Identifying risk factors and predicting incident cases of Diabetes

(a cohort study of adults in China)

Group 5
Anil Anderson
Arthi Hariharan
David Hu
Álvaro Quijano
Yumei Yang

Motivation

The total number of people living with diabetes is projected to rise to **643 million** by 2030 and **783 million** by 2045 (International Diabetes Federation, 2021).

Early diagnosis plays a crucial role in mitigating the effects of diabetes. Consequently, predictive modeling emerges as an indispensable tool, facilitating timely interventions and efficient allocation of resources.

According to Chen et al (2018), the question that remains unclear is whether there is an association between diabetes and **body mass index (BMI)**, and how this might be impacted by **age**

Objective To identify the risk factors for predicting Type 2 diabetes and assess the importance of BMI and age in diabetes using data from a cohort study of Chinese adults.

Data

Corresponds to a **population-based cohort** study of 211,833 participants (116,123 male and 95,710 female) in China.

The features included,

Characteristics, age, gender, family history, weight and height
Physiological metrics, systolic blood pressure (SBP), diastolic blood pressure (DBP)
Lipid profiles, cholesterol, triglyceride, high-density lipoprotein (HDL), and low-density lipoprotein (LDL)
Liver and Kidney function, alanine aminotransferase (ALT), aspartate aminotransferase (AST), blood urea nitrogen (BUN) and creatinine clearance rate (CCR)

Response variable, Type 2 Diabetes Mellitus (1/0)

Descriptive

	Diabetes (N = 4,147) (%/Stdev)	Non-diabetes (N = 207,659) (%/Stdev)
Age	54.7 (13.2)	41.8 (12.5)
Gender Female	1,174 (28.1%)	94,536 (45.5%)
Height (cm)	167 (8.45)	166 (8.32)
Weight (kg)	73.2 (13.1)	64.5 (12.1)
Family History	171 (4.1%)	4173 (2.0%)

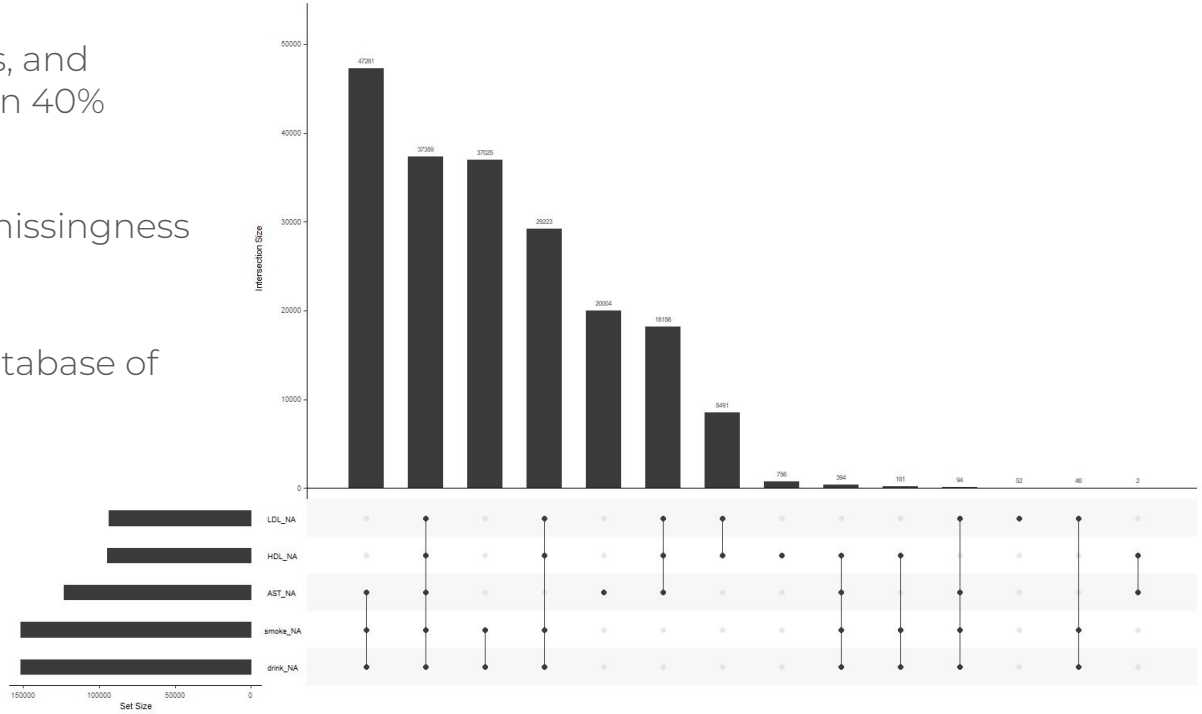
Missing data

LDL, HDL, AST, smoking status, and drinking status have more than 40% missingness

