Cohort analysis: Hypertension in children with T1D

In this report, we analyzed the Hypertension in children with T1D cohort to understand the elevated BP subjects. Specifically, we focused on :

- undercoving the determing factors, among potential clinical/demographic/family/immune/genetic variables, for elevated BP in children with T1D.
- developing a machine learning model to predict elevated BP in children with T1D.
- evaluating the model performance using various metrics for various models.

The updates made on May in addition to the previous report (April 20th):

- added imputed C-peptide
- added SNP-level genotype data
- added BMI quantile and BMI z-score according to CDC growth chart
- duration of diabetes (months)
- total daily insulin dose are scaled in units/kilogram/day.
- **added model validation using leave-one-out cross-validation (LOOCV) and bootstrapping.**
- Note that Insulin pump therapy (yes/no) data is missing from the dataset. We will not be able to include this variable in our analysis for now.

In conclusion, we evaluated ML models to predict elevated BP in children with T1D. The best model is random forest which demonstrated ROC performance of 0.64 (95% CI). Further details of the results can be found at the end of this report.

Variables of interest

Response Variable

elevated_BP:

- Case group (Y=1): if the subject has elevated BP, i.e. patients who have any BP >90th % at both dates that BP was measured.
- Control group (Y=0): otherwise.

Genotype Predictors: GRS2 scores and SNP-level genotype data

- SNP-level genotype data were extracte for the 72 SNPs from the GRS2 paper (Sharp et al. 2019). Each SNP have 3 genotype status: 0, 1, 2, corresponding to homozygous, heterozygous, and homozygous respectively.

- T1D genetic risk score 2 (GRS2) was calculated using the SNPs from GRS2 paper (Sharp et al. 2019).

- **Features generated:**

- `IID`, `GRS` (standardized score),
- SNP_1_A1, SNP_1_A2, SNP_1_A3, SNP_2_A1, SNP_2_A2, SNP_2_A3, ..., SNP_72_A1, SNP_72_A2, SNP_72_A3.

Predictors: Clinical Features

- added BMI z-score, BMI quantile (calculated with CDC growth chart by age/sex using CDC Rpackage cdcanthro: https://github.com/CDC-DNPAO/CDCAnthro)
- Total daily insulin dose (re-scaled to units of units/kilogram/day)
- added "imputed C-peptide" as a clinical feature, calculated using the formula: A1c+(Ins)x 4, where Ins is the insulin dose in units/kg/day
- Duration of diabetes (in days), calculated as the difference between the date of the first diabetes diagnosis and the date of the visit
- Age (months)
- BMI
- Personal history of other autoimmune disease (yes/no, type of autoimmune disease)
- Markers of diabetes control over the last 3 months: Hemoglobin A1C (%)

Predictors: Ancestry & Family History

- Ethnicity (self-reported ethnic origin of 4 grandparents)
- Family history of T1D (yes/no)
- Family history of T1D, T2D or other forms of diabetes in 1st or 2nd degree relatives (yes/no)

- **Features generated:**
  - `cluster_pred_ancestry` (predominant continent where grandparents were born)
  - `cluster_shannon` (ancestry diversity): high value reflect higher diversity.
  - `Cluster_family_diabetes` ((yes/no))
  - `cluster_family_12_degree_bin` (Yes/No for 1st/2nd degree relatives)

Predictors: Antibody Data

- Glutamic acid decarboxylase 65 (GAD65) autoantibody titers (nmol/L)
- Insulinoma-associated antigen 2 (IA-2) autoantibody titers (nmol/L)
- Zinc transporter 8 (ZnT8) autoantibody titers (nmol/L)

- **Features generated:**
  - `96GAD`
  - `IA2`
  - `ZnT8`

Data Processing

- We removed the features with high correlations (rho>0.8) to avoid multicollinearity.
- Note that the number of SNP features are very high and will lead to overfitting for some machine learning methods. Therefore we run univaraite association analysis to shrink SNP-level feature pool to 13

features by removing any SNP-level feature that have association test p-value > 0.1. The selected SNP-level features are:
    - "X6.29840255_A1" , "X6.33081532_A1" , "X6.126377573_A2", "X6.32415221_A2" , "X6.32415221_A1",
    - "X6.32634974_A2", "X6.32634974_A1", "X6.32644524_A2", "X6.32644524_A1", "X6.32658670_A2" , "X6.32705608_A1", "X6.33081408_A1", "X11.2159830_A1"
- We removed the subjects with missing values in the response variable (elevated_BP) and the predictors.
- In summary, we have 288 subjects and 34 predictive features.

Modeling Srategy

We applied and evaluated the following classifiers on a held-out test set using two evaluataion metrics: **ROC** curve and **Precision-Recall** curve.. The ROC AUC measures the model's ability to distinguish between the two classes, while the PR AUC focuses on the model's performance on the positive class (elevated BP).

We considered the follwing models:

1. Logistic Regression (standardized)

Standardized logistic regression fits a linear model linking predictor variables to the log-odds of the binary outcome. This model serves as our baseline: it's fast to train, yields easily interpretable odds-ratios, and sets a performance floor against which more complex learners can be judged.

1. LASSO Logistic (`glmnet`)

LASSO logistic regression introduces an L_1 penalty on the magnitude of coefficients, shrinking many to exactly zero. This regularization both guards against over-fitting (especially when the number of predictors is large relative to sample size) and performs automatic feature selection by excluding weakly predictive variables. We fit via glmnet::cv.glmnet() using AUC-optimized cross-validation to select the penalty strength (λ). The resulting sparse model highlights only the strongest predictors, simplifying interpretation and often improving generalization.

