Diabetes prediction with machine learning methods

Siwei Wu

Health data analytics and machine learning

Imperial College London

Background

Increasing DM makes a prediction model necessary; some prediction models have been developed.

High incidence of diabetes makes a prediction model in need:

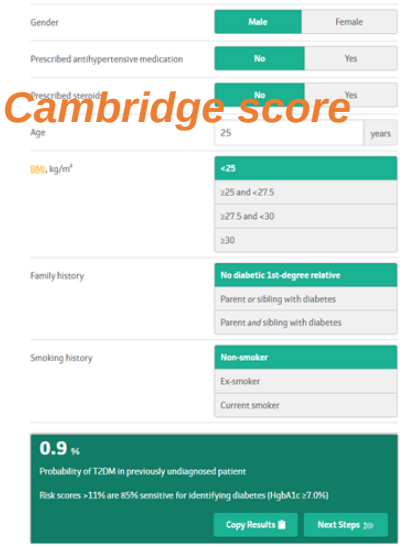
- Target people at higher risk
- Intervene their lifestyle to prevent or postpone the onset of diabetes

A variety of (logistic model based) risk scores but with drawbacks:

- **Variable selection sensitive to statistic power**
- Not robust to **collinearity**
- Can't model **non-linearity**

A new model: can machine learning help us?

FOS equation



Cambridge score

FOS T2DM risk score is based on logistic regression with a sparse set of predictors; Adding more predictors doesn't improve performance

	Model 1	Model 2-1	Model 2-2	Model 2-3	Model 2-4	Model 3-1	Model 3-2	Model 3-3	Model 4
AUC	**0.724**	**0.852**	**0.850**	**0.852**	**0.881**	**0.854**	**0.850**	**0.851**	**0.869**
Age									
Sex									
Parental history									
BMI									
Waist circumstance									
Blood pressure									
HDL-C									
Triglyceride									
Fasting glucose									
2-hour OGTT									
Fasting insulin									
C-reactive protein									
Gutt insulin sensitivity level									
HOMA insulin resistance index									
HOMA β-cell index									
Hormone therapy, smoking, alcohol, drug use, HbA1C									

Variables included as discrete

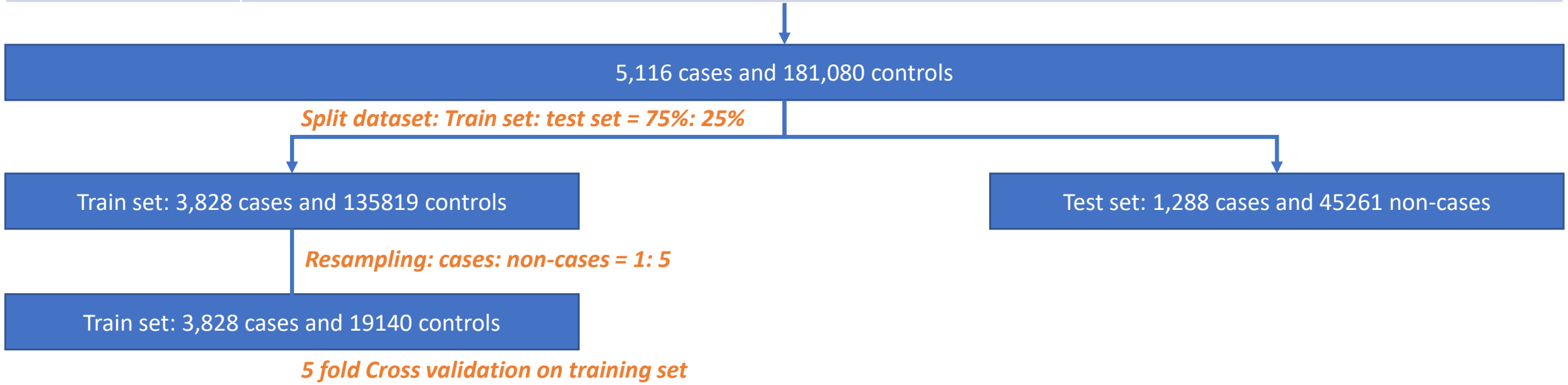
Variables included as continuous

Variables not included

METHODS

Data description: UK BIOBANK

Data pre-processing	
Inclusion and exclusion:	• Non-white people are excluded • Prevalent cases are excluded • People with glycated haemoglobin (HbA1c) > 42mmol/mol are excluded • People with fasting glucose > 7 mmol/L are excluded
Quality control:	• Only complete ones are included
Recoding:	• Alcohol frequency is recoded as an ordinal variable from a categorical variable("Never" -> 1, "Special occasions only"-> 2, "Once or twice a week"-> 3, "One to three times a month"-> 4, "Three or four times a week"-> 5, "Daily or almost daily"-> 6) • Smoking is recoded as an ordinal variable from a categorical variable ("No" -> 1, "Only occasionally" ->2, "Yes, on most or all days" -> 3)

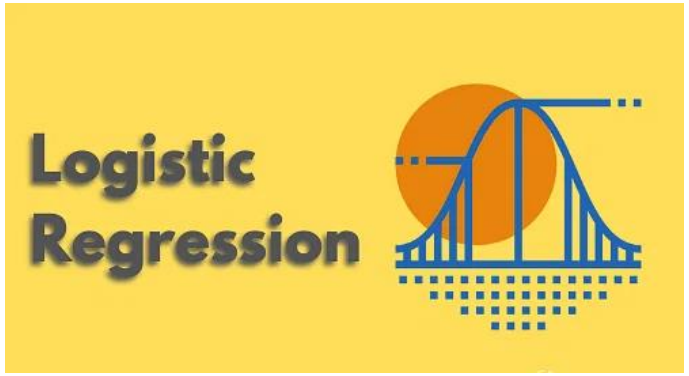


Data description: 45 predictors are included; more risk factors and biomarkers are included compared with FOS T2DM risk score

Category	Variable	Count
Demographic information	**Sex**, **Age**	2
Physical Measurement	**SBP**, **DBP**, **waist**, **BMI**	4
Life style	Physical activity, Smoking, Alcohol	3
Family history	**Family history of mother, Family history of father**	2
Co-mobidity	Hypertension, Dyslipidemia	2
Blood biochemical metrics (Sex hormone)	testosterone, shbg	2
Blood biochemical metrics (Glucose metabolism)	**Glucose**, HaB1C, IGF1	3
Blood biochemical metrics (Liver function)	Alkaline phosphate, Alanine aminotransferase(ALT), Aspartate aminotransferase(AST), Direct bilirubin, Gamma glutamyltransferase (GGT), Total bilirubin	6
Blood biochemical metrics (Inflammation)	C-Reactive protein	1
Blood biochemical metrics (Urate)	Urate	1
Blood biochemical metrics (Lipid metabolism)	Apolipoprotein A, Bio apolipoprotein B, Cholesterol, **HDL cholesterol**, ldl direct, Lipoprotein A, **Triglycerides**	7
Blood biochemical metrics (Nutrition)	Albumin, Total protein	2
Blood biochemical metrics (Skeleton mechanism)	Calcium, Phosphate, Vitamin D	3
Blood biochemical metrics (Renal function)	Urea, creatinine, Cystatin C	3
Urine biochemical metrics (Renal function)	Microalbumin, Creatine, Potassium, Sodium	4
Sum		45

