0.2765957  P(node) =0.1091754
    class counts:   136    52
   probabilities: 0.723 0.277

Node number 21: 28 observations
  predicted class=2  expected loss=0.1071429  P(node) =0.01626016
    class counts:     3    25
   probabilities: 0.107 0.893

Node number 22: 15 observations
  predicted class=1  expected loss=0  P(node) =0.008710801
    class counts:    15     0
   probabilities: 1.000 0.000

Node number 23: 202 observations
  predicted class=2  expected loss=0.2079208  P(node) =0.1173055
    class counts:    42   160
   probabilities: 0.208 0.792
```



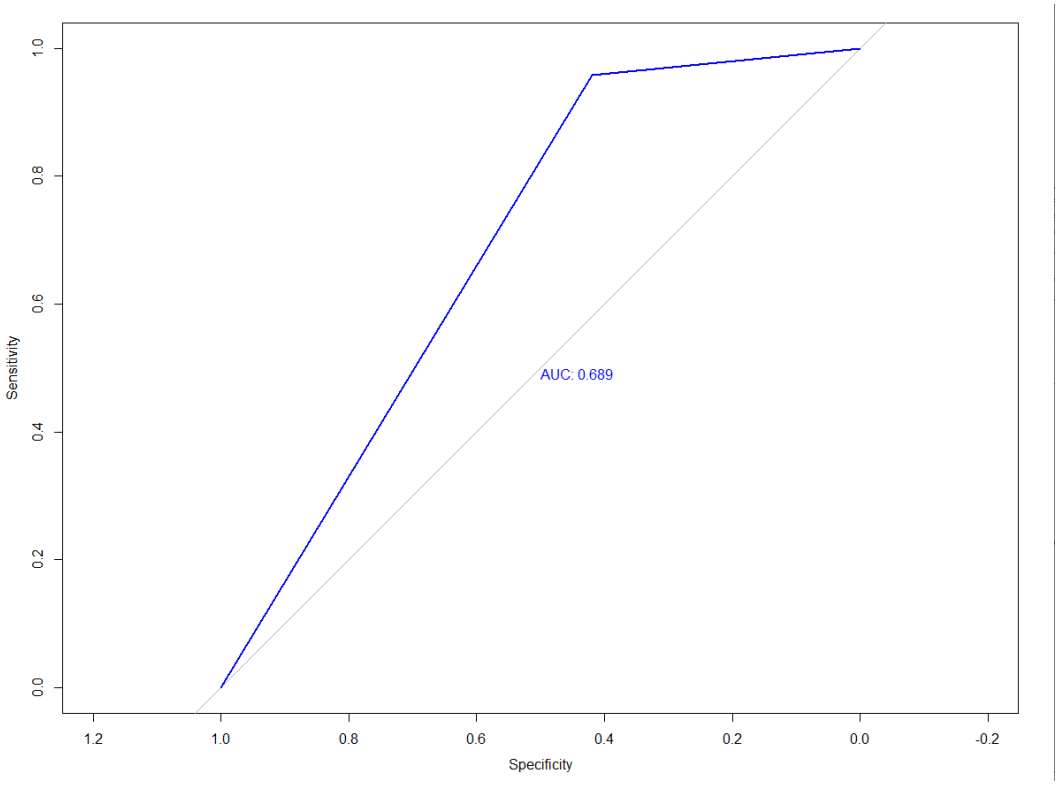
XGB trees



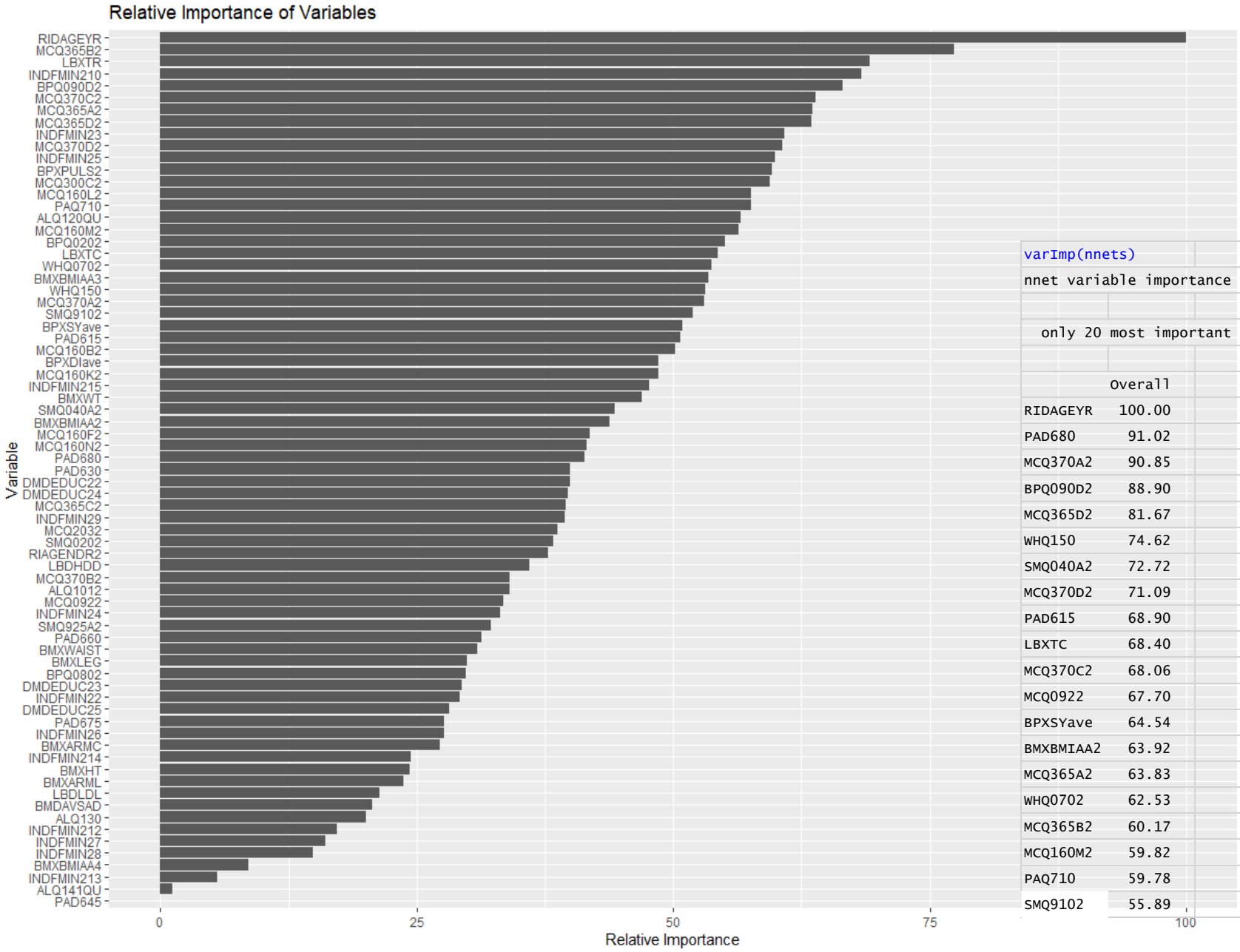
GBM trees



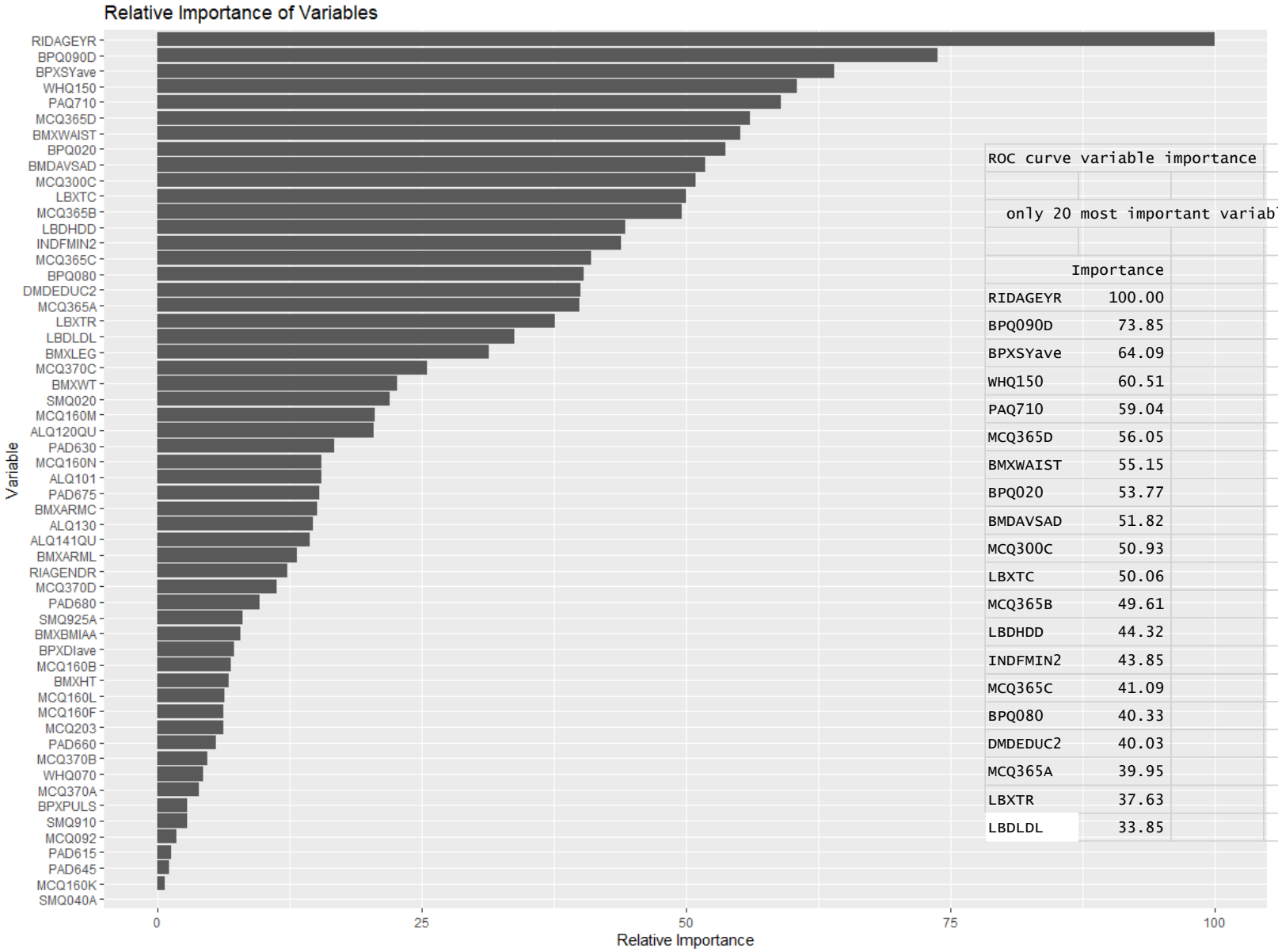
Adaboost trees



Appendix G – varImp results of NN and k-NN




```
varImp(nnets)
nnet variable importance

  only 20 most important variables shown (out of 73)