LDL-HDL, smoking-drinking missingness occur simultaneously

Data were extracted from a database of medical records

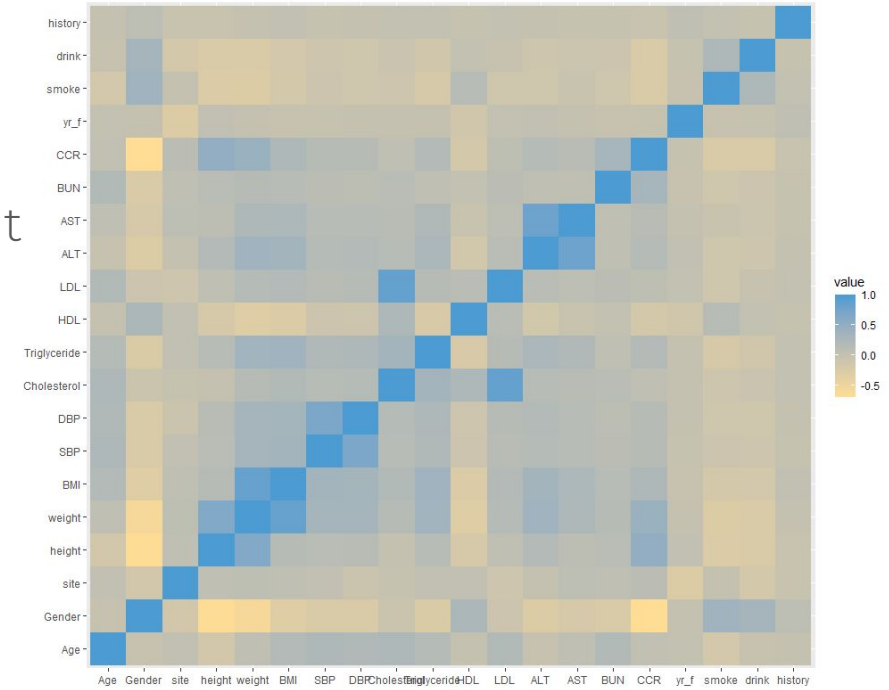
MCAR? - single imputation



Correlation

Highly correlated variables

- weight and height/BMI
- gender and CRR/height/weight
- cholesterol and LDL
- AST and ALT



Model selection

Removed HDL and LDL
- Correlated with cholesterol
- High degree of missingness

Removed weight: correlated with BMI and height

Data split: 80:20

Methods

Log-Likelihood based models

Penalized Logistic Regression (Ridge)
Bayesian Logistic Regression

Machine learning approach

Random Forest

Logistic Regression Likelihood

Logistic model:

$$\log\left(\frac{p_i}{1-p_i}\right) = x_i'\beta,$$

Likelihood:

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} = \prod_{i=1}^{n}\left(\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right)^{y_i}\left(\frac{1}{1+e^{x_i'\beta}}\right)^{1-y_i}$$

Log-likelihood:

$$l(\beta) = \sum_{i=1}^{n} y_i x_i'\beta - \log\left(1+e^{x_i'\beta}\right)$$

Penalized Logistic Regression (Ridge)

Penalized Likelihood function,

$$L^* = l(\beta) - \frac{\lambda}{2}\sum_k \beta_k^2$$

It lead us to an iterative solution as follows (Göksülük, D.,2011),

$$\hat{\boldsymbol{\beta}}^{t+1} = (\mathbf{X}'\mathbf{W}\mathbf{X} + \boldsymbol{\Lambda})^{-1}\mathbf{X}'\mathbf{W}\left\{\mathbf{X}\hat{\boldsymbol{\beta}}^{t} + \mathbf{W}^{-1}(\mathbf{Y} - \hat{\boldsymbol{\pi}})\right\}$$