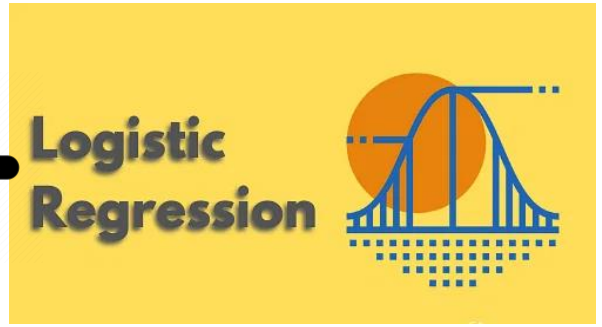
The FOS score is replicated as the baseline to compare with; 4 new models are trained to improve the prediction for DM;

Baseline model: replication of FOS

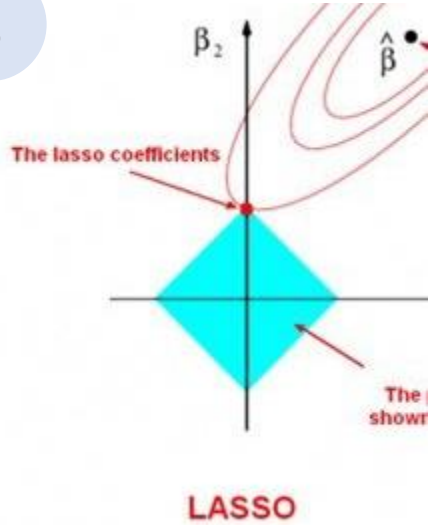


1

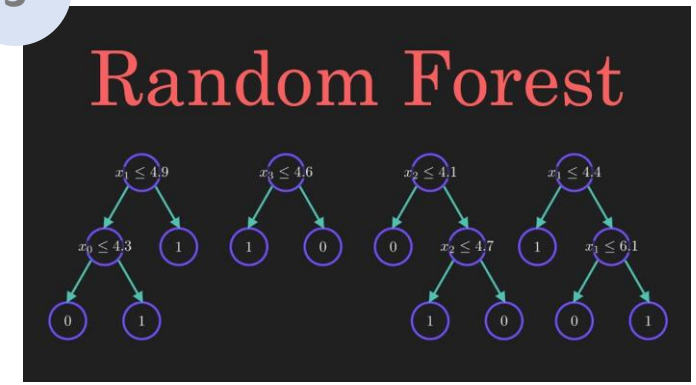
+



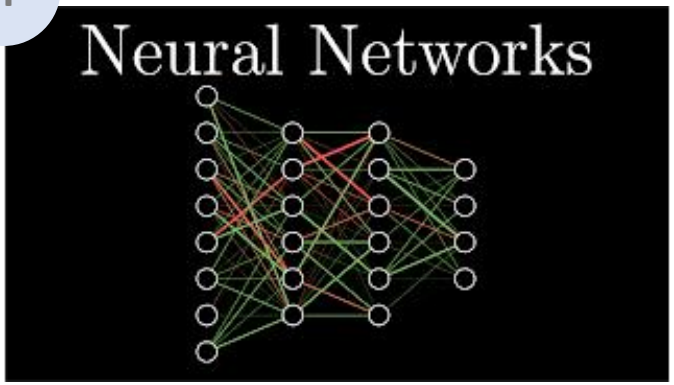
2



3



4



Model tuning: hyper-parameters of clustering is tuned based on silhouette score; hyper-parameter of random forest is pre-tuned on a decision tree

Methods	Data required	Hyper-parameter	Hyper-parameter tuning
Logistic	Only numeric data scaled	/	/
Logistic (on) Clustering	Only numeric data scaled	/	/
	Scaled data	• Number of clusters	• Pick number of clusters based on silhouette score (closest to 1) ;
Lasso	Scaled data (one hot categorical variable)	• Lambda	• Cross-validation; • Pick lambda.1se for prediction;
Random forest	Un-scaled data	• Impurity criterion; • Max depth; • Min samples each node; • Number of estimators;	• Cross-validation • Did **pilot tuning on a decision tree;** • **Narrow down the range of hyper-parameters tuning a forest model**
ANN	Scaled data	• Depth(layer) of network	• Cross-validation • Tuned manually

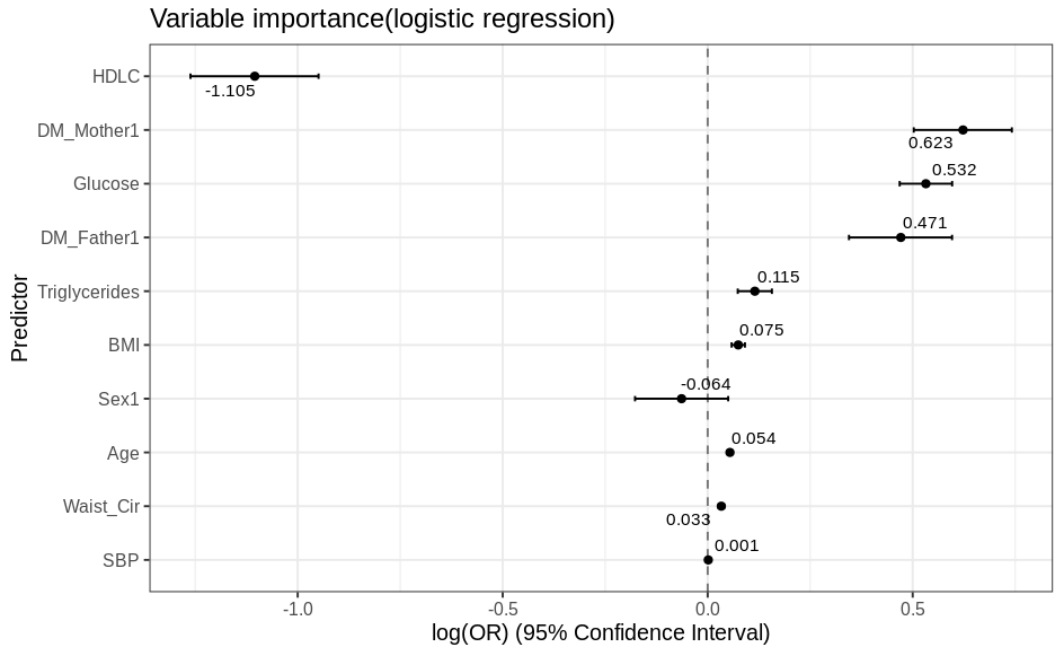
RESULT

Model 1: Logistic regression(replication of FOS): HDL-C is the most importance predictor; AUC of FOS diabetes score is 0.79