          Overall
RIDAGEYR  100.00
PAD680     91.02
MCQ370A2   90.85
BPQ090D2   88.90
MCQ365D2   81.67
WHQ150     74.62
SMQ040A2   72.72
MCQ370D2   71.09
PAD615     68.90
LBXTC      68.40
MCQ370C2   68.06
MCQ0922    67.70
BPXSYave   64.54
BMXBMIAA2  63.92
MCQ365A2   63.83
WHQ0702    62.53
MCQ365B2   60.17
MCQ160M2   59.82
PAQ710     59.78
SMQ9102    55.89
```



ROC curve variable importance

only 20 most important variables shown (out of 56)

	Importance
RIDAGEYR	100.00
BPQ090D	73.85
BPXSYave	64.09
WHQ150	60.51
PAQ710	59.04
MCQ365D	56.05
BMXWAIST	55.15
BPQ020	53.77
BMDAVSAD	51.82
MCQ300C	50.93
LBXTC	50.06
MCQ365B	49.61
LBDHDD	44.32
INDFMIN2	43.85
MCQ365C	41.09
BPQ080	40.33
DMDEDUC2	40.03
MCQ365A	39.95
LBXTR	37.63
LBDLDL	33.85

Appendix H – Multiclass models trained using imbalanced dataset

(some of the models)

DT

Confusion Matrix and Statistics

Prediction \ Reference	1	2	3	4
1	17	0	3	9
2	0	0	0	0
3	2	3	9	6
4	26	4	44	279

Overall Statistics

Accuracy	0.7587
95% CI	(0.7138, 0.7997)
No Information Rate	0.7313
P-Value [Acc > NIR]	0.118
Kappa	0.2961
Mcnemar's Test P-Value	NA

Statistics by Class:

	Class: 1	Class: 2	Class: 3	Class: 4
Precision	0.58621	NA	0.45000	0.7904
Recall	0.37778	0.00000	0.16071	0.9490
F1	0.45946	NA	0.23684	0.8624
Prevalence	0.11194	0.01741	0.13930	0.7313
Detection Rate	0.04229	0.00000	0.02239	0.6940
Detection Prevalence	0.07214	0.00000	0.04975	0.8781
Balanced Accuracy	0.67208	0.50000	0.56446	0.6319

RF

Confusion Matrix and Statistics

Prediction \ Reference	1	2	3	4
1	9	0	2	0
2	0	0	0	0
3	0	0	0	0
4	36	7	54	294

Overall Statistics

Accuracy	0.7537
95% CI	(0.7086, 0.7951)
No Information Rate	0.7313
P-Value [Acc > NIR]	0.1696
Kappa	0.1377
Mcnemar's Test P-Value	NA

Statistics by Class:

	Class: 1	Class: 2	Class: 3	Class: 4
Precision	0.81818	NA	NA	0.7519
Recall	0.20000	0.00000	0.0000	1.0000
F1	0.32143	NA	NA	0.8584
Prevalence	0.11194	0.01741	0.1393	0.7313
Detection Rate	0.02239	0.00000	0.0000	0.7313
Detection Prevalence	0.02736	0.00000	0.0000	0.9726
Balanced Accuracy	0.59720	0.50000	0.5000	0.5509

NN

Confusion Matrix and Statistics

Prediction \ Reference	1	2	3	4
1	14	3	4	13
2	0	0	0	0
3	0	0	0	0
4	31	4	52	281

Overall Statistics

Accuracy	0.7338
95% CI	(0.6878, 0.7764)
No Information Rate	0.7313
P-Value [Acc > NIR]	0.481
Kappa	0.1709
Mcnemar's Test P-Value	NA

Statistics by Class:

	Class: 1	Class: 2	Class: 3	Class: 4
Precision	0.41176	NA	NA	0.7636
Recall	0.31111	0.00000	0.0000	0.9558
F1	0.35443	NA	NA	0.8489
Prevalence	0.11194	0.01741	0.1393	0.7313
Detection Rate	0.03483	0.00000	0.0000	0.6990
Detection Prevalence	0.08458	0.00000	0.0000	0.9154
Balanced Accuracy	0.62754	0.50000	0.5000	0.5751

NB

Confusion Matrix and Statistics

Prediction \ Reference	1	2	3	4
1	0	0	0	0
2	0	0	0	0
3	0	0	0	3
4	45	7	56	291

Overall Statistics

Accuracy	0.7239
95% CI	(0.6774, 0.767)
No Information Rate	0.7313
P-Value [Acc > NIR]	0.6558
Kappa	-0.0111
Mcnemar's Test P-Value	NA

Statistics by Class:

	Class: 1	Class: 2	Class: 3	Class: 4
Precision	NA	NA	0.000000	0.7293
Recall	0.0000	0.00000	0.000000	0.9898
F1	NA	NA	NaN	0.8398
Prevalence	0.1119	0.01741	0.139303	0.7313
Detection Rate	0.0000	0.00000	0.000000	0.7239
Detection Prevalence	0.0000	0.00000	0.007463	0.9925
Balanced Accuracy	0.5000	0.50000	0.495665	0.4949

Appendix I – Hyperparameters tuning and results

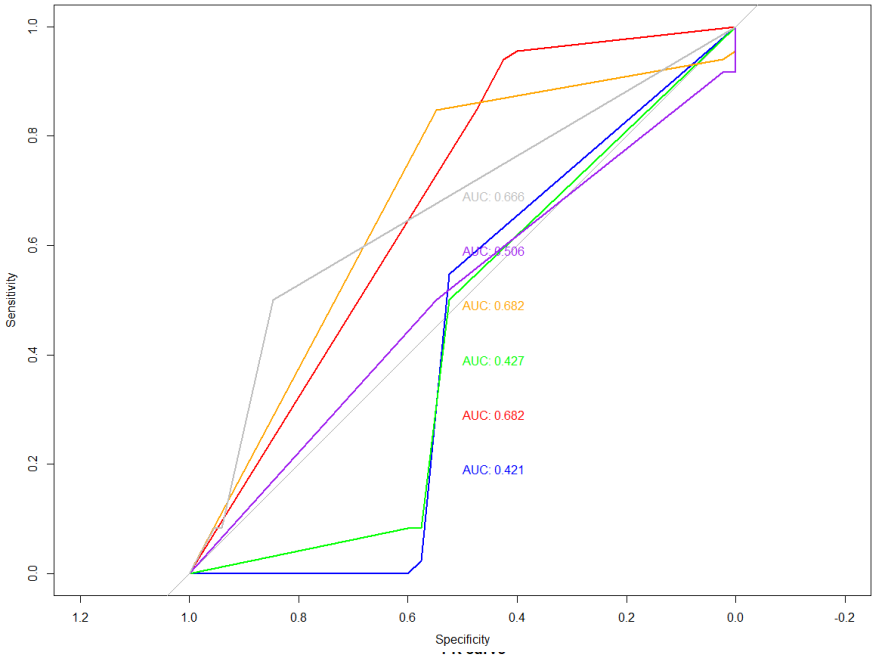
DT

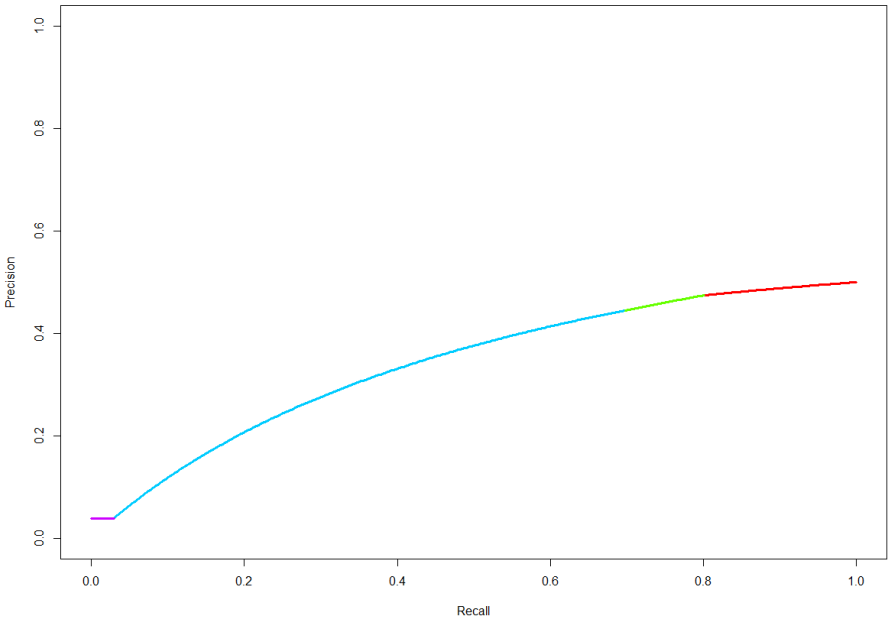
```
Call:
rpart(formula = DIQ010B ~ ., data = balancedtrain_B1, method = "class",
    model = TRUE, parms = list(split = "information"), control = rpart.control(minsplit = 20,
        minbucket = 7, usesurrogate = 0, maxsurrogate = 0))
  n= 939

          CP nsplit rel error    xerror       xstd
1  0.23295455      0 1.0000000 1.0568182 0.01765624
2  0.08380682      1 0.7670455 0.7883523 0.02139965
3  0.07812500      2 0.6832386 0.6860795 0.02175457
4  0.03977273      4 0.5269886 0.5639205 0.02150252
5  0.03409091      5 0.4872159 0.5411932 0.02137343
6  0.02698864      6 0.4531250 0.5227273 0.02124890
7  0.02556818      7 0.4261364 0.5085227 0.02114093
8  0.01420455      8 0.4005682 0.4772727 0.02086521
9  0.01207386      9 0.3863636 0.4232955 0.02025964
10 0.01136364     11 0.3622159 0.4176136 0.02018590
11 0.01000000     13 0.3394886 0.4133523 0.02012929

Variable importance
   LBXTR   LBDLDL  BPQ090D BPXSYave    LBXTC BPXDIave   PAD680   PAD675  SMQ925A ALQ120QU RIDAGEYR
      24       23       10       10        7        6        5        5        4        4        3

Node number 1: 939 observations,    complexity param=0.2329545
  predicted class=1  expected loss=0.7497338  P(node) =1
    class counts:   235   234   235   235
   probabilities: 0.250 0.249 0.250 0.250
  left son=2 (412 obs) right son=3 (527 obs)
  Primary splits:
      LBXTR    < 0.1012422  to the right, improve=153.25340, (0 missing)
      LBDLDL   < 0.3652765  to the right, improve=150.05280, (0 missing)
      BPXSYave < 0.3248855  to the right, improve=135.88060, (0 missing)
      BPQ090D  splits as  LR, improve= 95.28437, (0 missing)
      RIDAGEYR < 0.4504734  to the right, improve= 86.16024, (0 missing)

Node number 2: 412 observations,    complexity param=0.08380682
  predicted class=2  expected loss=0.5339806  P(node) =0.4387646
    class counts:    60   192   132    28
   probabilities: 0.146 0.466 0.320 0.068
  left son=4 (285 obs) right son=5 (127 obs)
  Primary splits:
      BPXSYave < 0.3322235  to the right, improve=61.39871, (0 missing)
      LBXTC    < 0.2555522  to the left,  improve=45.61302, (0 missing)
      PAD675   < 0.1001977  to the right, improve=42.27651, (0 missing)
      INDFMIN2 splits as  RRRRLLLLLL-RRL, improve=36.51103, (0 missing)
      ALQ120QU < 0.1518916  to the right, improve=36.38813, (0 missing)

Node number 3: 527 observations,    complexity param=0.078125
  predicted class=4  expected loss=0.6072106  P(node) =0.5612354
    class counts:   175    42   103   207
   probabilities: 0.332 0.080 0.195 0.393
  left son=6 (426 obs) right son=7 (101 obs)
  Primary splits:
      LBDLDL   < 0.3472812  to the left,  improve=97.14584, (0 missing)
      LBXTR    < 0.0981294  to the left,  improve=96.63174, (0 missing)
      RIDAGEYR < 0.4504734  to the right, improve=74.72916, (0 missing)
      BPQ090D  splits as  LR, improve=73.99264, (0 missing)
      BPXSYave < 0.3131817  to the right, improve=52.46960, (0 missing)

Node number 4: 285 observations,    complexity param=0.02698864
  predicted class=2  expected loss=0.3649123  P(node) =0.3035144
    class counts:    32   181    62    10
   probabilities: 0.112 0.635 0.218 0.035
  left son=8 (27 obs) right son=9 (258 obs)
  Primary splits:
      LBXTC    < 0.2573388  to the left,  improve=42.88545, (0 missing)
      PAD675   < 0.1013333  to the left,  improve=41.79717, (0 missing)
      BMXLEG   < 0.25       to the left,  improve=34.79425, (0 missing)
      BPXDIave < 0.6220333  to the left,  improve=34.37668, (0 missing)
      INDFMIN2 splits as  RRRRLLLLLL-RRL, improve=29.96326, (0 missing)
```

```
Node number 5: 127 observations
  predicted class=3  expected loss=0.4488189  P(node) =0.1352503
    class counts:    28    11    70    18
   probabilities: 0.220 0.087 0.551 0.142

Node number 6: 426 observations,    complexity param=0.078125
  predicted class=4  expected loss=0.5586854  P(node) =0.4536741
    class counts:   169     1    68   188
   probabilities: 0.397 0.002 0.160 0.441
  left son=12 (150 obs) right son=13 (276 obs)
  Primary splits:
      BPQ090D  splits as  LR, improve=63.64100, (0 missing)
      LBDLDL   < 0.3277808  to the left,  improve=62.78975, (0 missing)
      RIDAGEYR < 0.6856679  to the right, improve=60.18801, (0 missing)
      LBXTR    < 0.09784008 to the left,  improve=50.00703, (0 missing)
      MCQ365D  splits as  LR, improve=43.24987, (0 missing)

Node number 7: 101 observations,    complexity param=0.03409091
  predicted class=2  expected loss=0.5940594  P(node) =0.1075612
    class counts:     6    41    35    19
   probabilities: 0.059 0.406 0.347 0.188
  left son=14 (61 obs) right son=15 (40 obs)
  Primary splits:
      BPXDIave < 0.4806226  to the right, improve=37.73456, (0 missing)
      BPXSYave < 0.3239816  to the right, improve=30.38431, (0 missing)
      RIDAGEYR < 0.4667408  to the right, improve=23.13376, (0 missing)
      PAD615   < 0.2203646  to the left,  improve=19.38726, (0 missing)
      MCQ370B  splits as  RL, improve=18.25352, (0 missing)

Node number 8: 27 observations
  predicted class=1  expected loss=0.2222222  P(node) =0.02875399
    class counts:    21     2     4     0
   probabilities: 0.778 0.074 0.148 0.000