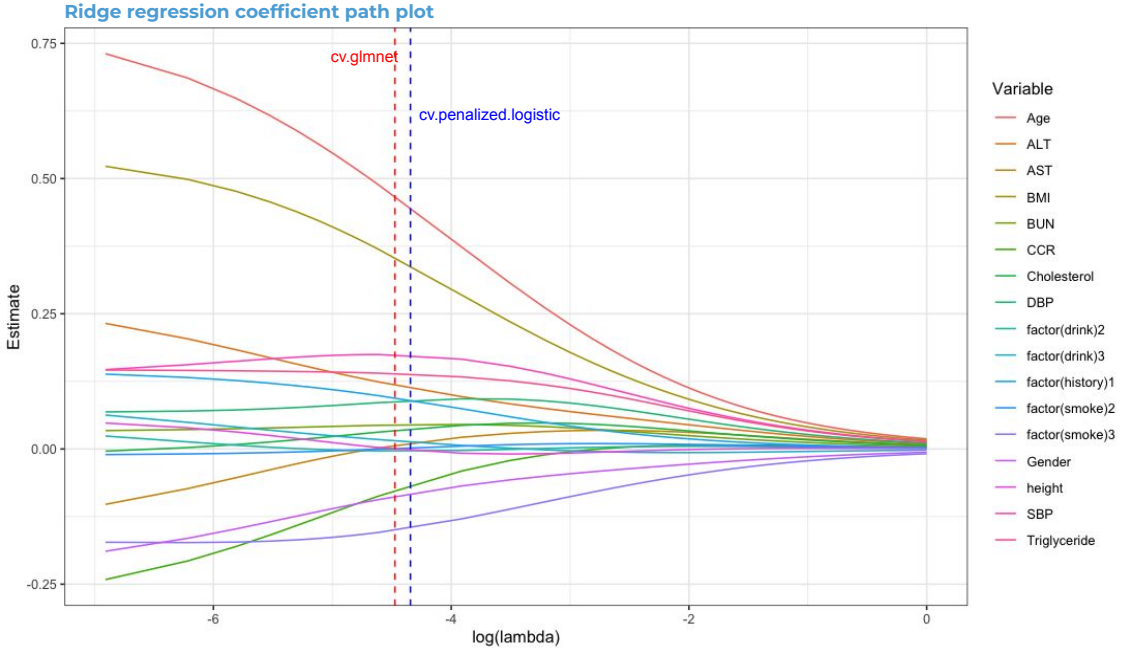
Where,

$$\mathbf{W} = \mathbf{diag}(\pi_i(1-\pi_i))$$

$$\boldsymbol{\Lambda} = \mathbf{diag}(\lambda)$$

Note that for $\lambda = 0$ it corresponds to the NR solution for GLM

Penalized Logistic Regression (Ridge)

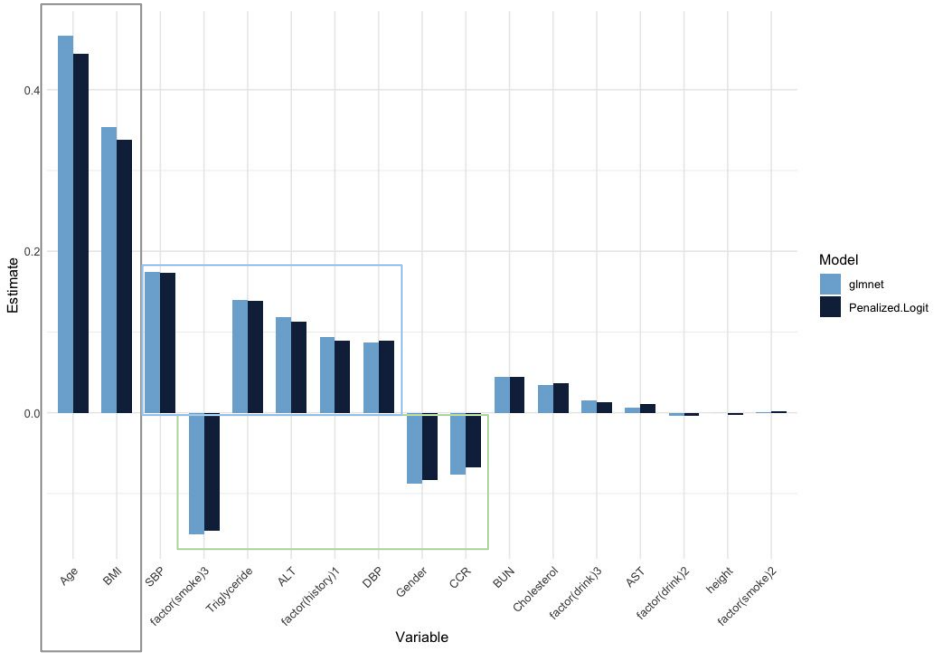


Using 5-fold cross-validation to minimize the misclassification rate.