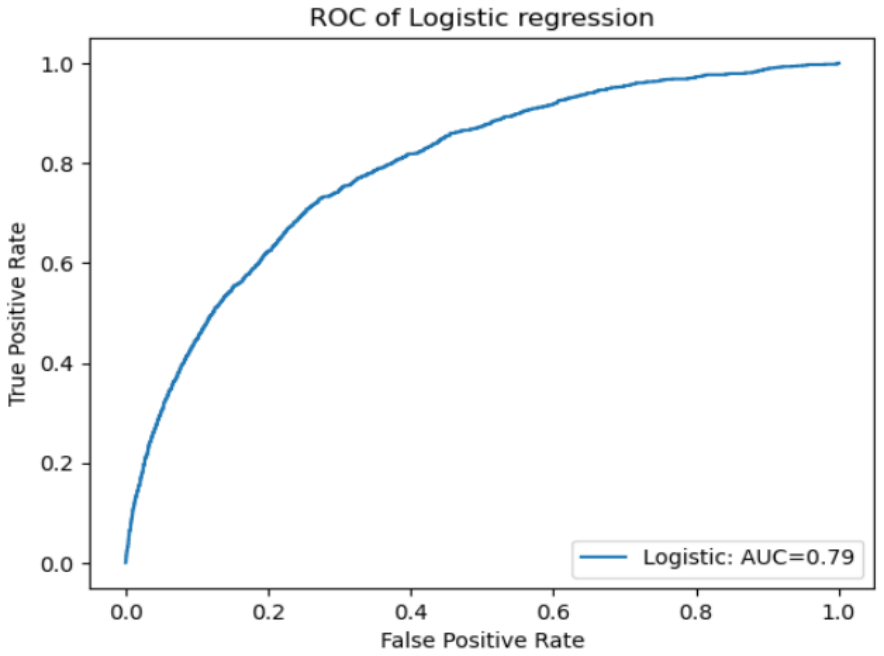
Regression model:

*Logit(Probability of DM) = – 1.105*HDLC + 0.623*Mother with DM + 0.532*Glu + 0.471*Father with DM + 0.115*Cholesterol + 0.075*BMI - 0.064*male + 0.054*age + 0.033*waist circumstance + 0.001*SBP*

Variable importance: the most important predictor is HDLC with the largest coefficient, follow by family history and glucose

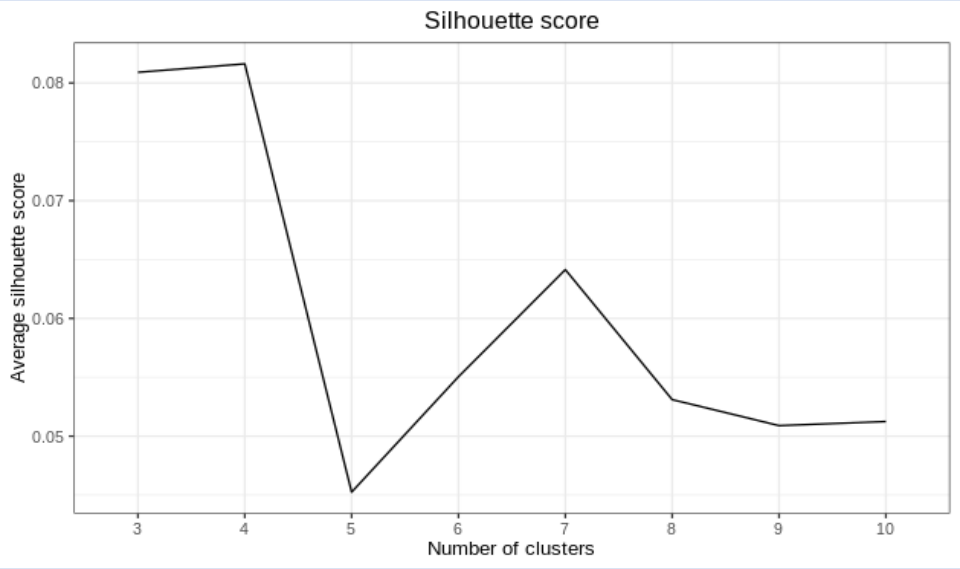


The AUC of FOS diabetes score is 0.79



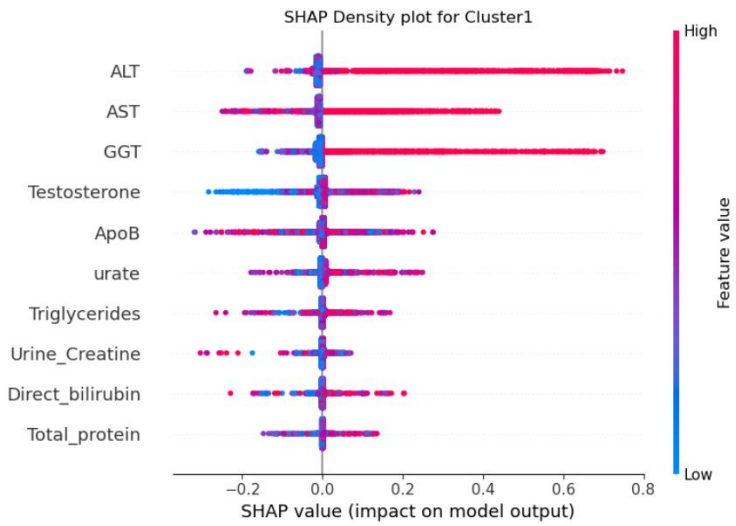
Model 2: Logistic regression on clustering: k-means clustering on 29 biomarkers; 4 clusters can result in the silhouette score closest to 1

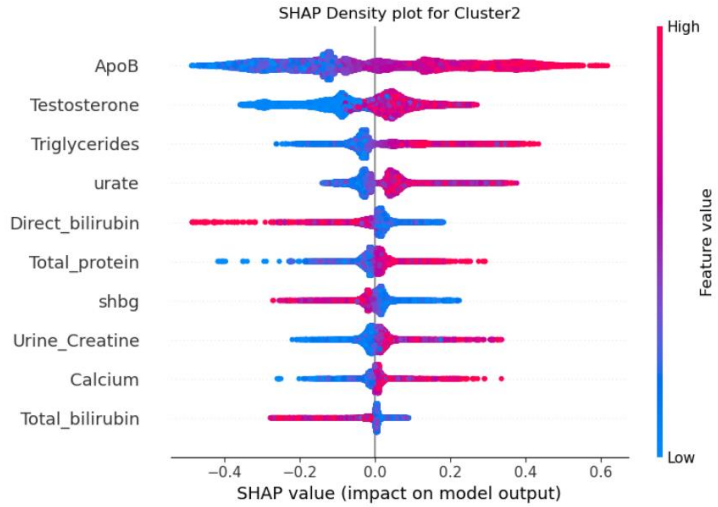
Motivation:	**Logistic regression struggles with too many predictors(overfitting) But these predictors are also informative and we want to include them**
Clustering:	**KMeans Clustering on the 29 biomarkers (out of 32) available in UK Biobank; Number of clusters is decided based on average silhouette score;**
Hyper-parameter tuning:	**Average silhouette score gets closest to 1 when number of clusters is four**



Model 2(Logistic regression on clusters): Clustering is interpreted by a decision tree and using SHAP value.