Node number 9: 258 observations,    complexity param=0.02556818
  predicted class=2  expected loss=0.3062016  P(node) =0.2747604
    class counts:    11   179    58    10
   probabilities: 0.043 0.694 0.225 0.039
  left son=18 (236 obs) right son=19 (22 obs)
  Primary splits:
      PAD675   < 0.103065   to the left,  improve=29.05295, (0 missing)
      BPXDIave < 0.6280932  to the left,  improve=28.66979, (0 missing)
      PAD680   < 0.4434526  to the left,  improve=26.45480, (0 missing)
      INDFMIN2 splits as  RRRRLLLLLL-RRL, improve=23.93603, (0 missing)
      SMQ925A  splits as  LR, improve=20.48314, (0 missing)

Node number 12: 150 observations
  predicted class=1  expected loss=0.2466667  P(node) =0.1597444
    class counts:   113     0    12    25
   probabilities: 0.753 0.000 0.080 0.167

Node number 13: 276 observations,    complexity param=0.03977273
  predicted class=4  expected loss=0.4094203  P(node) =0.2939297
    class counts:    56     1    56   163
   probabilities: 0.203 0.004 0.203 0.591
  left son=26 (83 obs) right son=27 (193 obs)
  Primary splits:
      LBDLDL   < 0.3431497  to the left,  improve=49.99173, (0 missing)
      LBXTR    < 0.09751553 to the left,  improve=41.91309, (0 missing)
      RIDAGEYR < 0.65006    to the right, improve=34.55268, (0 missing)
      MCQ365D  splits as  LR, improve=19.51376, (0 missing)
      INDFMIN2 splits as  RLLRLRRRRRR-RR, improve=18.50263, (0 missing)

Node number 14: 61 observations
  predicted class=2  expected loss=0.3278689  P(node) =0.06496273
    class counts:     6    41    11     3
```

```
Node number 19: 22 observations
  predicted class=3  expected loss=0.1818182  P(node) =0.02342918
    class counts:     2     0    18     2
   probabilities: 0.091 0.000 0.818 0.091

Node number 26: 83 observations,    complexity param=0.01420455
  predicted class=3  expected loss=0.4457831  P(node) =0.08839191
    class counts:    19     0    46    18
   probabilities: 0.229 0.000 0.554 0.217
  left son=52 (16 obs) right son=53 (67 obs)
  Primary splits:
      RIDAGEYR < 0.8125668  to the right, improve=17.74844, (0 missing)
      BPXSYave < 0.3516961  to the right, improve=15.15554, (0 missing)
      WHQ150   < 0.5895522  to the right, improve=14.59365, (0 missing)
      LBXTR    < 0.0689441  to the right, improve=14.00509, (0 missing)
      BMXWAIST < 0.3254438  to the right, improve=13.16691, (0 missing)

Node number 27: 193 observations
  predicted class=4  expected loss=0.2487047  P(node) =0.2055378
    class counts:    37     1    10   145
   probabilities: 0.192 0.005 0.052 0.751

Node number 30: 30 observations
  predicted class=3  expected loss=0.2333333  P(node) =0.03194888
    class counts:     0     0    23     7
   probabilities: 0.000 0.000 0.767 0.233

Node number 31: 10 observations
  predicted class=4  expected loss=0.1  P(node) =0.01064963
    class counts:     0     0     1     9
   probabilities: 0.000 0.000 0.100 0.900

Node number 36: 193 observations,    complexity param=0.01136364
  predicted class=2  expected loss=0.1502591  P(node) =0.2055378
    class counts:     7   164    21     1
   probabilities: 0.036 0.850 0.109 0.005
  left son=72 (182 obs) right son=73 (11 obs)
  Primary splits:
      SMQ925A  splits as  LR, improve=24.050160, (0 missing)
      PAD645   < 0.0107899  to the right, improve=15.083430, (0 missing)
      BMXWAIST < 0.587342   to the right, improve=10.299340, (0 missing)
      BPXDIave < 0.5992129  to the left,  improve= 9.995180, (0 missing)
      INDFMIN2 splits as  LL-LLRRRRR---R, improve= 9.926434, (0 missing)

Node number 37: 43 observations,    complexity param=0.01207386
  predicted class=3  expected loss=0.5581395  P(node) =0.0457934
    class counts:     2    15    19     7
   probabilities: 0.047 0.349 0.442 0.163
  left son=74 (17 obs) right son=75 (26 obs)
  Primary splits:
      ALQ120QU < 0.3055154  to the right, improve=22.46305, (0 missing)
      BMXHT    < 0.3539519  to the right, improve=19.10355, (0 missing)
      BMXARML  < 0.5774367  to the left,  improve=19.07806, (0 missing)
      PAQ710   < 0.2981002  to the right, improve=14.67425, (0 missing)
      LBDHDD   < 0.1478765  to the left,  improve=13.22824, (0 missing)

Node number 52: 16 observations
  predicted class=1  expected loss=0.1875  P(node) =0.0170394
    class counts:    13     0     3     0
   probabilities: 0.812 0.000 0.188 0.000

Node number 53: 67 observations
```

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4
         1  32   2   4  42
         2   0   1  22  19
         3  12   4  25 227
         4   1   0   5   6

Overall Statistics

               Accuracy : 0.1592
                 95% CI : (0.1248, 0.1987)
    No Information Rate : 0.7313
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0237

 Mcnemar's Test P-Value : <2e-16

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Precision              0.4000 0.023810  0.09328  0.50000
Recall                 0.7111 0.142857  0.44643  0.02041
F1                     0.5120 0.040816  0.15432  0.03922
Prevalence             0.1119 0.017413  0.13930  0.73134
Detection Rate         0.0796 0.002488  0.06219  0.01493
Detection Prevalence   0.1990 0.104478  0.66667  0.02985
Balanced Accuracy      0.7883 0.519530  0.37206  0.48243
```

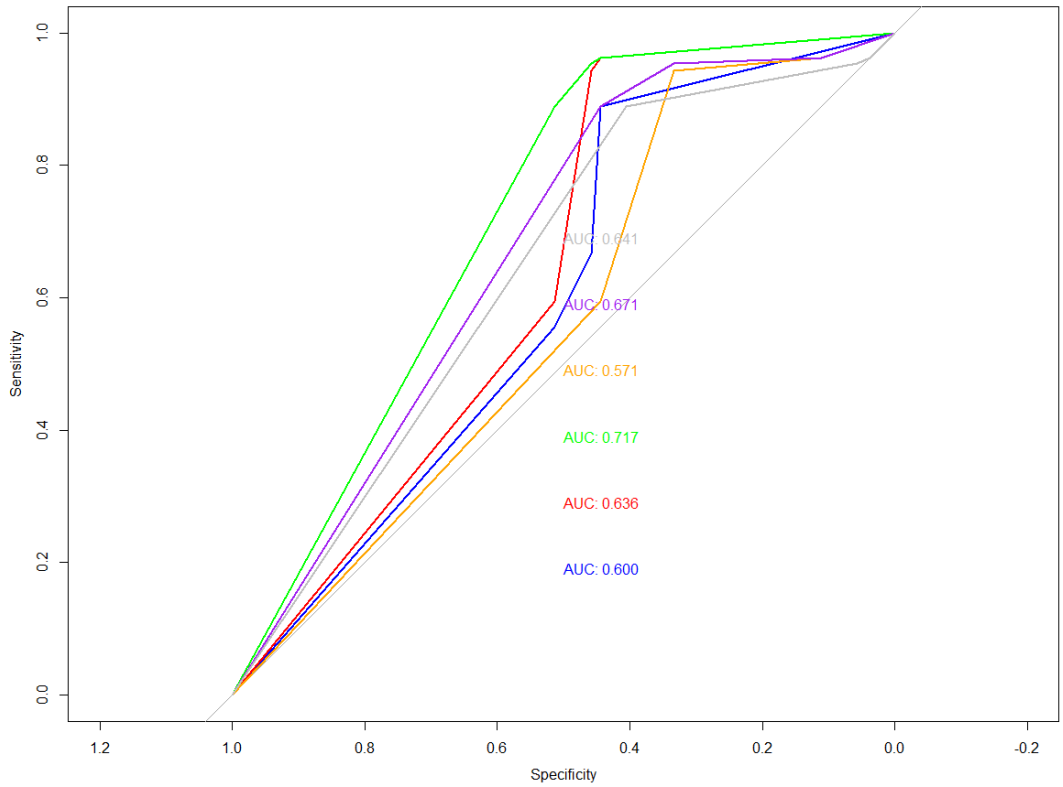


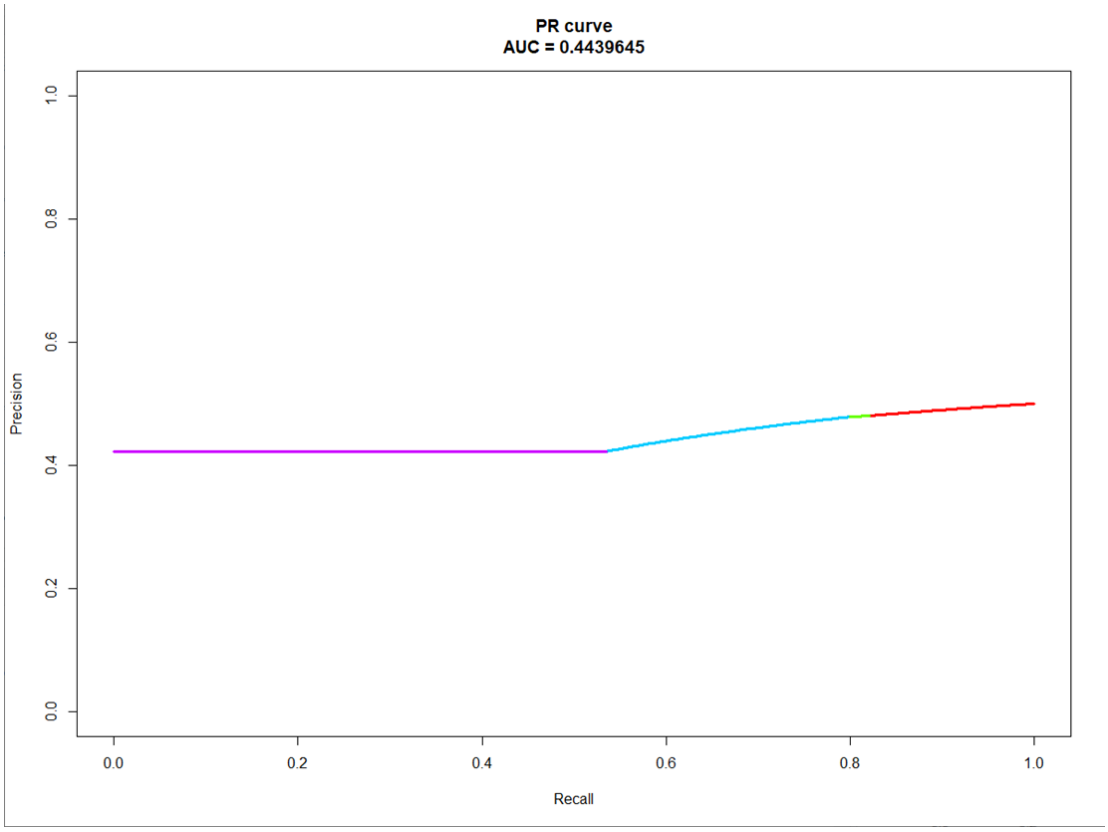


RF

```
Random Forest

939 samples
 56 predictor
  4 classes: '1', '2', '3', '4'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 751, 751, 751, 751, 752
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
  1     0.7220560  0.6294440
  2     0.7976903  0.7302576
  3     0.8253669  0.7671579
  4     0.8338890  0.7785204

Accuracy was used to select the optimal model using the
The final value used for the model was mtry = 4.
```

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4
         1  32   1   4  35
         2   1   2   1   5
         3   4   2  37  63
         4   8   2  14 191

Overall Statistics

               Accuracy : 0.6517
                 95% CI : (0.6029, 0.6983)
    No Information Rate : 0.7313
    P-Value [Acc > NIR] : 0.9998