λ = 0.013

λ (glmnet package)= 0.0114

Penalized Logistic Regression



- Body Mass Index (MI) and age are the most important features (risk factors) associated to diabetes.
    - *Chen et al (2018) demonstrated that BMI and age are associated to incident diabetes.*
- High Systolic and Diastolic Pressure (SBP/DBP), Triglycerides, ALA, family history are also positively associated.
- Never having smoked, gender, and high creatinine levels (CCR) are protective factors against diabetes.
- Ridge penalization does not shrink coefficients to zero

Penalized Logistic Regression

- Sensitivity: 0.71
- Specificity: 0.85

Predicted/Actual	No	Yes
No	29393	130
Yes	12114	729

A Bayesian Approach to Logistic Regression

Prior in hierarchical form:

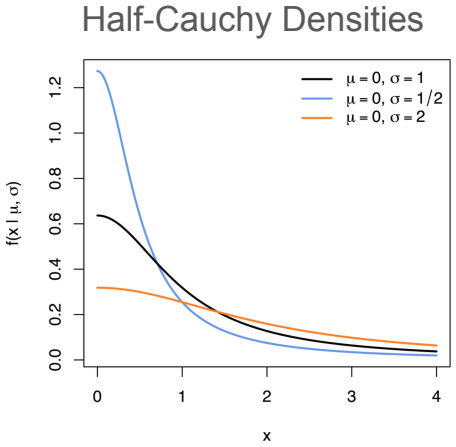
$$\begin{aligned}\beta_0 &\sim N(0, 10^2)\\ \beta_j|\lambda &\sim N(0, \lambda^2) \text{ for } j = 2, \ldots, p\\ \lambda &\sim \text{Half-Cauchy}(0, 1)\end{aligned}$$

Prior as an equation:

$$\pi(\beta, \lambda) = \frac{1}{\sqrt{2\pi(10)^2}} e^{-\beta_0^2/2(10)^2} \prod_{j=2}^{p} \frac{1}{\sqrt{2\pi\lambda^2}} e^{-\beta_j^2/2\lambda^2} \frac{2}{\pi(1+\lambda^2)}$$

Posterior:

$$p(\beta, \lambda|y) \propto L(\beta)\pi(\beta, \lambda)$$



Adaptive Metropolis Algorithm

- Initialize all coefficients at 0 and lambda at 1
- Propose a new set of coefficients and a new lambda by drawing from

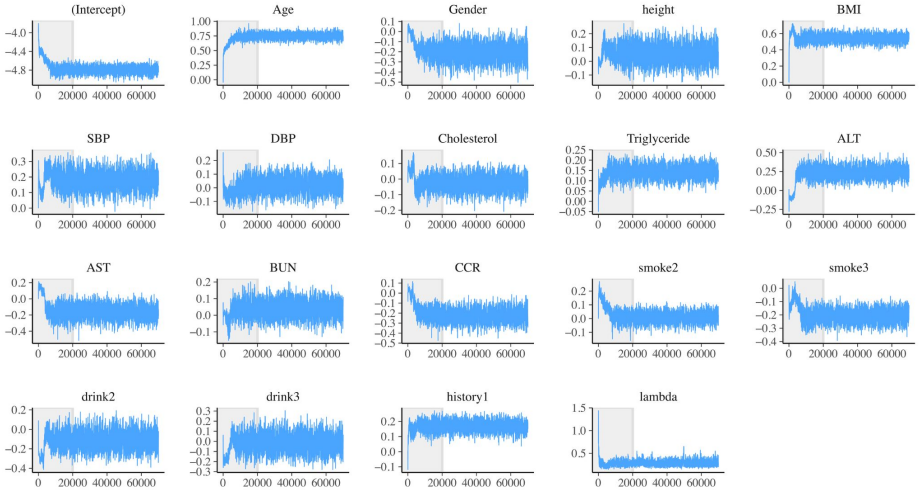
$$(\beta, \lambda)_{(i)}^{\text{proposal}} \sim N\left((\beta, \lambda)_{(i)}^{\text{current}}, \Sigma_{(i)}^{*}\right)$$

$$\Sigma_{(i)}^{*} = \frac{2.4^2}{d}\left(\hat{\Sigma}_{(i)} + \varepsilon I_d\right)$$