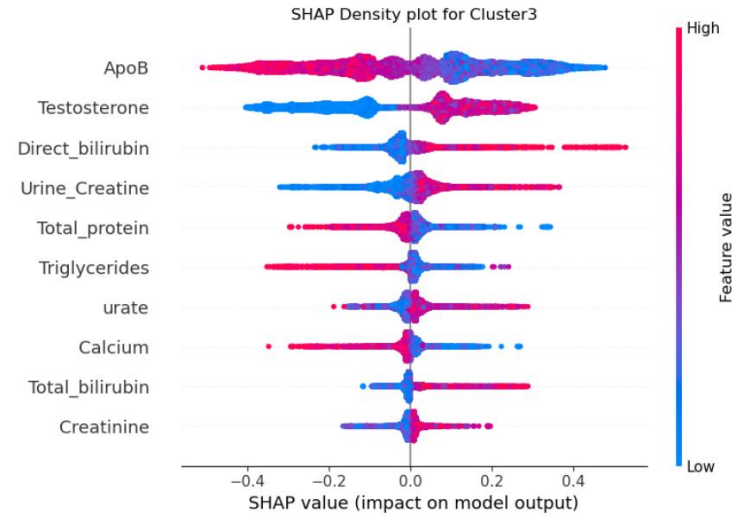
Cluster 1:
Higher ALT
Higher AST
Higher GGT

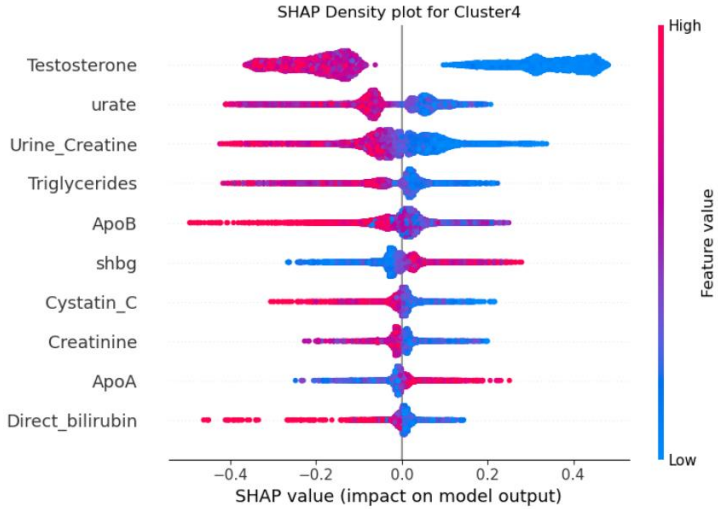




Cluster 2:
High ApoB
High Testosterone
High Triglycerides

Cluster 3:
Low ApoB
High Testosterone
Higher bilirubin
Lower triglyceride





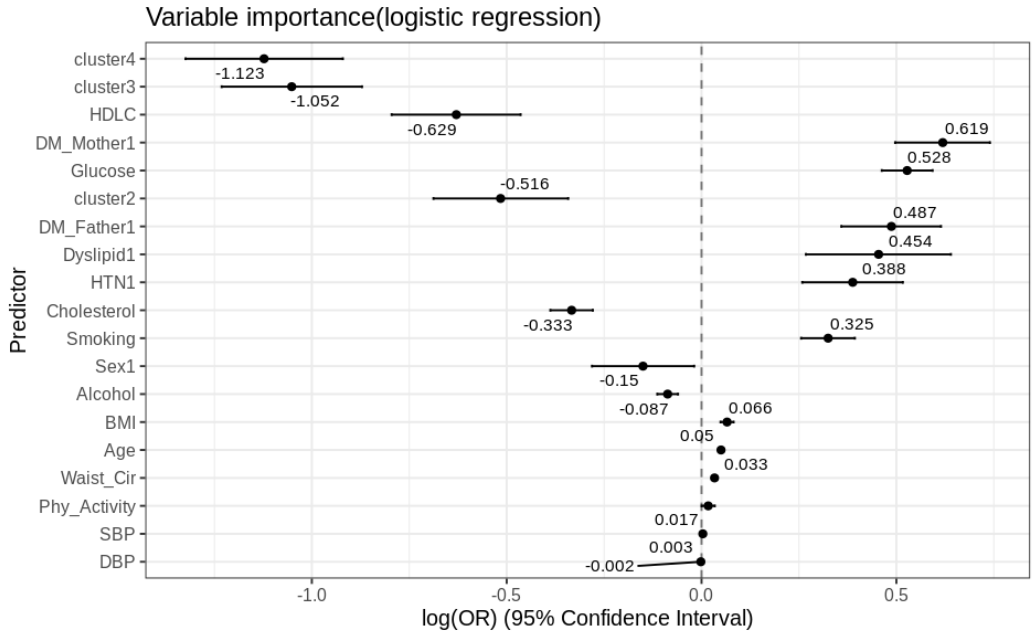
Cluster 4
Low Testosterone
Low urate
Low uric creatine

Model 2(Logistic regression on clusters): the clusters on biomarkers play an important role; AUC improves to 0.80

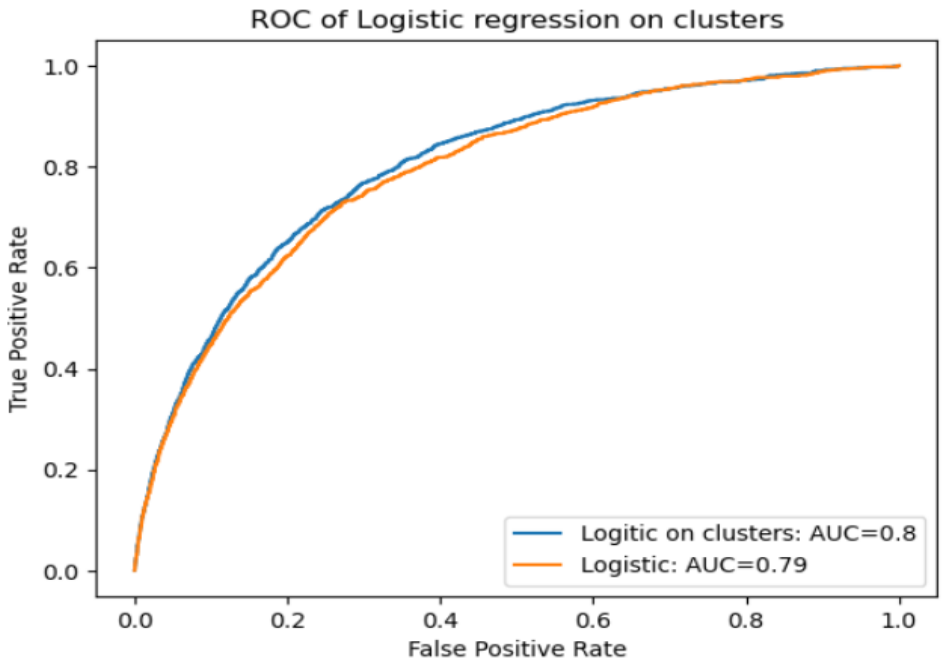
Regression model:

Logit(Probability of DM) = -1.12*cluster4 - 1.05*cluster3 -0.629*HDLC + + 0.619*Mother with DM + 0.528*Glu + -0.516*cluster2 + 0.487*Father with DM +0.454*dyslipidemia + 0.388*HTN + ...