                  Kappa : 0.3687

 Mcnemar's Test P-Value : 5.265e-09

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Precision              0.4444 0.222222  0.34906   0.8884
Recall                 0.7111 0.285714  0.66071   0.6497
F1                     0.5470 0.250000  0.45679   0.7505
Prevalence             0.1119 0.017413  0.13930   0.7313
Detection Rate         0.0796 0.004975  0.09204   0.4751
Detection Prevalence   0.1791 0.022388  0.26368   0.5348
Balanced Accuracy      0.7995 0.633996  0.73065   0.7137
```

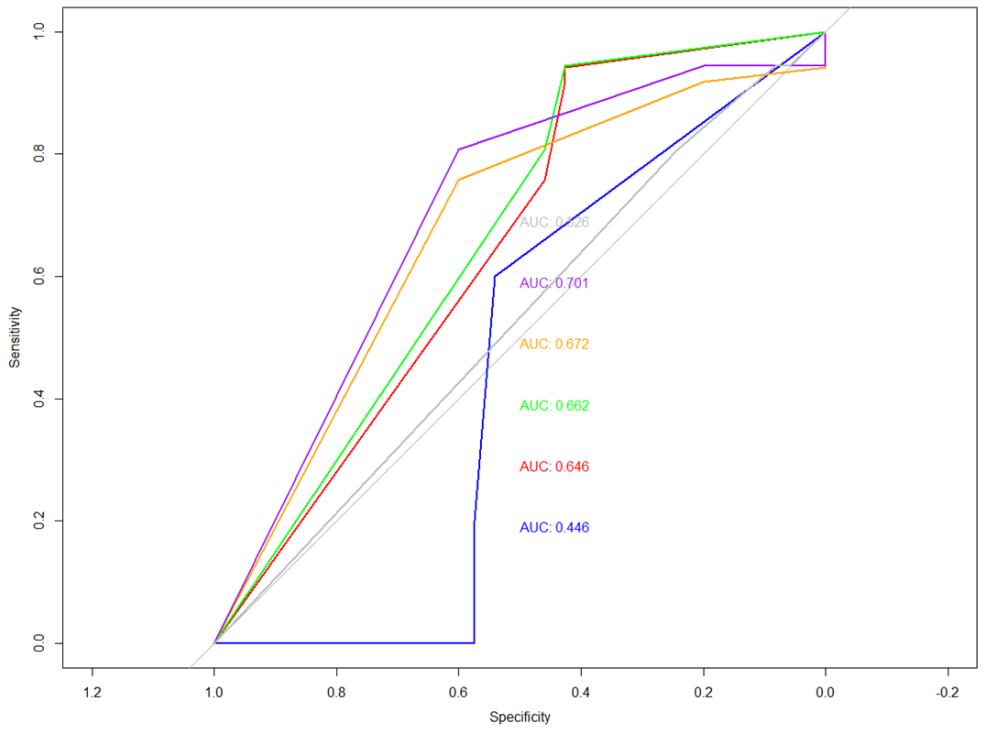

Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
AUC: 0.641
AUC: 0.671
AUC: 0.571
AUC: 0.717
AUC: 0.636
AUC: 0.600

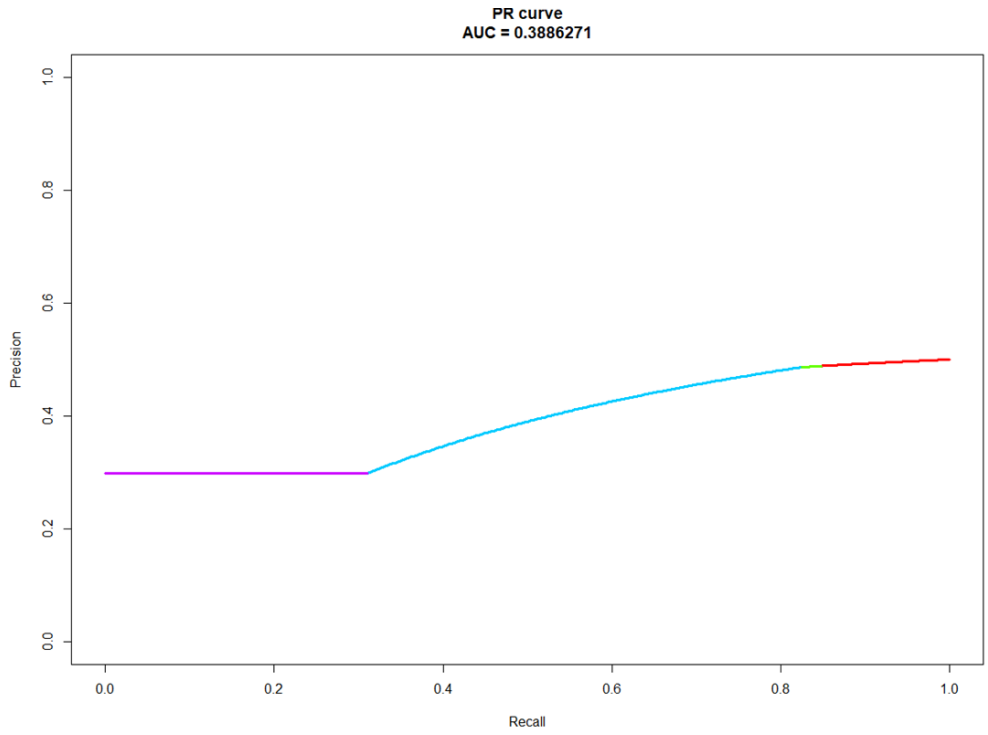

PR curve
AUC = 0.4439645
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

XGB trees

```
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 846, 845, 846, 845, 845, 846, ...
Resampling results across tuning parameters:
Tuning parameter 'gamma' was held constant at a value of 0
Tuning parameter 'min_child_weight' was held constant at a value of 1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were nrounds = 100, max_depth = 3, eta = 0.4, gamma = 0, colsample_bytree = 0.8, min_child_weight = 1 and subsample = 0.75.
```

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4
         1  26   0   2  33
         2   0   2   4   4
         3  12   5  33 156
         4   7   0  17 101

Overall Statistics

               Accuracy : 0.403
                 95% CI : (0.3547, 0.4527)
    No Information Rate : 0.7313
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1269

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Precision            0.42623 0.200000  0.16019   0.8080
Recall               0.57778 0.285714  0.58929   0.3435
F1                   0.49057 0.235294  0.25191   0.4821
Prevalence           0.11194 0.017413  0.13930   0.7313
Detection Rate       0.06468 0.004975  0.08209   0.2512
Detection Prevalence 0.15174 0.024876  0.51244   0.3109
Balanced Accuracy    0.73987 0.632731  0.54464   0.5607
```





GBM trees

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 846, 845, 846, 845, 845, 846, ...

Resampling results across tuning parameters:

interaction.depth	n.trees	Accuracy	Kappa
1	50	0.7327204	0.6436161
1	100	0.7614231	0.6818775
1	150	0.7891190	0.7188331
2	50	0.7753564	0.7004634
2	100	0.8072515	0.7429720
2	150	0.8093794	0.7458364
3	50	0.7933297	0.7244079
3	100	0.8050443	0.7400592
3	150	0.8167359	0.7556180

Tuning parameter 'shrinkage' was held constant at a value of 0.1

Tuning parameter 'n.minobsinnode' was held constant at a value of 10

Accuracy was used to select the optimal model using the largest value.

The final values used for the model were n.trees = 150, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10.

Confusion Matrix and Statistics

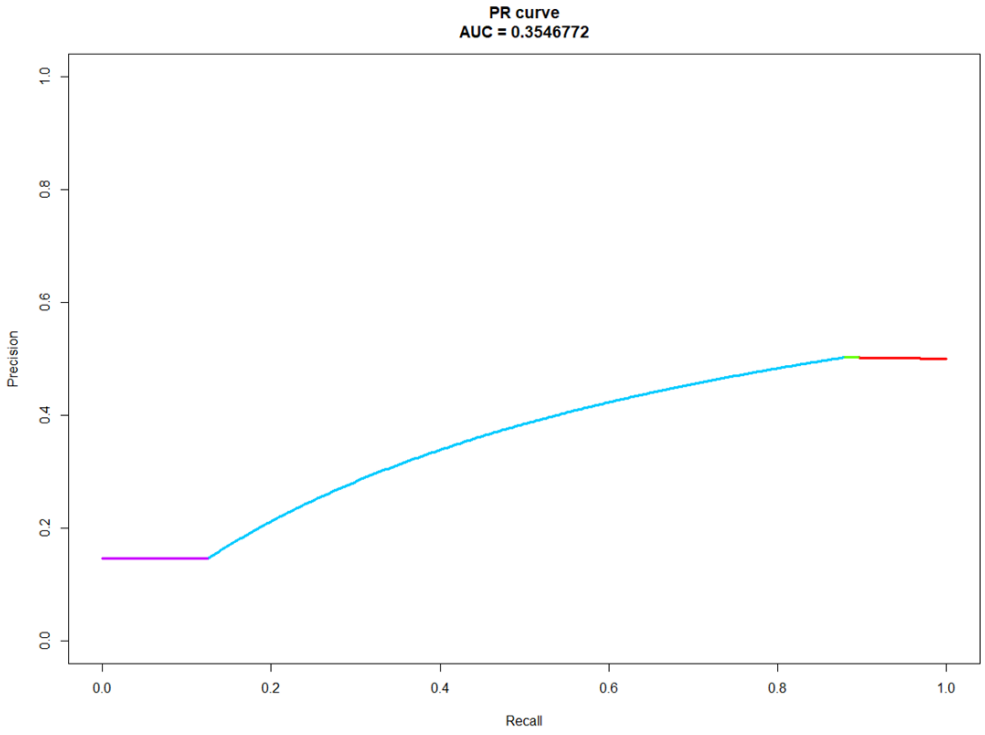
Prediction	Reference 1	2	3	4
1	17	0	4	21
2	1	2	2	2
3	27	5	45	226
4	0	0	5	45

Overall Statistics

Accuracy	0.2711
95% CI	(0.2283, 0.3174)
No Information Rate	0.7313
P-Value [Acc > NIR]	1
Kappa	0.0798
Mcnemar's Test P-Value	<2e-16

Statistics by Class:

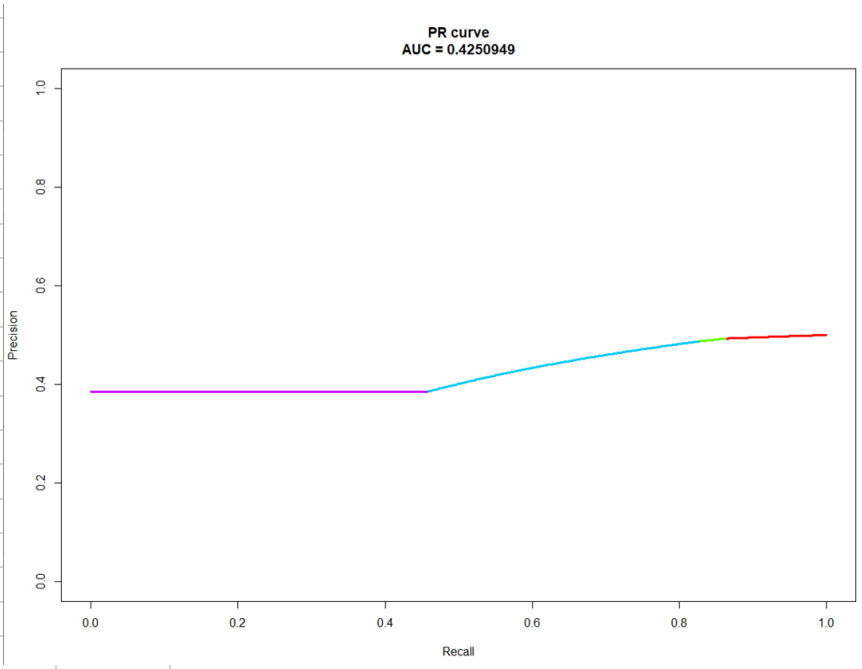
	Class: 1	Class: 2	Class: 3	Class: 4
Precision	0.40476	0.285714	0.1485	0.9000
Recall	0.37778	0.285714	0.8036	0.1531
F1	0.39080	0.285714	0.2507	0.2616
Prevalence	0.11194	0.017413	0.1393	0.7313
Detection Rate	0.04229	0.004975	0.1119	0.1119
Detection Prevalence	0.10448	0.017413	0.7537	0.1244
Balanced Accuracy	0.65387	0.636528	0.5290	0.5534



NN

```
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 939, 939, 939, 939, 939, 939, ...
Resampling results across tuning parameters:

  size  decay  Accuracy   Kappa
  1     0.000  0.4082440  0.2159860
  1     0.001  0.4086406  0.2164542
  1     0.100  0.5269095  0.3702181
  3     0.000  0.6144996  0.4861270
  3     0.001  0.5974596  0.4629852
  3     0.100  0.6499823  0.5328934
  5     0.000  0.6041312  0.4723176
  5     0.001  0.6077107  0.4767020
  5     0.100  0.6430317  0.5238790
  7     0.000  0.5977052  0.4628322
  7     0.001  0.6056173  0.4741502
  7     0.100  0.6558893  0.5409668
  9     0.000  0.5893983  0.4525378
  9     0.001  0.6070056  0.4756937
  9     0.100  0.6579200  0.5437058

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were size = 9 and decay = 0.1.
```