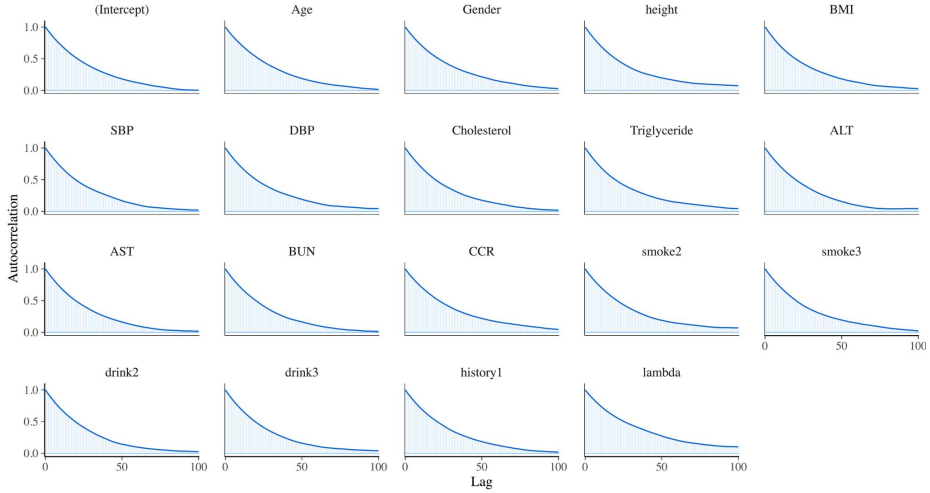
- Accept proposal values with probability min(1, R), where R is the ratio of the unnormalized posterior at the proposal parameters and at the current parameters
- 70,000 posterior draws
- Downsample training set to 30,000 rows

MCMC Diagnostics

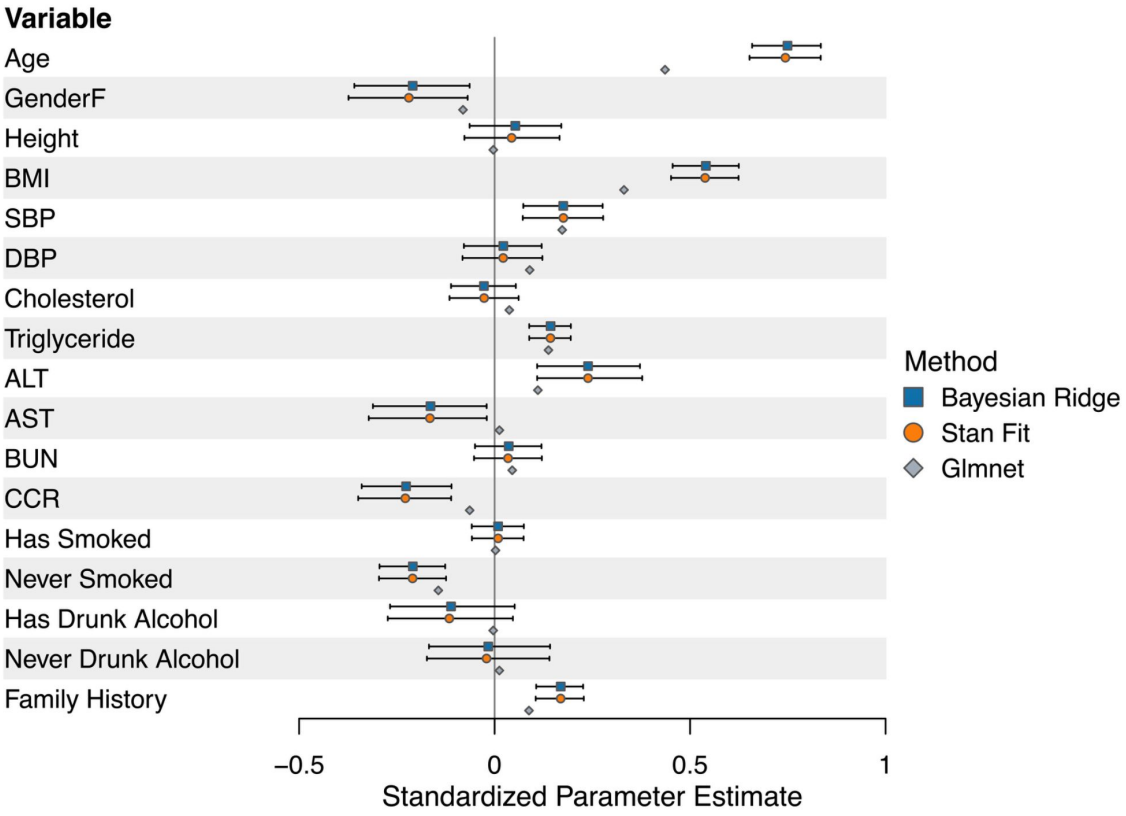
Trace plots:



ACF plots after burn-in:



Posterior Means and Credible Intervals



Takeaways:

- Consistency between our fit and Stan's fit
- Glmnet (Ridge) shrinks coefficients to 0 more strongly
- Age and BMI are coefficients with greatest magnitude

Random Forest (Overview)

- Ensemble of decision trees
- Build Process
  - Bootstrapping observations
  - Random subset of features
  - (Optional)Hyperparameter tuning
- Prediction Process
  - Each tree contributes a probability
- Advantages
  - Model diversity
  - Robust against overfitting