Clusters play an important role in prediction; those in cluster 4 and cluster 3 are at a higher risk of diabetes

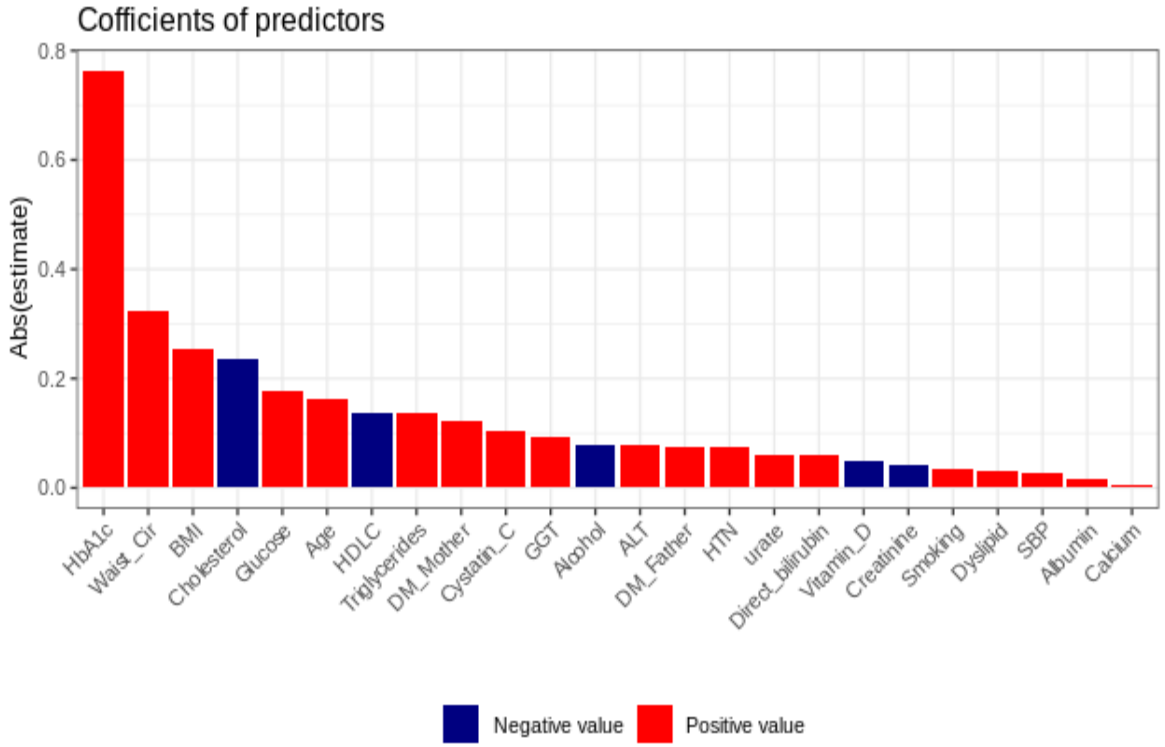


The AUC improves to 0.80 compared to 0.79 of FOS diabetes score

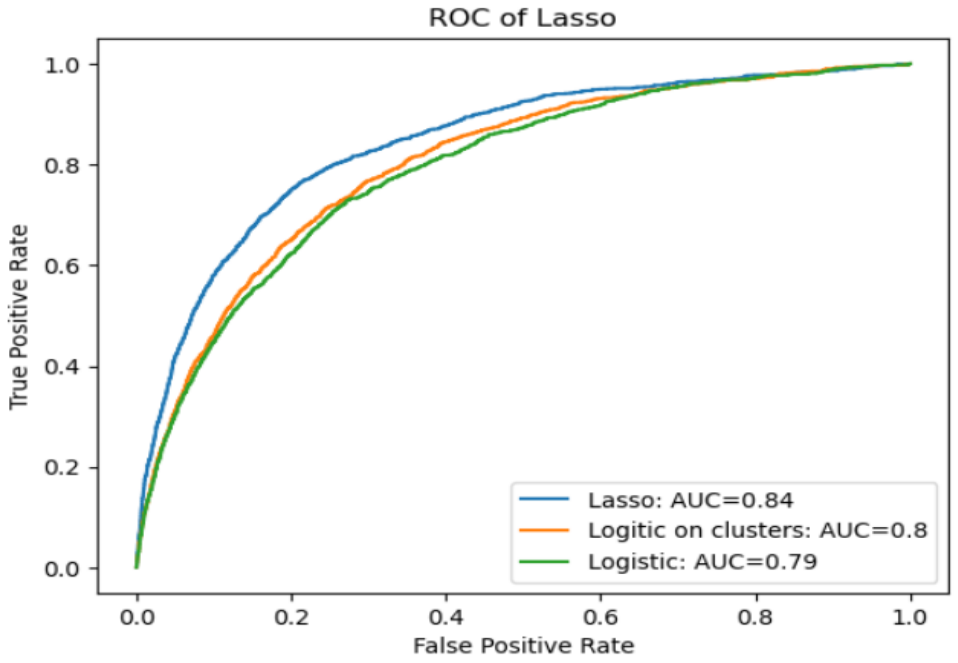


Model 3(LASSO): 23 variables are selected by Lasso; HbA1c, waist circumstance and BMI are the three with largest effect size; ROC increases to 0.84

- **23 variables are selected by Lasso;**
- **Higher HbA1c, waist circumstance and BMI mostly increase the risk of diabetes**
- **Lambda is set as Lambda.1se=0.059 (See Appendix)**