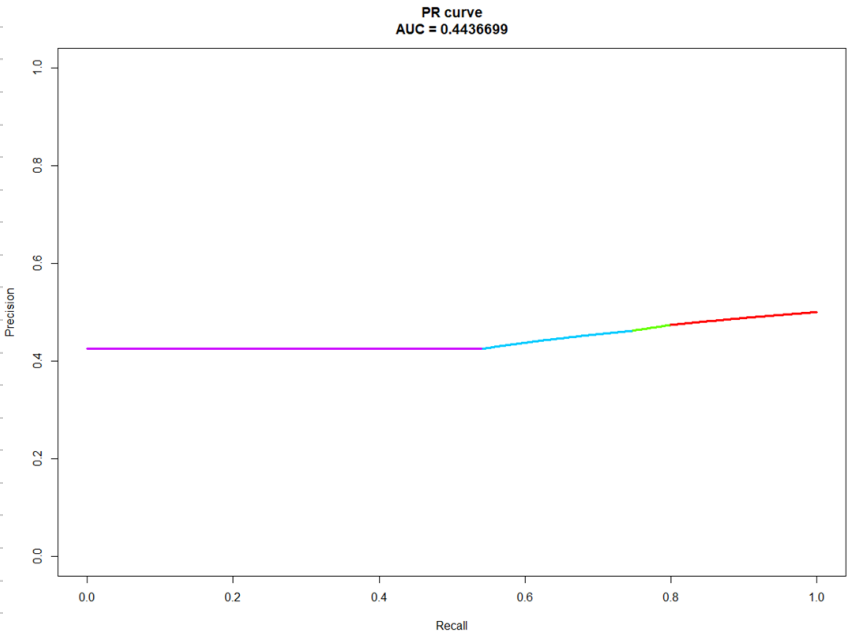
```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4
         1  17   2   5  31
         2   2   1   4   8
         3  14   3  23 108
         4  12   1  24 147

Overall Statistics

               Accuracy : 0.4677
                 95% CI : (0.418, 0.5178)
    No Information Rate : 0.7313
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1098

 Mcnemar's Test P-Value : 1.831e-13

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Precision             0.30909 0.066667  0.15541   0.7989
Recall                0.37778 0.142857  0.41071   0.5000
F1                    0.34000 0.090909  0.22549   0.6151
Prevalence            0.11194 0.017413  0.13930   0.7313
Detection Rate        0.04229 0.002488  0.05721   0.3657
Detection Prevalence  0.13682 0.037313  0.36816   0.4577
Balanced Accuracy     0.63567 0.553707  0.52472   0.5787
```



SVM

```
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 844, 845, 844, 846, 845, 846, ...
Resampling results across tuning parameters:

  C     Accuracy   Kappa
  0.00       NaN       NaN
  0.01  0.7015701  0.6020842
  0.05  0.7000864  0.6001068
  0.10  0.6993528  0.5991184
  0.25  0.6964130  0.5951957
  0.50  0.6936321  0.5914812
  0.75  0.6890502  0.5853709
  1.00  0.6896931  0.5862297
  1.25  0.6877744  0.5836716
  1.50  0.6899206  0.5865322
  1.75  0.6910932  0.5880977
  2.00  0.6908803  0.5878214
  5.00  0.6916352  0.5888175

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was C = 0.01.
```

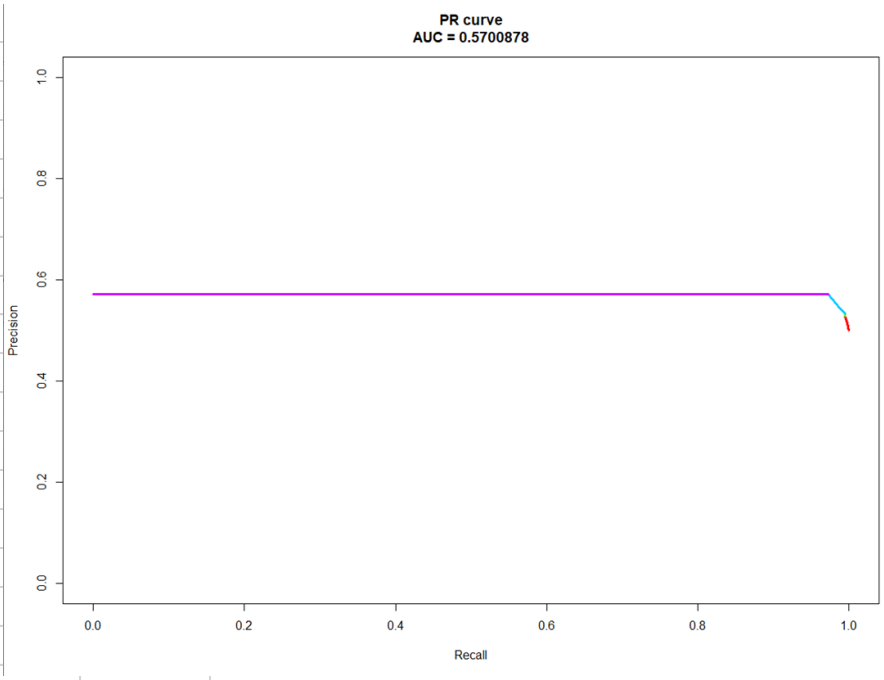
```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4
         1  32   2   6  41
         2   2   1   7  11
         3   0   3  16  64
         4  11   1  27 178

Overall Statistics

               Accuracy : 0.5647
                 95% CI : (0.5146, 0.6138)
    No Information Rate : 0.7313
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2128

 Mcnemar's Test P-Value : 1.036e-08

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Precision              0.3951 0.047619   0.1928   0.8203
Recall                 0.7111 0.142857   0.2857   0.6054
F1                     0.5079 0.071429   0.2302   0.6967
Prevalence             0.1119 0.017413   0.1393   0.7313
Detection Rate         0.0796 0.002488   0.0398   0.4428
Detection Prevalence   0.2015 0.052239   0.2065   0.5398
Balanced Accuracy      0.7869 0.546112   0.5460   0.6222
```



NB

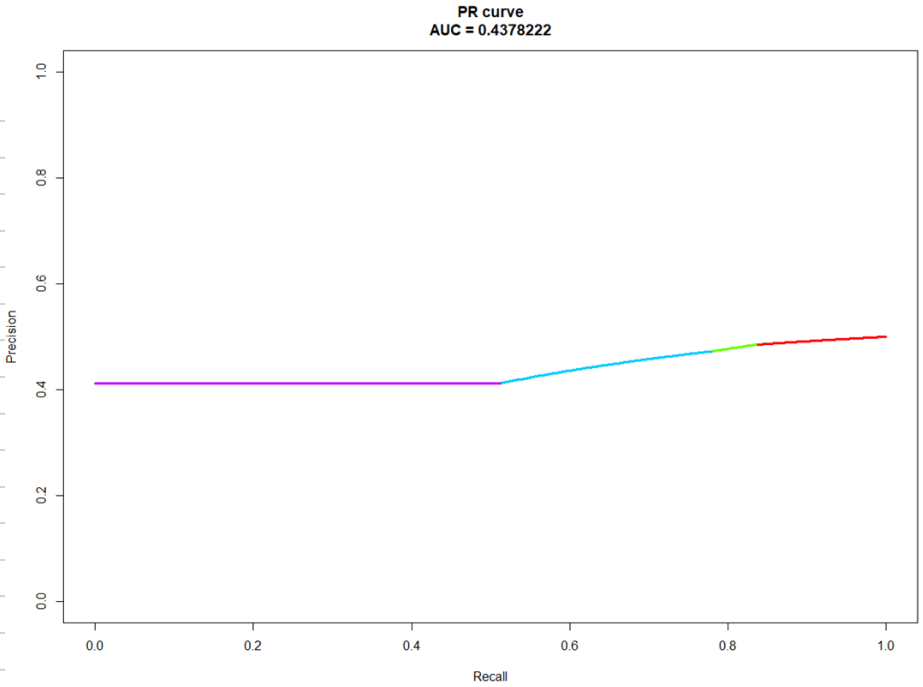
```
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 846, 845, 846, 845, 845, 846, ...
Resampling results across tuning parameters:

  usekernel  Accuracy   Kappa
  FALSE      0.2790739  0.03953739
   TRUE      0.4516325  0.26850035

Tuning parameter 'laplace' was held constant at a value of 0
Tuning parameter 'adjust' was held constant at a value of 1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were laplace = 0, usekernel = TRUE and adjust = 1.
```

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4
         1   1   0   1   0
         2   0   0   0   0
         3   0   0   1   8
         4  44   7  54 286

Overall Statistics

               Accuracy : 0.7164
                 95% CI : (0.6696, 0.76)
    No Information Rate : 0.7313
    P-Value [Acc > NIR] : 0.7689

                  Kappa : 0.005

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Precision            0.500000       NA 0.111111   0.7315
Recall               0.022222  0.00000 0.017857   0.9728
F1                   0.042553       NA 0.030769   0.8350
Prevalence           0.111940  0.01741 0.139303   0.7313
Detection Rate       0.002488  0.00000 0.002488   0.7114
Detection Prevalence 0.004975  0.00000 0.022388   0.9726
Balanced Accuracy    0.509711  0.50000 0.497368   0.5003
```



k-NN

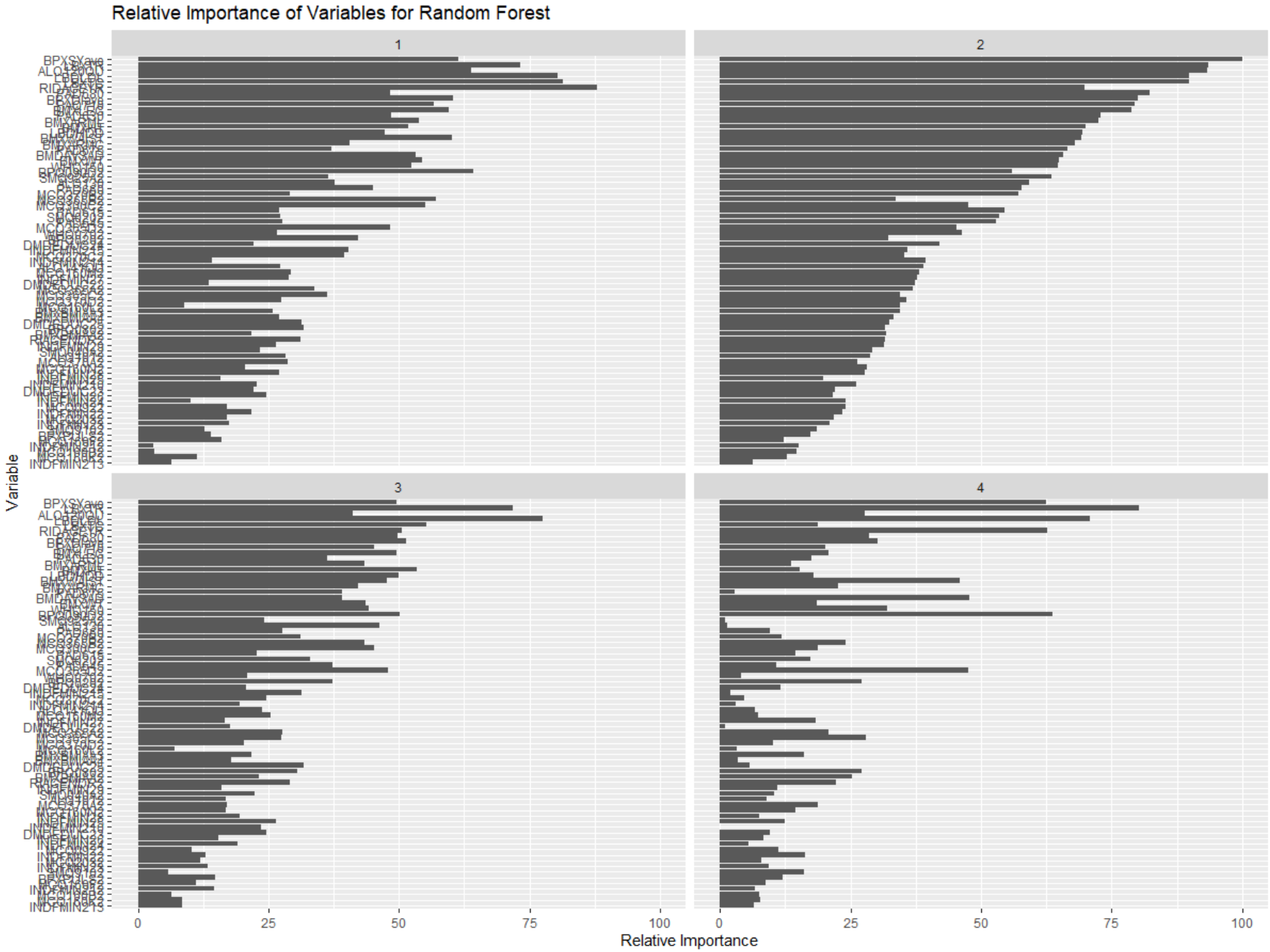
```
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 846, 847, 845, 844, 843, 845, ...
Resampling results across tuning parameters:

  kmax  Accuracy   Kappa
  5     0.7628070  0.6837204
  7     0.7617317  0.6822833
  9     0.7617317  0.6822833