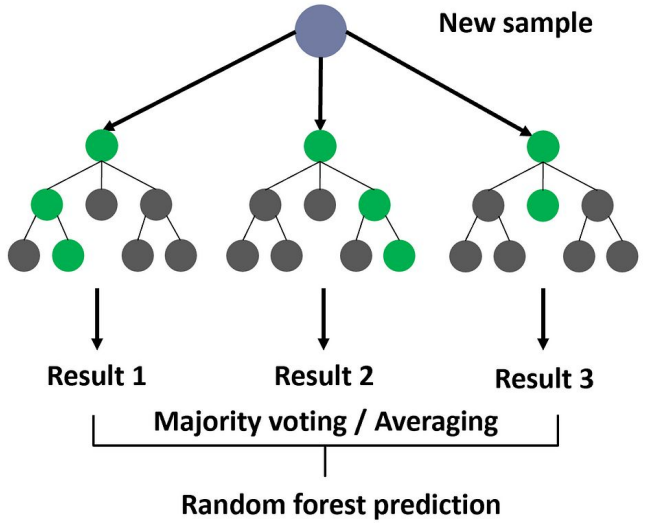


Image Credit: Dr. Yehoshua, Medium

Random Forest (Procedure)

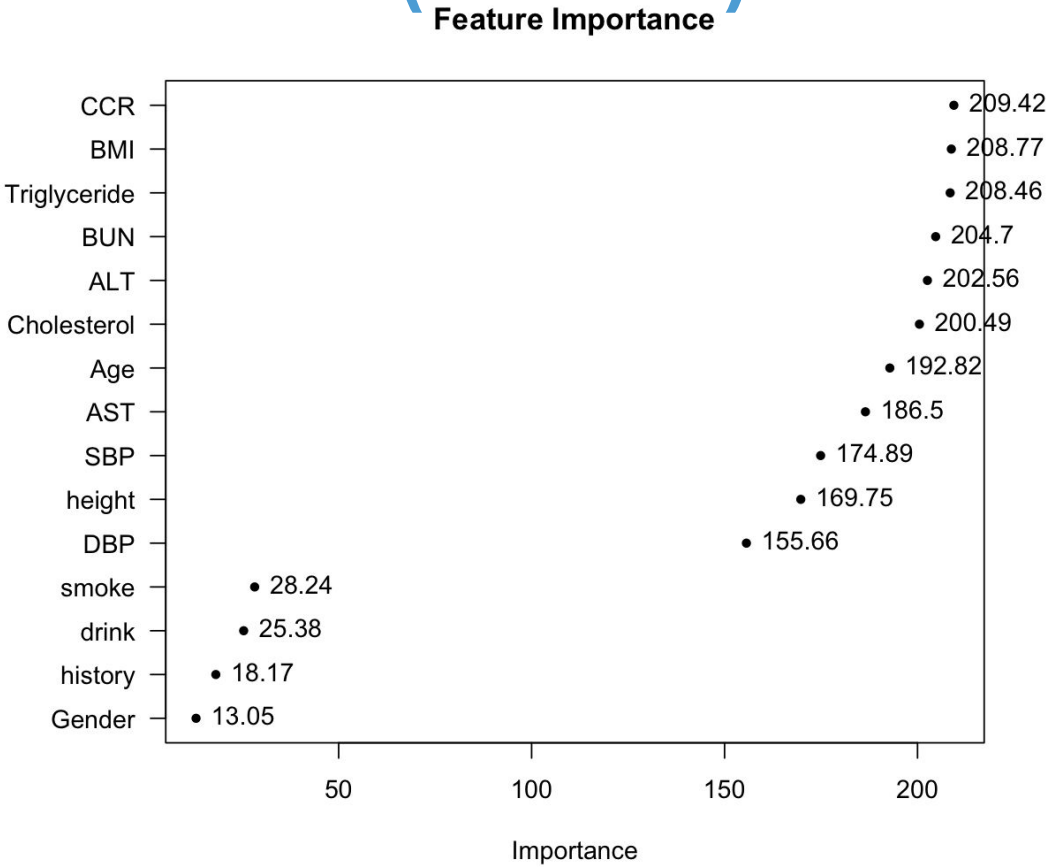
- ‘Ranger’ package in R
- Same training/test data
- Out of bag error
  - Similar to LOOCV
- Class Weights
  - No Diabetes: 1
  - Diabetes: 0.0196
- Threshold
  - Set as proportion of diabetes: 0.0196

Random Forest (Results)

- Sensitivity: 0.85
- Specificity: 0.70

Predicted/Actual	No	Yes
No	29145	126
Yes	12362	733

Random Forest (Results)