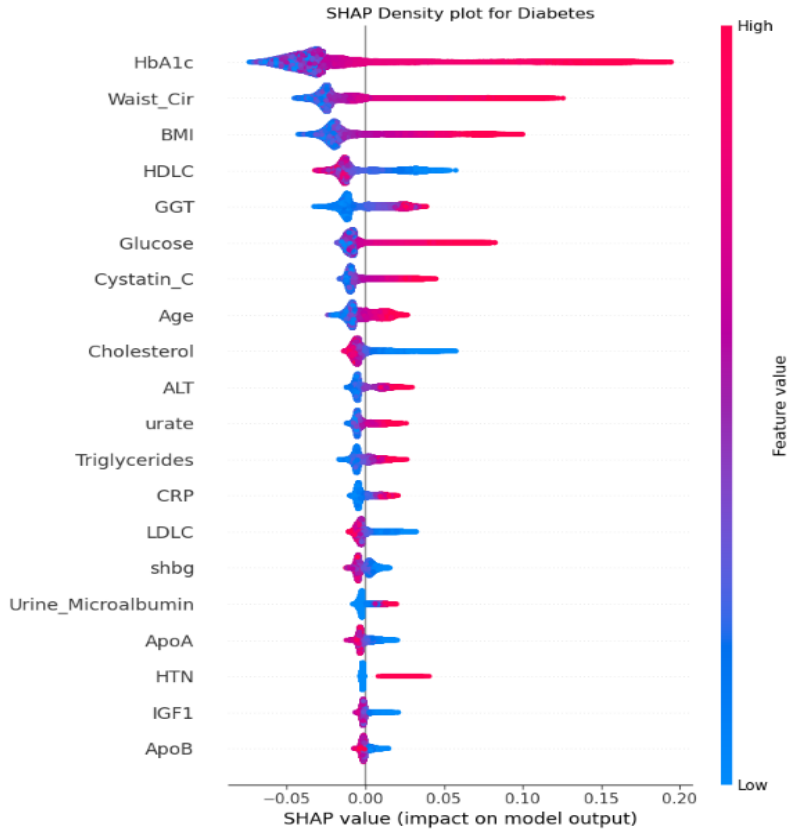
- **A significant improvement in AUC**
- **Less information is lost with lasso selecting all variables important to prediction**

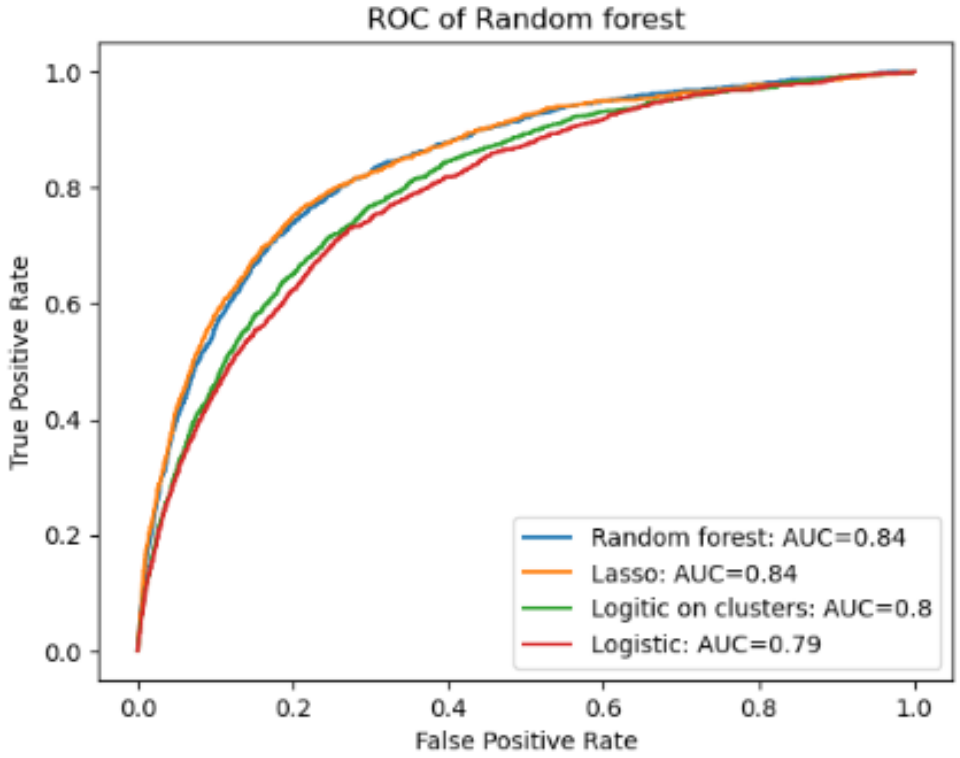


Model 4(Random forest): higher HbA1c, waist circumstance, and BMI are most predictive of diabetes; AUC of random forest is 0.84

- **Hyper-parameters:** criterion=Gini; depth=14; min samples per leaf = 64; num estimators=200
- HbA1c, waist circumstance, and BMI are most predictive of diabetes

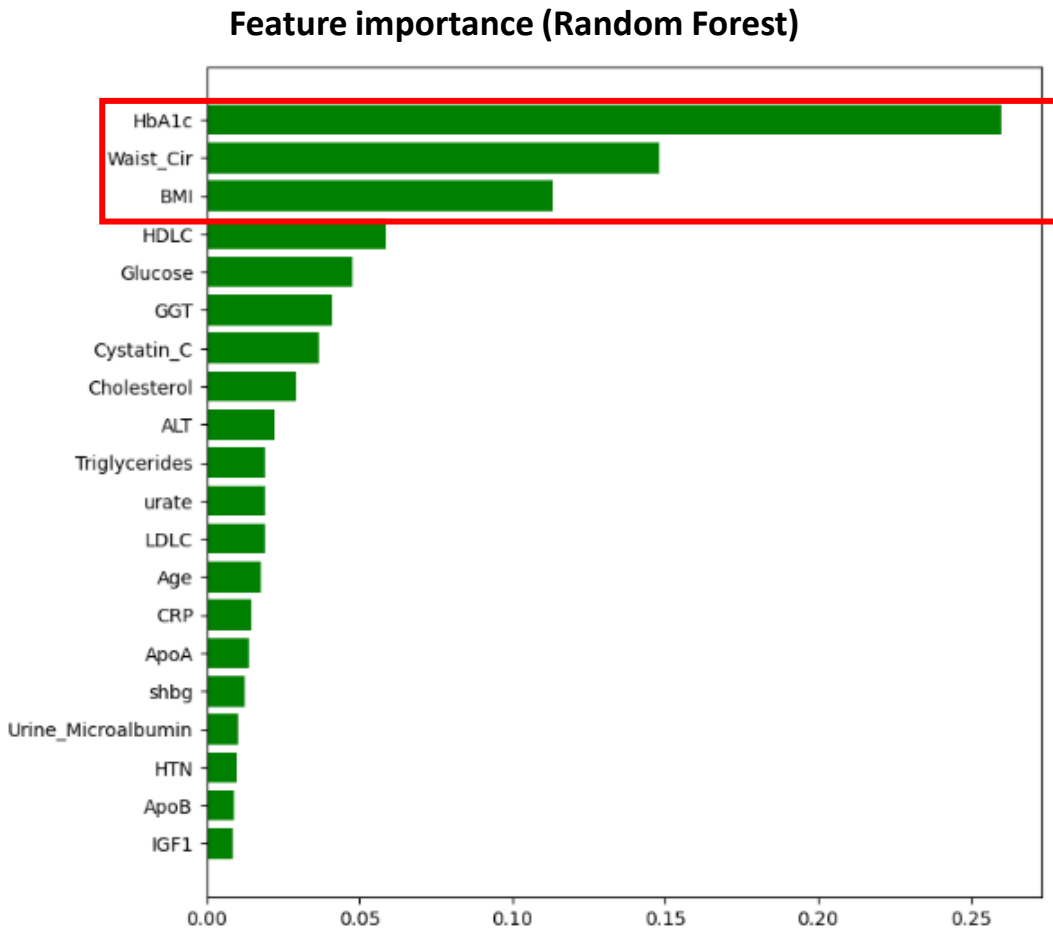
- AUC of random forest = 0.84, equivalent to LASSO