Tuning parameter 'distance' was held constant at a value of 2
Tuning parameter 'kernel' was held constant at a value of optimal
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were kmax = 5, distance = 2 and kernel = optimal.
```



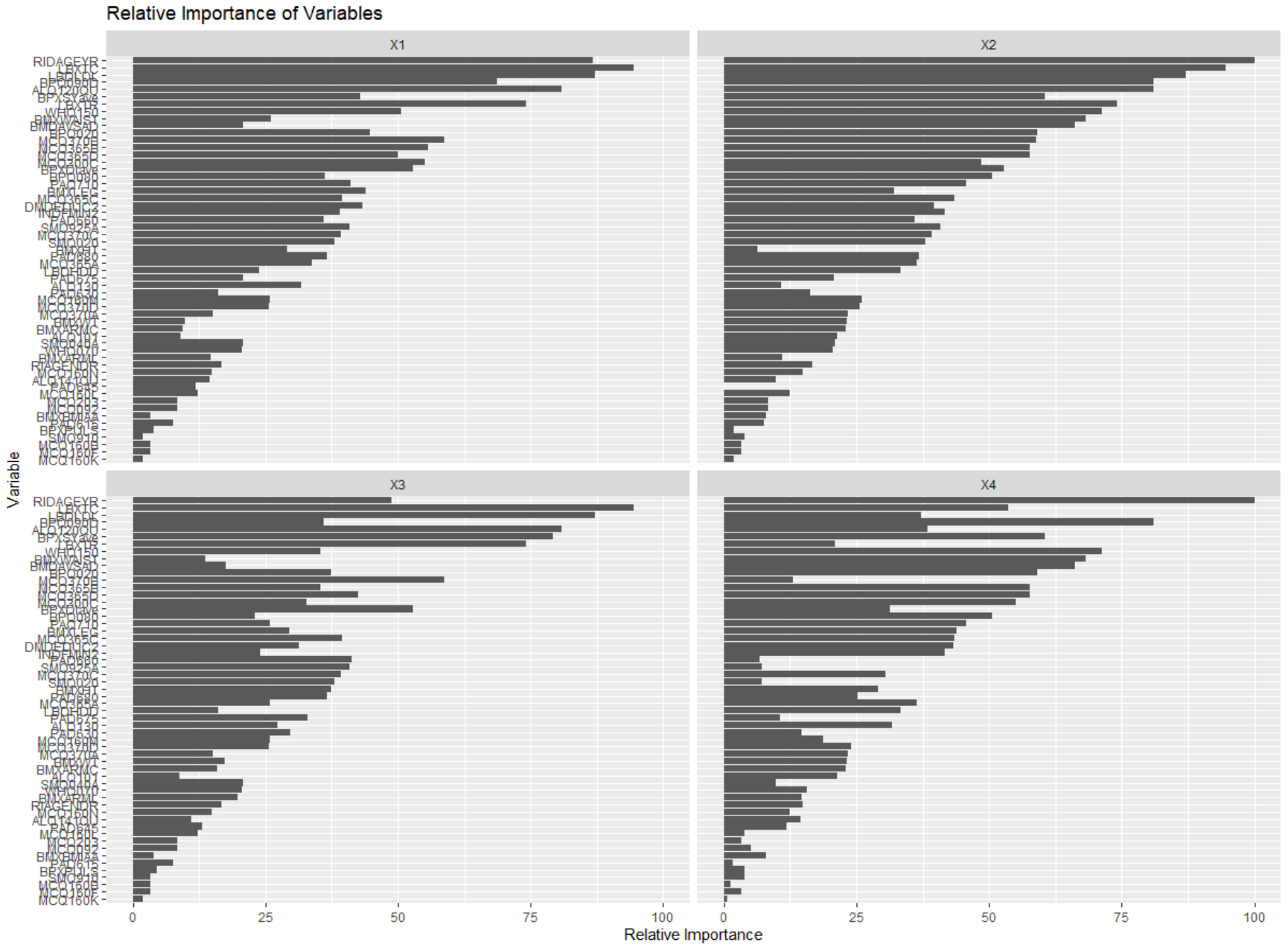

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4
         1  16   1   6  43
         2   3   3   5  12
         3   6   0  20  81
         4  20   3  25 158

Overall Statistics

               Accuracy : 0.49
                 95% CI : (0.4402, 0.5401)
    No Information Rate : 0.7313
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1034

 Mcnemar's Test P-Value : 6.253e-09

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Precision              0.2424 0.130435  0.18692   0.7670
Recall                 0.3556 0.428571  0.35714   0.5374
F1                     0.2883 0.200000  0.24540   0.6320
Prevalence             0.1119 0.017413  0.13930   0.7313
Detection Rate         0.0398 0.007463  0.04975   0.3930
Detection Prevalence   0.1642 0.057214  0.26617   0.5124
Balanced Accuracy      0.6077 0.688969  0.55285   0.5465
```

Appendix J – varImp() plots for RF and k-NN

RF


Relative Importance of Variables for Random Forest
1
2
3
4
Variable
0
25
50
75
100
Relative Importance

k-NN


Relative Importance of Variables
X1
X2
X3
X4
Variable
RIDAGEYR
LBXTC
LBDLDL
BPQ090D
ALQ120Q
BPXSY2
LBXTR
WHD150
BMXWAIST
BMDAVSAD
BPQ020
MCQ370B
MCQ365B
MCQ365D
MCQ300C
BPXDI2
BPQ080
PAD710
BMXLEG
MCQ365C
DMDEDUC2
INDFMIN2
PAD680
SMQ925A
MCQ370C
SMQ020
BMXWT
PAD680
MCQ184A
LBDHDD
PAD575
ALQ130
PAD630
MCQ195M
MCQ370D
MCQ370A
BMXWT
BMXARMC
ALQ101
SMQ040A
WHQ070
BMXBMI
RIAGENDR
MCQ160N
ALQ140Q
PAD645
MCQ160L
MCQ220
BMXBMIA
PAD615
BPXPLS
SMQ910
MCQ160B
MCQ160F
MCQ160K
0
25
50
75
100
Relative Importance

Appendix K – varImp() results for robustness tests (RF and k-NN)

Robustness test – Exclude cholesterol

```
> varImp(rfB1_maxtrees_tc)
rf variable importance

  variables are sorted by maximum importance across the classes
  only 20 most important variables shown (out of 69)

             1      2     3      4
BPXSYave 55.28 100.00 49.03 63.263
ALQ120QU 57.31  91.58 34.38 24.890
RIDAGEYR 84.66  72.11 48.19 54.470
BPXDIave 55.95  79.46 41.00 19.389
PAD680   45.28  79.33 44.06 29.671
PAQ710   51.81  74.90 43.25 19.264
BMXLEG   55.62  72.56 45.99 14.094
BMXHT    49.09  70.45 46.33 11.495
BMXWAIST 54.26  69.30 42.92 39.173
BMXARML  49.78  66.53 43.01 14.366
BMXARMC  40.76  66.32 36.52 15.392
BMXWT    48.76  65.50 39.02 16.288
PAD675   34.79  65.19 33.84  1.388
BMDAVSAD 52.01  64.35 32.94 49.967
PAD630   45.97  63.72 29.49 13.383
BPQ090D2 63.10  54.38 43.19 54.861
WHQ150   51.30  62.72 32.94 30.264
SMQ925A2 35.36  60.27 19.31  2.136
MCQ370B2 31.71  57.38 25.09 12.539
PAD660   35.62  54.02 26.15  4.272
```

k-NN

```
ROC curve variable importance

  variables are sorted by maximum importance across the classes
  only 20 most important variables shown (out of 52)

             X1     X2    X3      X4
RIDAGEYR 86.69 100.00 48.83 100.000
BPQ090D  68.71  81.03 35.93  81.033
ALQ120QU 80.92  80.92 80.92  38.287
BPXSYave 42.80  60.45 79.15  60.449
WHQ150   50.69  71.30 35.44  71.300
BMXWAIST 26.05  68.16 13.66  68.160
BMDAVSAD 20.78  66.25 17.55  66.254
BPQ020   44.73  58.99 37.38  58.990
MCQ370B  58.79  58.79 58.79  12.958
MCQ365D  49.91  57.69 42.46  57.693
MCQ365B  55.75  57.69 35.38  57.693
MCQ300C  55.10  48.62 32.80  55.100
BPXDIave 52.78  52.78 52.78  31.296
BPQ080   36.30  50.56 23.05  50.561
PAQ710   41.12  45.68 25.88  45.675
BMXLEG   43.86  32.06 29.41  43.860
MCQ365C  39.44  43.43 39.44  43.430
DMDEDUC2 43.36  39.62 31.29  43.355
INDFMIN2 39.12  41.67 24.01  41.672
PAD660   36.03  36.03 41.27   6.674
```

Robustness test – Categorical cholesterol

```
> varImp(sa3_rf)
rf variable importance

  variables are sorted by maximum importance across the classes
  only 20 most important variables shown (out of 73)

             1      2     3      4
BPXSYave 58.87 100.00 48.25 58.381
ALQ120QU 69.07  98.33 41.13 34.596
RIDAGEYR 94.63  81.23 51.50 60.708
PAD680   57.10  85.15 43.84 18.583
PAQ710   59.52  81.68 48.49 18.372
BPXDIave 63.54  81.23 49.47 17.786
BMXLEG   65.66  78.60 49.83 17.021
BMXWAIST 64.76  78.28 46.96 47.557
BMXARML  57.20  76.80 48.30 23.083
BMXWT    57.35  75.53 47.67 26.756
PAD630   51.39  75.08 41.21  7.989
MCQ300C2 74.83  55.36 58.02 27.343
BMXHT    59.97  74.55 54.26 22.307
BMDAVSAD 61.76  74.12 43.83 46.414
BMXARMC  48.82  73.82 40.31 35.114
PAD675   44.57  73.65 44.57 14.172
WHQ150   56.79  72.43 46.24 26.165
SMQ925A2 33.60  72.09 29.09 24.602
BPQ090D2 71.21  59.71 55.21 68.254
PAD615   29.99  66.59 24.84 25.342
```

```
> varImp(knn_sa3, scale=FALSE)
ROC curve variable importance

  variables are sorted by maximum importance across the classes
  only 20 most important variables shown (out of 56)

             X1     X2     X3     X4
RIDAGEYR 0.7647 0.8102 0.6531 0.8102
ALQ120QU 0.7641 0.7641 0.7641 0.6037
BPQ090D  0.7277 0.7638 0.6222 0.7638
BMXWAIST 0.5826 0.7385 0.5643 0.7385
MCQ300C  0.7255 0.7021 0.6223 0.7255
BMDAVSAD 0.5696 0.7234 0.5481 0.7234
BPXSYave 0.6197 0.6902 0.7233 0.6902
WHQ150   0.6313 0.7017 0.6313 0.7017
MCQ370B  0.6951 0.6951 0.6951 0.5468
MCQ365D  0.6809 0.6596 0.6500 0.6809
BPQ020   0.6085 0.6638 0.6014 0.6638
BPQ080   0.6191 0.6638 0.5757 0.6638
BPXDIave 0.6598 0.6598 0.6598 0.6162
MCQ365B  0.6553 0.6574 0.6162 0.6574
DMDEDUC2 0.6448 0.6162 0.5931 0.6448
MCQ370C  0.6375 0.6375 0.6375 0.6319
SMQ925A  0.6362 0.6362 0.6362 0.5213
BMXLEG   0.6357 0.6016 0.5857 0.6357
INDFMIN2 0.6356 0.6343 0.5785 0.6356
MCQ365C  0.6209 0.6340 0.6209 0.6340
```