Top 5

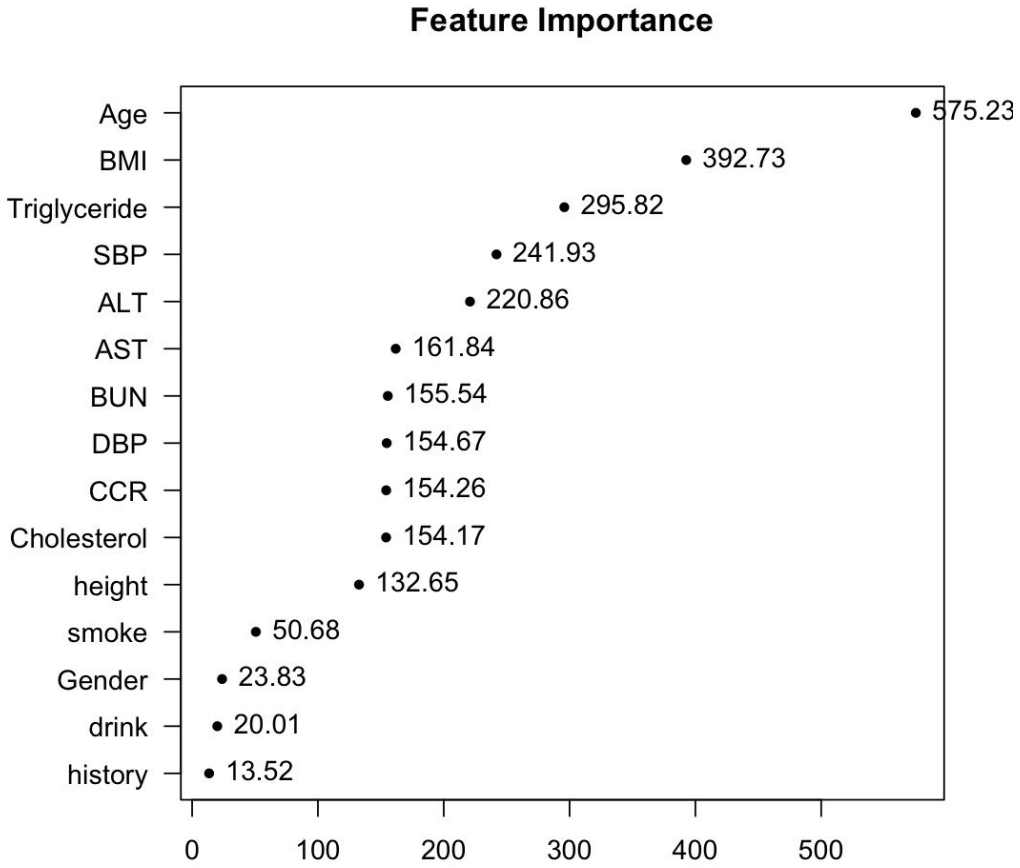
- CCR
- BMI
- Triglyceride
- BUN
- ALT

Random Forest (Results with Downsampling)

- Sensitivity: 0.85
- Specificity: 0.72

Predicted/Actual	No	Yes
No	29881	132
Yes	11626	727

Random Forest (Results with Downsampling)



Top 5

- AGE
- BMI
- Triglyceride
- SBP
- ALT

Feature importance

Since we standardized the independent variables, for both, Penalized and Standard Logistic regression, the feature importance is based on the magnitude of the estimates.

For the Random Forest algorithm, the feature importance is calculated through setting importance = ‘impurity’. It calculates decrease in node impurity for each feature in random forest model.

Important Variables

Penalized Logistic Regression:

Age,BMI,Systolic BP, Triglycerides, ALT, Family history

Bayesian Logistic Regression:

Age, BMI, Systolic BP, Triglycerides, ALT, Family history

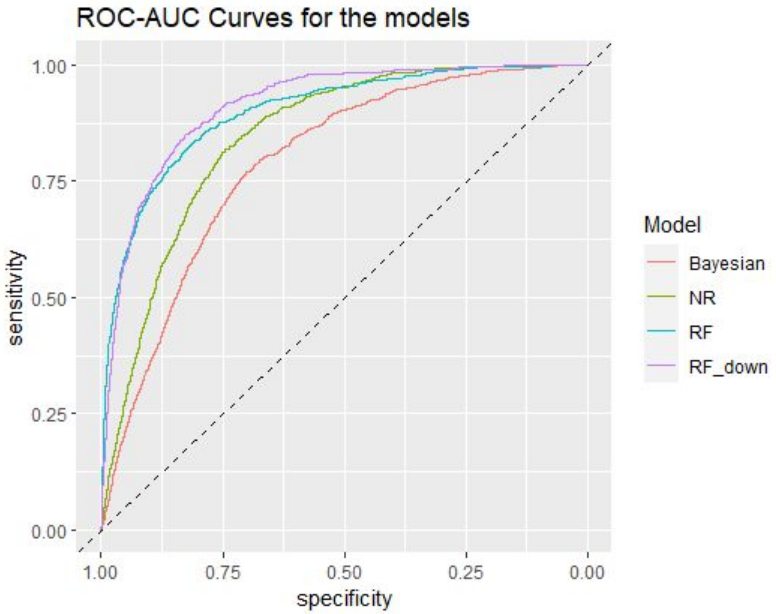
Random Forest:

Class weights: CCR, BMI, Triglycerides, BUN, ALT

Downsampling: AGE, BMI, Triglycerides, SBP, ALT

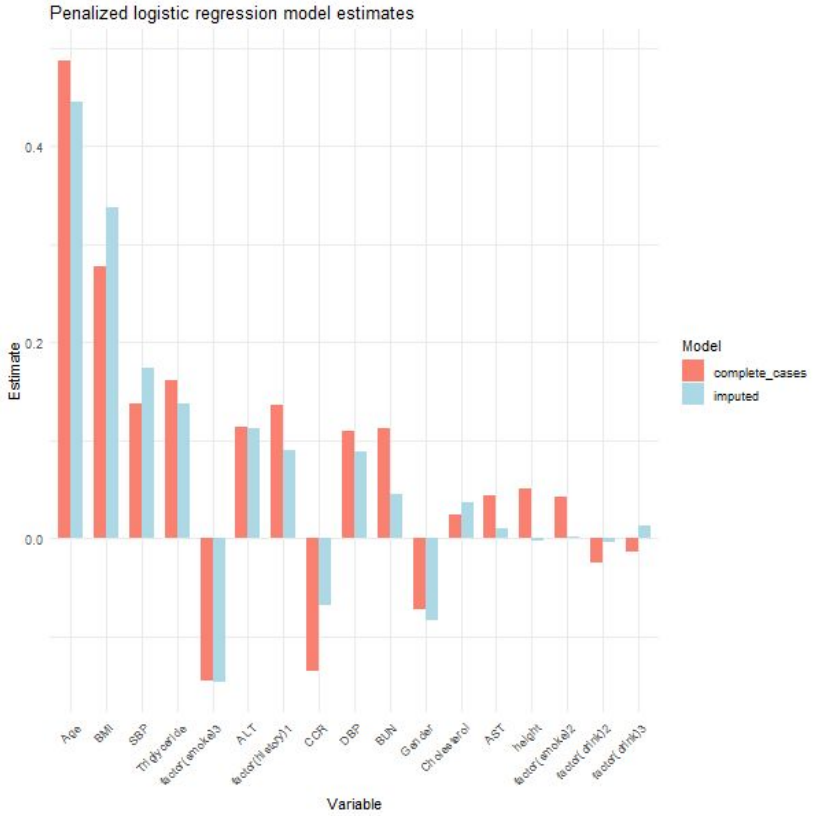
Model Evaluation

	Penalized Logistic Regression Bayesian approach		Random forest	
Metrics	Bayesian approach	Newton-Raphson	Class weights	Down-sampling
Precision	0.9802	0.9949	0.9964	0.9961
Recall	0.9804	0.7496	0.7677	0.8285
F1 score	0.9803	0.8550	0.8672	0.9046
AUC	0.7904	0.8466	0.9001	0.9134



Sensitivity Analysis

- Penalized Logistic regression
  - Using complete cases data and imputed data
  - Comparison of beta estimates



Conclusions

- Bayesian approach for solving logistic regression is better than other models
- Age, BMI, Triglycerides, ALT values are the most important risk factors for predicting Type 2 diabetes
- Age and BMI are positively associated with incidence of diabetes.

References

Chen, Y., Zhang, X. P., Yuan, J., Cai, B., Wang, X. L., Wu, X. L., ... & Li, X. Y. (2018). Association of body mass index and age with incident diabetes in Chinese adults: a population-based cohort study. *BMJ open*, *8*(9), e021768.

Göksülük, D. (2011). Penalized logistic regression. Yüksek Lisans.

Heikki Haario, Eero Saksman, Johanna Tamminen "An adaptive Metropolis algorithm," Bernoulli, Bernoulli 7(2), 223-242, (April 2001)

International Diabetes Federation. (2021). Diabetes Facts and Figures. Retrieved from https://idf.org/about-diabetes/diabetes-facts-figures/

Yeho, R. Random Forests. Medium. Retrieved [April 28, 2024] from [URL]