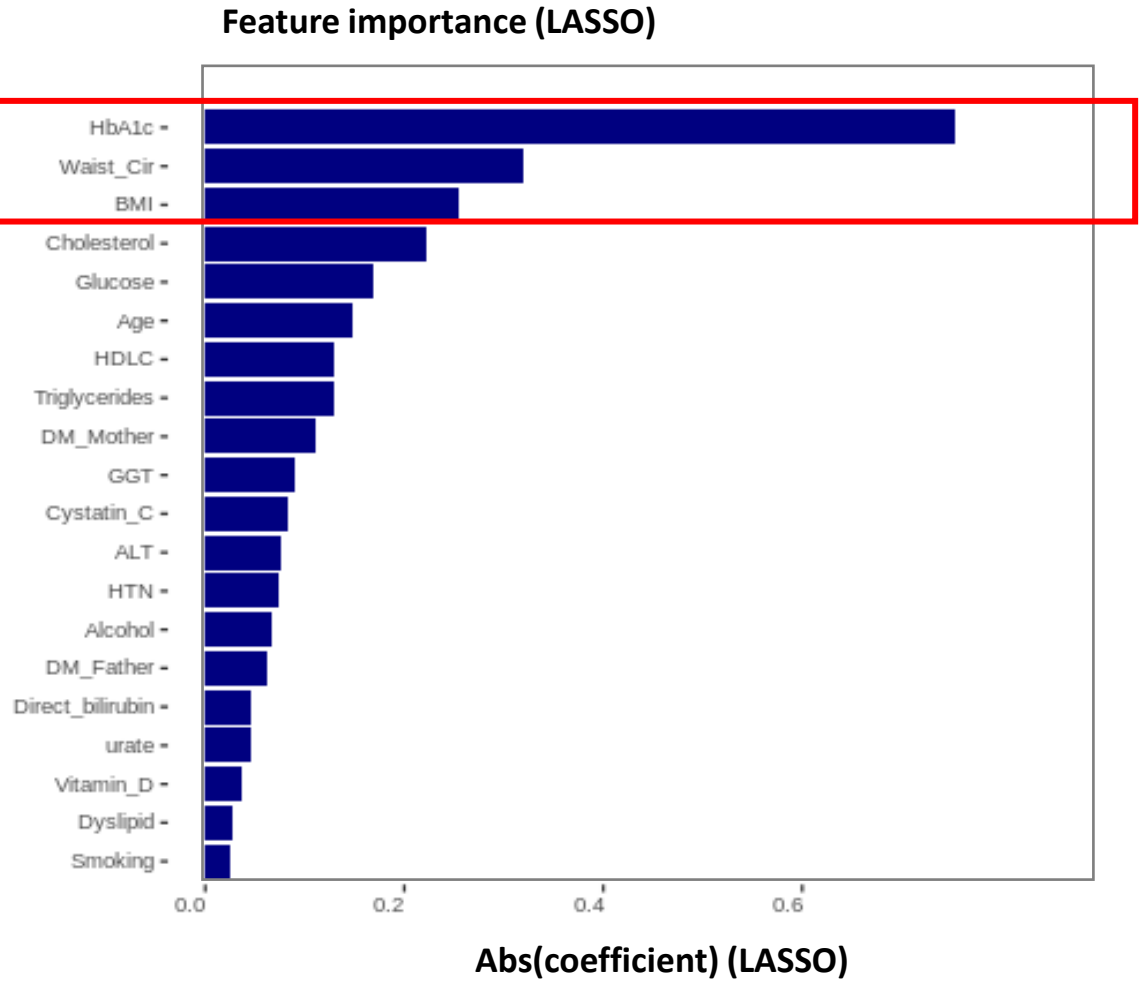




A comparison of variable importance between random forest and lasso: important features are similar across two models

- HbA1c, waist circumstance, and BMI are top predictors for both random forest and lasso