Robustness test – Laboratory data

```
> varImp(balancedB2_rf)
rf variable importance

  variables are sorted by maximum importance across the classes
  only 20 most important variables shown (out of 77)

             1      2     3      4
BPXSYave 63.00 100.00 40.92 65.378
RIDAGEYR 94.58  74.61 50.56 78.417
LBDLDL   80.93  92.10 70.94 73.276
LBXTC    78.34  91.45 49.53 20.780
ALQ120QU 65.70  91.26 39.87 10.018
LBXTR    66.50  89.98 68.65 70.694
LBXSGTSI 54.13  81.08 46.77 46.348
BMXLEG   58.35  81.06 49.13 15.204
BPXDIave 62.33  80.10 40.91 31.680
PAQ710   57.62  79.58 51.55 13.855
LBDHDD   51.45  79.03 50.88  8.527
PAD680   51.50  78.54 53.57 30.725
LBXSPH   52.38  76.85 41.49 26.718
LBXSKSI  60.22  75.07 41.94 29.631
PAD630   49.29  74.98 40.63  9.203
BMXHT    49.97  72.49 48.39 11.230
BMXARML  50.83  72.30 43.94 19.612
BMXWAIST 57.81  70.31 45.33 33.735
LBXSCA   57.47  69.68 42.89 13.277
SMQ925A2 39.06  69.40 30.35  5.650
```

k-NN

```
ROC curve variable importance

  variables are sorted by maximum importance across the classes
  only 20 most important variables shown (out of 60)

             X1     X2     X3     X4
RIDAGEYR 0.7753 0.8494 0.6800 0.8494
LBXTC    0.8066 0.8066 0.8066 0.6873
LBDLDL   0.7891 0.7891 0.7891 0.6305
ALQ120QU 0.7575 0.7575 0.7575 0.6148
BPQ090D  0.7234 0.7553 0.6436 0.7553
BPXSYave 0.6492 0.7082 0.7459 0.7082
WHQ150   0.6487 0.7406 0.6109 0.7406
LBXTR    0.7236 0.7236 0.7236 0.5742
MCQ300C  0.6979 0.6702 0.6183 0.6979
BMDAVSAD 0.5511 0.6884 0.5392 0.6884
BPQ080   0.6553 0.6872 0.5797 0.6872
BMXWAIST 0.5765 0.6868 0.5521 0.6868
MCQ365D  0.6532 0.6447 0.6863 0.6532
BPXDIave 0.6858 0.6858 0.6858 0.6343
BPQ020   0.6128 0.6809 0.5801 0.6809
MCQ365B  0.6489 0.6660 0.5841 0.6660
LBXSKSI  0.6139 0.6654 0.6139 0.6654
MCQ370B  0.6653 0.6653 0.6653 0.5191
LBXSPH   0.6596 0.6596 0.6596 0.5602
PAQ710   0.6536 0.6353 0.6090 0.6536
```

Robustness test – Physical activity

```
rf variable importance

  variables are sorted by maximum importance across the classes
  only 20 most important variables shown (out of 73)

             1      2     3     4
LBXTR    71.03 100.00 63.97 78.21
BPXSYave 54.47  98.18 45.96 52.28
LBDLDL   77.39  92.85 68.89 63.92
ALQ120QU 53.93  91.05 45.03 20.23
LBXTC    77.77  91.04 45.26 18.59
RIDAGEYR 89.43  82.22 50.50 61.61
PAD680   47.06  82.13 40.09 25.31
BPXDIave 58.90  81.99 44.42 23.67
PAQ710   55.28  81.15 41.56 27.39
BMXLEG   58.75  77.95 42.75 11.82
LBDHDD   49.42  77.47 49.43 11.10
BMXARML  53.96  75.76 44.89 26.52
BMXWAIST 63.67  74.93 49.16 35.67
BMXWT    58.45  73.56 46.56 20.88
BMXHT    54.68  73.08 50.28 12.12
WHQ150   54.72  71.21 42.95 22.13
BMXARMC  41.79  70.44 44.29 16.40
BMDAVSAD 52.44  69.70 43.78 40.56
BPQ090D2 67.18  55.72 43.09 63.51
SMQ925A2 39.24  65.35 19.11 10.21
```

k-NN

```
ROC curve variable importance

  variables are sorted by maximum importance across the classes
  only 20 most important variables shown (out of 56)

             X1     X2     X3     X4
RIDAGEYR 0.7739 0.8254 0.6645 0.8254
LBXTC    0.8020 0.8020 0.8020 0.6901
BPQ090D  0.7255 0.7723 0.6247 0.7723
LBDLDL   0.7708 0.7708 0.7708 0.6494
ALQ120QU 0.7450 0.7450 0.7450 0.5906
BPXSYave 0.6583 0.6686 0.7439 0.6686
WHQ150   0.6454 0.7334 0.6047 0.7334
LBXTR    0.7270 0.7270 0.7270 0.5884
BMXWAIST 0.5864 0.7187 0.5720 0.7187
BMDAVSAD 0.5574 0.7122 0.5321 0.7122
MCQ365B  0.6830 0.7000 0.6332 0.7000
MCQ300C  0.6979 0.6936 0.6012 0.6979
BPXDIave 0.6843 0.6843 0.6843 0.6206
BPQ020   0.6228 0.6787 0.6228 0.6787
MCQ365D  0.6723 0.6787 0.6415 0.6787
MCQ370B  0.6652 0.6652 0.6652 0.5532
BPQ080   0.6532 0.6638 0.5970 0.6638
PAQ710   0.6332 0.6500 0.5927 0.6500
BMXLEG   0.6478 0.6008 0.5843 0.6478
MCQ365C  0.6401 0.6401 0.6401 0.6383
```

Robustness test – Smoking

rf variable importance

variables are sorted by maximum importance across the classes
only 20 most important variables shown (out of 73)

	1	2	3	4
LBXTR	72.81	100.00	74.19	85.600
BPXSYave	59.33	97.90	45.52	53.856
ALQ120QU	65.73	94.84	41.60	23.062
RIDAGEYR	94.07	81.84	50.10	70.794
LBXTC	79.92	91.51	56.26	5.666
LBDLDL	82.88	90.68	75.54	70.154
BPXDIave	63.79	86.62	41.06	20.213
PAQ710	66.35	84.87	51.04	20.565
BMXLEG	54.02	82.15	49.87	16.730
PAD680	57.38	81.96	47.91	26.234
BMXWT	62.38	80.21	45.74	28.967
LBDHDD	53.38	79.66	60.06	22.064
BMXHT	55.47	79.48	53.05	22.632
BMXWAIST	63.12	76.84	50.54	35.592
PAD630	52.15	76.65	51.16	11.151
MCQ300C2	76.57	52.00	57.65	11.403
BMDAVSAD	61.70	76.45	45.93	47.786
BMXARML	54.91	76.37	43.20	27.291
PAD675	47.21	74.07	35.34	11.890
BMXARMC	47.98	72.97	44.55	32.058

k-NN

ROC curve variable importance

variables are sorted by maximum importance across the classes
only 20 most important variables shown (out of 56)

	X1	X2	X3	X4
RIDAGEYR	0.7769	0.8332	0.6533	0.8332
LBXTC	0.7900	0.7900	0.7900	0.6930
ALQ120QU	0.7643	0.7643	0.7643	0.6083
LBXTR	0.7505	0.7505	0.7505	0.5856
BPQ090D	0.7213	0.7447	0.6350	0.7447
LBDLDL	0.7435	0.7435	0.7435	0.6399
BPXSYave	0.6444	0.6754	0.7399	0.6754
WHQ150	0.6630	0.7335	0.6068	0.7335
MCQ300C	0.7213	0.6872	0.6524	0.7213
BPQ020	0.6319	0.6894	0.5885	0.6894
MCQ370B	0.6887	0.6887	0.6887	0.5553
BMDAVSAD	0.5872	0.6849	0.5563	0.6849
BMXWAIST	0.5912	0.6847	0.5501	0.6847
BPQ080	0.6404	0.6787	0.5735	0.6787
MCQ370C	0.6780	0.6780	0.6780	0.6213
PAQ710	0.6487	0.6744	0.6097	0.6744
MCQ365B	0.6681	0.6702	0.6183	0.6702
MCQ365D	0.6532	0.6617	0.6457	0.6617
MCQ365C	0.6529	0.6529	0.6529	0.6404
BPXDIave	0.6493	0.6493	0.6493	0.6132

Robustness Test – Blood pressure

rf variable importance

variables are sorted by maximum importance across the classes
only 20 most important variables shown (out of 73)

	1	2	3	4
LBXTR	68.86	100.00	75.52	88.172
ALQ120QU	63.17	99.39	42.10	38.466
LBDLDL	84.10	94.74	80.95	80.574
LBXTC	83.59	93.59	59.06	19.561
RIDAGEYR	90.06	76.08	60.26	62.501
BPXSYaveC2	35.13	87.73	25.60	47.216
PAQ710	57.75	84.33	50.90	21.541
BMXLEG	63.04	81.16	59.38	11.720
BMXARML	52.53	80.52	48.94	18.260
PAD680	56.93	80.11	48.08	16.082
LBDHDD	48.89	77.70	50.41	11.127
BPQ090D2	70.70	65.30	61.45	75.311
BMXARMC	41.71	75.30	42.40	26.089
BMXHT	59.92	75.17	60.27	14.541
PAD630	51.12	72.38	38.51	13.525
BMXWT	54.29	72.34	46.98	27.179
BMXWAIST	56.53	72.31	51.89	39.318
PAD675	42.61	71.60	38.17	9.321
MCQ370B2	37.50	69.96	30.15	18.822
SMQ925A2	35.19	67.51	19.34	2.878

k-NN

ROC curve variable importance

variables are sorted by maximum importance across the classes
only 20 most important variables shown (out of 56)

	X1	X2	X3	X4
RIDAGEYR	0.7905	0.8142	0.6697	0.8142
LBXTC	0.7797	0.7797	0.7797	0.6798
LBDLDL	0.7751	0.7751	0.7751	0.6432
BPQ090D	0.7447	0.7723	0.6863	0.7723
ALQ120QU	0.7544	0.7544	0.7544	0.6044
LBXTR	0.7483	0.7483	0.7483	0.5798
WHQ150	0.6600	0.7294	0.6059	0.7294
MCQ300C	0.7128	0.6851	0.6180	0.7128
BPXSYaveC	0.6509	0.6509	0.6999	0.6213
MCQ370B	0.6994	0.6994	0.6994	0.5319
BPQ080	0.6511	0.6915	0.5756	0.6915
BMXWAIST	0.5741	0.6906	0.5573	0.6906
BMDAVSAD	0.5681	0.6896	0.5597	0.6896
MCQ365D	0.6766	0.6617	0.6863	0.6766
BPQ020	0.6234	0.6745	0.5801	0.6745
MCQ365B	0.6745	0.6660	0.5882	0.6745
PAD660	0.6303	0.6303	0.6533	0.5299
BMXLEG	0.6511	0.6012	0.5914	0.6511
PAQ710	0.6436	0.6506	0.6001	0.6506
MCQ370C	0.6482	0.6482	0.6482	0.6106

Robustness test – excluding height and weight

rf variable importance

variables are sorted by maximum importance across the classes
only 20 most important variables shown (out of 71)

	1	2	3	4
BPXSYave	61.52	100.00	51.82	63.690
LBXTR	72.65	94.66	70.58	80.440
LBDLDL	80.29	90.60	79.03	73.357
ALQ120QU	59.94	90.35	37.20	30.734
RIDAGEYR	89.50	74.40	51.54	64.472
LBXTC	82.56	87.39	53.87	17.405
BPXDIave	61.77	82.58	49.69	29.264
PAD680	48.46	81.01	50.84	32.992
PAQ710	56.43	79.09	43.39	19.405
BMXARML	52.51	76.22	45.03	17.492
BMXLEG	57.22	76.14	51.44	21.141
PAD630	51.19	71.84	37.47	18.670
BMXWAIST	59.03	70.51	46.09	42.852
PAD675	39.32	69.46	34.09	7.595
BMXARMC	41.25	67.39	43.42	18.303
BMDAVSAD	54.41	67.25	43.53	48.189
BPQ090D2	66.82	58.02	50.42	63.908
LBDHDD	46.26	66.66	52.60	15.438
SMQ925A2	37.96	65.67	26.60	0.000
WHQ150	53.84	64.35	42.28	33.762

k-NN

ROC curve variable importance

variables are sorted by maximum importance across the classes
only 20 most important variables shown (out of 54)

	x1	x2	x3	x4
RIDAGEYR	0.7867	0.8303	0.6624	0.8303
LBXTC	0.8124	0.8124	0.8124	0.6784
LBDLDL	0.7879	0.7879	0.7879	0.6239
BPQ090D	0.7277	0.7681	0.6201	0.7681
ALQ120QU	0.7677	0.7677	0.7677	0.6278
BPXSYave	0.6426	0.7005	0.7619	0.7005
LBXTR	0.7452	0.7452	0.7452	0.5711
WHQ150	0.6685	0.7361	0.6185	0.7361
BMXWAIST	0.5876	0.7258	0.5470	0.7258
BMDAVSAD	0.5703	0.7196	0.5597	0.7196
BPQ020	0.6489	0.6957	0.6248	0.6957
MCQ370B	0.6951	0.6951	0.6951	0.5447
MCQ365D	0.6660	0.6915	0.6415	0.6915
MCQ365B	0.6851	0.6915	0.6182	0.6915
MCQ300C	0.6830	0.6617	0.6098	0.6830
BPXDIave	0.6754	0.6754	0.6754	0.6049
BPQ080	0.6213	0.6681	0.5778	0.6681
PAQ710	0.6371	0.6521	0.5871	0.6521
BMXLEG	0.6461	0.6074	0.5987	0.6461
MCQ365C	0.6316	0.6447	0.6316	0.6447

Robustness test – BMXBMIO

rf variable importance

variables are sorted by maximum importance across the classes
only 20 most important variables shown (out of 73)

	1	2	3	4
BPXSYave	64.76	100.00	48.42	52.438
ALQ120QU	61.91	95.42	40.08	17.746
LBXTR	72.67	93.36	73.37	77.097
RIDAGEYR	89.13	77.83	53.56	58.967
LBXTC	85.90	88.50	49.25	24.468
LBDLDL	82.24	88.17	74.81	67.928
BPXDIave	64.37	83.05	38.83	20.801
PAQ710	51.93	78.15	46.27	10.582
PAD680	45.98	78.14	42.01	22.102
PAD630	52.69	77.77	39.92	11.666
BMXARML	49.02	76.48	40.54	22.807
BMXLEG	52.53	75.90	45.23	11.603
LBDHDD	43.82	75.38	51.27	22.136
BMXWAIST	62.68	74.85	46.79	22.107
BMXHT	49.51	73.41	52.07	16.036
PAD675	46.54	73.37	38.24	7.058
BMDAVSAD	60.16	73.22	43.91	29.368
BPQ090D2	72.53	60.96	62.04	63.102
BMXARMC	48.76	71.91	38.16	23.948
BMXWT	53.25	71.43	45.38	17.112

k-NN

ROC curve variable importance

variables are sorted by maximum importance across the classes
only 20 most important variables shown (out of 56)

	x1	x2	x3	x4
RIDAGEYR	0.8008	0.8280	0.6871	0.8280
LBXTC	0.7960	0.7960	0.7960	0.7055
BPXSYave	0.6317	0.6881	0.7694	0.6881
BPQ090D	0.7447	0.7617	0.6499	0.7617
ALQ120QU	0.7507	0.7507	0.7507	0.6305
LBDLDL	0.7464	0.7464	0.7464	0.6647
LBXTR	0.7237	0.7237	0.7237	0.5747
WHQ150	0.6641	0.7172	0.6133	0.7172
MCQ300C	0.6957	0.6809	0.6076	0.6957
BPQ020	0.6468	0.6957	0.5885	0.6957
MCQ370B	0.6951	0.6951	0.6951	0.5319
BMXWAIST	0.5991	0.6833	0.5891	0.6833
MCQ365D	0.6809	0.6511	0.6265	0.6809
BPXDIave	0.6800	0.6800	0.6800	0.6120
MCQ365B	0.6787	0.6489	0.6204	0.6787
BMDAVSAD	0.5900	0.6709	0.5446	0.6709
DMDEDUC2	0.6144	0.6139	0.6618	0.6144
MCQ370C	0.6525	0.6525	0.6525	0.6064
BPQ080	0.6362	0.6511	0.5714	0.6511
SMQ925A	0.6468	0.6468	0.6468	0.5298

Robustness test – BMXBMIO excluding height and weight

rf variable importance

variables are sorted by maximum importance across the classes
only 20 most important variables shown (out of 71)

	1	2	3	4
ALQ120QU	63.20	100.00	42.64	23.531
BPXSYave	59.40	99.82	48.29	50.101
LBXTR	74.04	97.68	72.90	80.257
RIDAGEYR	95.04	79.96	58.38	63.387
LBXTC	90.76	90.07	53.12	21.159
LBDLDL	82.90	90.17	75.25	69.398
BPXDIave	64.24	88.09	37.34	18.746
BMXLEG	51.56	83.33	46.81	13.932
PAD680	47.19	81.80	43.27	20.336
PAQ710	56.48	79.97	49.83	10.051
BMXARML	46.25	78.70	42.26	19.230
PAD630	51.30	77.83	42.93	12.796
PAD675	46.90	76.86	37.05	7.882
BMDAVSAD	57.74	76.70	47.91	33.167
WHQ150	60.04	76.31	40.84	32.710
BMXARMC	49.19	75.91	43.27	21.823
SMQ925A2	39.81	75.86	29.63	7.822
BMXWAIST	63.30	75.86	46.45	26.949
BPQ090D2	75.11	64.89	68.80	64.384
LBDHDD	45.19	74.95	53.73	23.092

k-NN

ROC curve variable importance

variables are sorted by maximum importance across the classes
only 20 most important variables shown (out of 54)

	X1	X2	X3	X4
RIDAGEYR	0.8008	0.8280	0.6871	0.8280
LBXTC	0.7960	0.7960	0.7960	0.7055
BPXSYave	0.6317	0.6881	0.7694	0.6881
BPQ090D	0.7447	0.7617	0.6499	0.7617
ALQ120QU	0.7507	0.7507	0.7507	0.6305
LBDLDL	0.7464	0.7464	0.7464	0.6647
LBXTR	0.7237	0.7237	0.7237	0.5747
WHQ150	0.6641	0.7172	0.6133	0.7172
BPQ020	0.6468	0.6957	0.5885	0.6957
MCQ300C	0.6957	0.6809	0.6076	0.6957
MCQ370B	0.6951	0.6951	0.6951	0.5319
BMXWAIST	0.5991	0.6833	0.5891	0.6833
MCQ365D	0.6809	0.6511	0.6265	0.6809
BPXDIave	0.6800	0.6800	0.6800	0.6120
MCQ365B	0.6787	0.6489	0.6204	0.6787
BMDAVSAD	0.5900	0.6709	0.5446	0.6709
DMDEDUC2	0.6144	0.6139	0.6618	0.6144
MCQ370C	0.6525	0.6525	0.6525	0.6064
BPQ080	0.6362	0.6511	0.5714	0.6511
SMQ925A	0.6468	0.6468	0.6468	0.5298

Robustness test – BMXBMI

rf variable importance

variables are sorted by maximum importance across the classes
only 20 most important variables shown (out of 71)

	1	2	3	4
BPXSYave	60.49	100.00	39.21	57.77
ALQ120QU	65.40	92.01	41.82	26.31
LBXTR	66.07	90.65	64.58	79.58
LBXTC	86.07	87.61	52.79	20.37
LBDLDL	78.15	85.71	71.95	68.78
RIDAGEYR	83.90	73.21	52.62	54.67
BPXDIave	53.53	78.06	43.45	19.11
PAD680	48.34	76.70	42.40	19.48
BMXLEG	49.00	76.29	48.43	14.34
PAQ710	56.92	74.88	46.82	16.30
BMXARML	46.45	73.09	41.58	20.43
BMXWAIST	59.00	72.02	40.81	38.65
BMXHT	51.09	70.94	47.11	14.97
PAD630	47.70	70.37	37.74	13.49
BMXARMC	40.98	68.49	34.58	20.87
BMDAVSAD	53.07	68.47	39.04	42.66
SMQ925A2	40.05	68.13	27.18	16.64
BMXWT	53.45	66.57	38.41	23.40
LBDHDD	50.50	66.50	44.93	17.11
BMXBMI	44.61	66.03	34.14	20.45

k-NN

ROC curve variable importance

variables are sorted by maximum importance across the classes
only 20 most important variables shown (out of 56)

	X1	X2	X3	X4
LBXTC	0.8052	0.8052	0.8052	0.6943
RIDAGEYR	0.7643	0.8021	0.6604	0.8021
ALQ120QU	0.7755	0.7755	0.7755	0.5916
BPQ090D	0.7340	0.7532	0.6564	0.7532
LBDLDL	0.7502	0.7502	0.7502	0.6383
BPXSYave	0.6495	0.6767	0.7411	0.6767
LBXTR	0.7239	0.7239	0.7239	0.5826
BMXWAIST	0.5930	0.7083	0.5603	0.7083
MCQ300C	0.7000	0.6660	0.6180	0.7000
BMDAVSAD	0.5719	0.6989	0.5625	0.6989
WHQ150	0.6451	0.6976	0.6166	0.6976
MCQ365B	0.6766	0.6830	0.6140	0.6830
MCQ370B	0.6780	0.6780	0.6780	0.5404
BPQ020	0.6319	0.6766	0.6035	0.6766
BPQ080	0.6319	0.6766	0.5841	0.6766
BPXDIave	0.6683	0.6683	0.6683	0.5985
LBDHDD	0.5962	0.6418	0.5516	0.6418
MCQ365D	0.6340	0.6362	0.6393	0.6362
SMQ925A	0.6383	0.6383	0.6383	0.5298
PAD660	0.6127	0.6127	0.6363	0.5304

Robustness test – BMXBMI excluding height and weight

rf variable importance

variables are sorted by maximum importance across the classes
only 20 most important variables shown (out of 69)

	1	2	3	4
BPXSYave	57.01	100.00	37.48	56.325
ALQ120QU	62.49	93.52	43.04	23.882
LBXTR	65.85	92.64	65.59	77.667
LBXTC	82.67	91.06	53.97	15.606
LBDLDL	77.66	85.83	71.31	68.062
RIDAGEYR	83.45	69.91	51.84	49.030
BPXDIave	58.31	78.58	41.19	18.733
PAQ710	48.57	77.65	46.77	14.145
BMXLEG	48.03	76.31	52.45	6.046
PAD680	47.24	73.58	40.43	16.705
PAD630	45.69	70.52	38.36	5.306
BMXARML	43.46	70.34	45.32	17.111
SMQ925A2	34.32	69.53	25.12	10.924
BMXWAIST	56.54	68.31	39.92	34.268
LBDHDD	49.26	68.18	43.24	9.931
BMXARMC	39.22	68.03	36.64	17.787
BMDAVSAD	50.12	67.04	35.29	40.213
PAD675	42.20	66.33	36.76	9.635
BMXBMI	45.10	65.60	38.58	12.046
MCQ300C2	65.31	49.10	48.34	22.479

k-NN

ROC curve variable importance

variables are sorted by maximum importance across the classes
only 20 most important variables shown (out of 54)

	X1	X2	X3	X4
LBXTC	0.8052	0.8052	0.8052	0.6943
RIDAGEYR	0.7643	0.8021	0.6604	0.8021
ALQ120QU	0.7755	0.7755	0.7755	0.5916
BPQ090D	0.7340	0.7532	0.6564	0.7532
LBDLDL	0.7502	0.7502	0.7502	0.6383
BPXSYave	0.6495	0.6767	0.7411	0.6767
LBXTR	0.7239	0.7239	0.7239	0.5826
BMXWAIST	0.5930	0.7083	0.5603	0.7083
MCQ300C	0.7000	0.6660	0.6180	0.7000
BMDAVSAD	0.5719	0.6989	0.5625	0.6989
WHQ150	0.6451	0.6976	0.6166	0.6976
MCQ365B	0.6766	0.6830	0.6140	0.6830
MCQ370B	0.6780	0.6780	0.6780	0.5404
BPQ020	0.6319	0.6766	0.6035	0.6766
BPQ080	0.6319	0.6766	0.5841	0.6766
BPXDIave	0.6683	0.6683	0.6683	0.5985
LBDHDD	0.5962	0.6418	0.5516	0.6418
MCQ365D	0.6340	0.6362	0.6393	0.6362
SMQ925A	0.6383	0.6383	0.6383	0.5298
PAD660	0.6127	0.6127	0.6363	0.5304

Robustness Test – Excluding diagnosed

```
> varImp(balancedB234_rf)
```

rf variable importance

variables are sorted by maximum importance across the classes
only 20 most important variables shown (out of 73)

	1	2	3
BPXSYave	100.00	53.49	52.40
LBXTR	85.35	62.16	82.81
ALQ120QU	82.99	58.62	27.99
BPXDIave	78.18	48.02	28.29
LBXTC	77.31	58.32	11.63
PAD680	76.37	54.46	31.07
PAQ710	76.22	54.57	10.08
LBDLDL	76.15	71.57	74.40
BMXARML	72.54	52.97	19.65
PAD675	72.13	47.91	11.38
RIDAGEYR	71.52	42.03	38.55
BMXLEG	70.48	46.89	19.39
LBDHDD	69.17	51.17	20.05
BMXHT	68.67	60.71	17.14
BMXWAIST	67.67	54.96	34.61
PAD630	66.54	43.65	19.91
SMQ925A2	66.53	36.23	9.69
MCQ370B2	65.40	45.48	25.01
BMXWT	65.39	49.08	27.81
BMXARMC	64.83	47.24	18.64

```
> varImp(knn_balB234, scale=FALSE)
```

ROC curve variable importance

variables are sorted by maximum importance across the classes
only 20 most important variables shown (out of 56)

	X1	X2	X3
LBXTR	0.8242	0.6678	0.8242
BPXSYave	0.8219	0.7522	0.8219
RIDAGEYR	0.7403	0.6559	0.7403
BMDAVSAD	0.6916	0.6234	0.6916
MCQ370B	0.6898	0.6898	0.6826
ALQ120QU	0.6887	0.6887	0.6631
BPQ090D	0.6820	0.6422	0.6820
LBDLDL	0.6772	0.6056	0.6772
MCQ365D	0.6658	0.6549	0.6658
BMXWAIST	0.6653	0.6217	0.6653
PAD680	0.6593	0.5807	0.6593
WHQ150	0.6461	0.5891	0.6461
LBDHDD	0.6310	0.5720	0.6310
PAD660	0.6299	0.6299	0.5958
SMQ925A	0.6250	0.6105	0.6250
SMQ020	0.6150	0.6077	0.6150
BMXHT	0.6149	0.6149	0.5770
DMDEDUC2	0.6143	0.6143	0.6091
LBXTC	0.6109	0.6109	0.6100
MCQ300C	0.6077	0.6077	0.5562