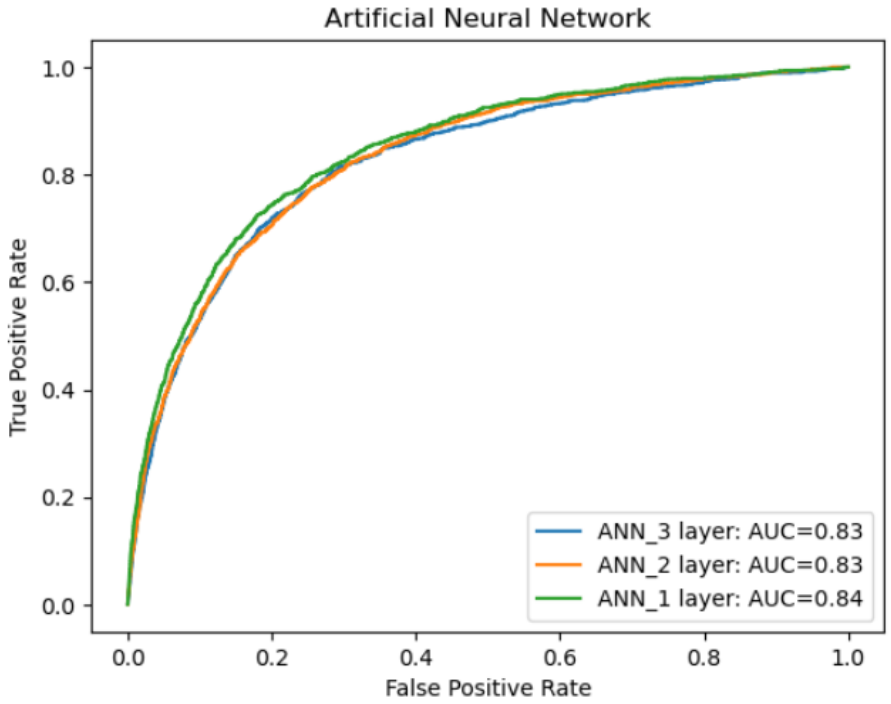


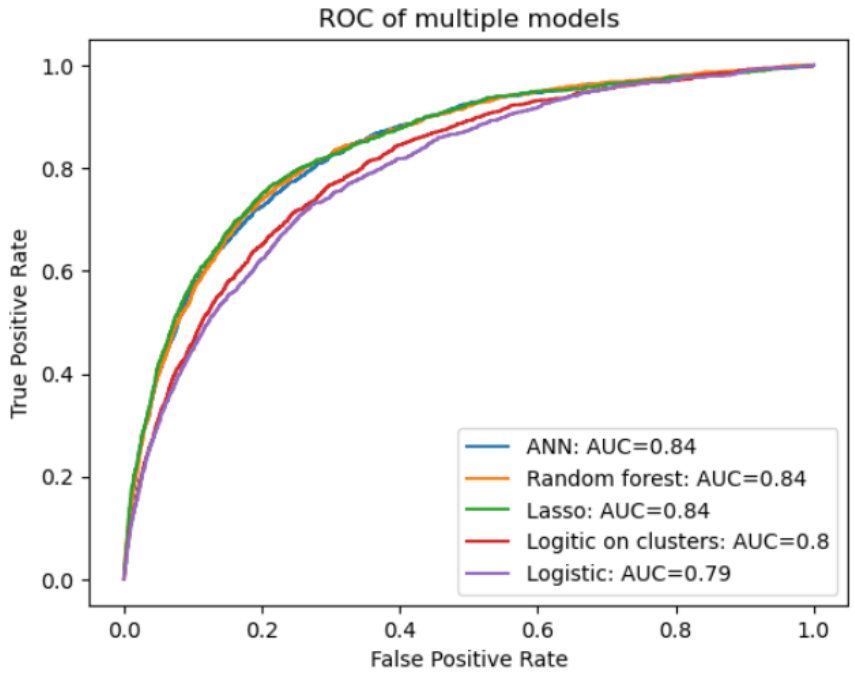


ANN with only 1 hidden layer outperforms those with multiple hidden layers; It doesn't outperform random forest or lasso

- Hyper-parameters are tuned manually with cross-validation
- Network with only 1 hidden layer outperforms those with more hidden layers

- The AUC is 0.84
- Neural network doesn't outperform random forest or lasso





CONCLUSION

Conclusion

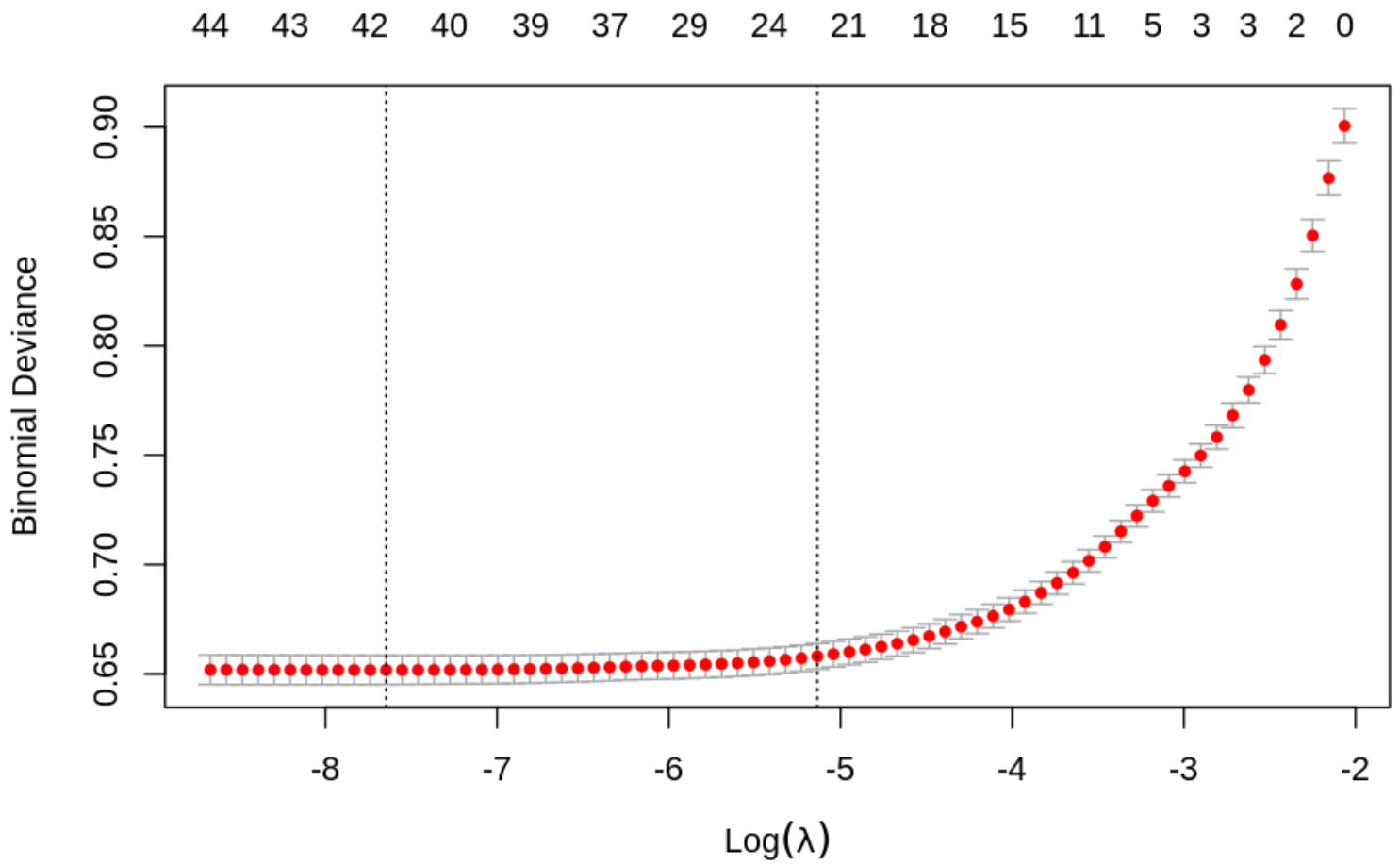
- Replicated the Framingham equation as the baseline.
- Built four new models: logistic regression on clusters, LASSO, random forest, and ANN.
- All the four models outperform the Framingham equation.
- LASSO, random forest, and ANN performs equally with the same AUC(0.84)
- LASSO and random forest are preferred for sake of interpretability.

Limitation & future work

- More observations can be included if I imputed them.
- More predictors can be included to our model like polygenetic risk score for type 2 diabetes, which can be helpful to our prediction.
- Significance test on AUC can be performed to get a confident conclusion whether new models are significantly better than the FOS
- The ANN is not well tuned. It may outperform other methods if the hyper-parameters are well calibrated.

QUESTIONS?

Appendix: lambda.1se for lasso is 0.059 with cross validation



Appendix: error metrics for prediction model

	accuracy	F1 score	Precision	Recall
Logistic	0.95	0.19	0.18	0.20
Logistic + clustering	0.95	0.20	0.18	0.23
LASSO	0.92	0.25	0.17	0.48
Random forest	0.96	0.21	0.26	0.18
ANN	0.92	0.24	0.16	0.49