1. Random Forest (caret)

Random forests build an ensemble of decision trees, each grown on a bootstrap-sample of the training set and considering a random subset of predictors at each split. By averaging across many decorrelated trees, the model captures non-linear relationships and high-order interactions while reducing variance. We used caret::train(method = "rf") with repeated cross-validation to tune the number of variables tried at each split (mtry). Random forests are robust to outliers, handle mixed-type features, and provide built-in measures of variable importance.

1. XGBoost (mlr)

XGBoost implements gradient boosting of decision trees, sequentially fitting each new tree to the residual errors of its predecessors. This method excels at capturing complex interactions and skewed distributions by optimizing a regularized objective (including L_1/L_2 penalties) under a fast, distributed framework. We wrapped it in mlr, tuning tree depth, learning rate, and sampling parameters via randomized search on PR-AUC.

XGBoost often outperforms other algorithms when appropriately tuned, at the cost of more hyperparameter complexity. However, it turns out our data is not large enough to benefit from the XGBoost algorithm and the perofmance is dominated by the age factor. Therefore, we do not recommend using XGBoost for this dataset.

1. Linear SVM (caret)

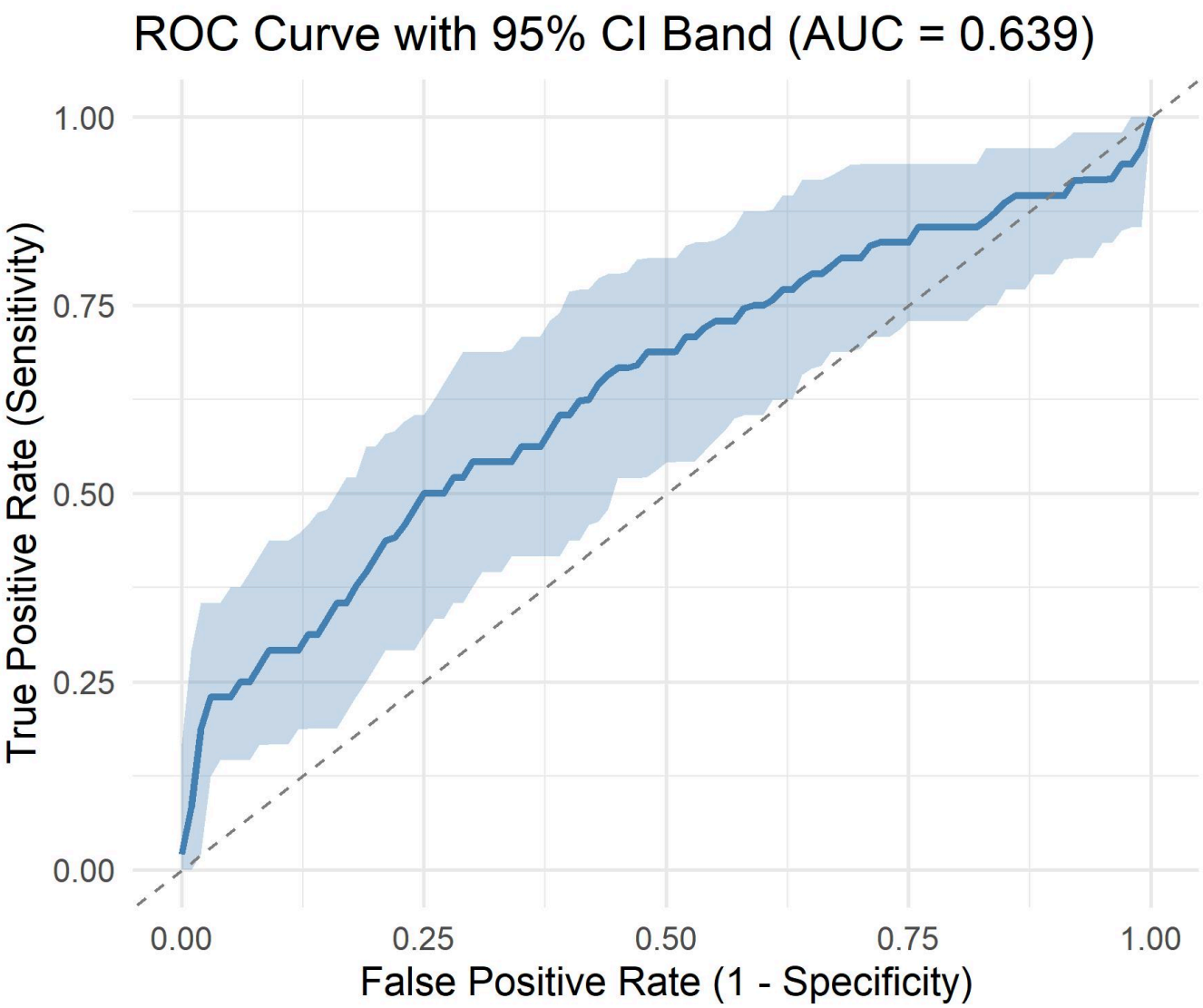
A linear Support Vector Machine seeks the hyperplane that maximizes the margin between classes in feature space. It's effectively a large-margin classifier that's robust to high-dimensional data and can be less sensitive to outliers than logistic regression. We trained with caret::train(method = "svmLinear"), tuning the cost parameter (C) via cross-validation to balance margin width against misclassification error. Although its decision boundary is linear, SVMs can yield strong performance in cases where classes are well-separated in a high-dimensional feature representation. However, due to limited samle size of this project, the performance of SVM is not as good as the Random Forest model on our dataset.

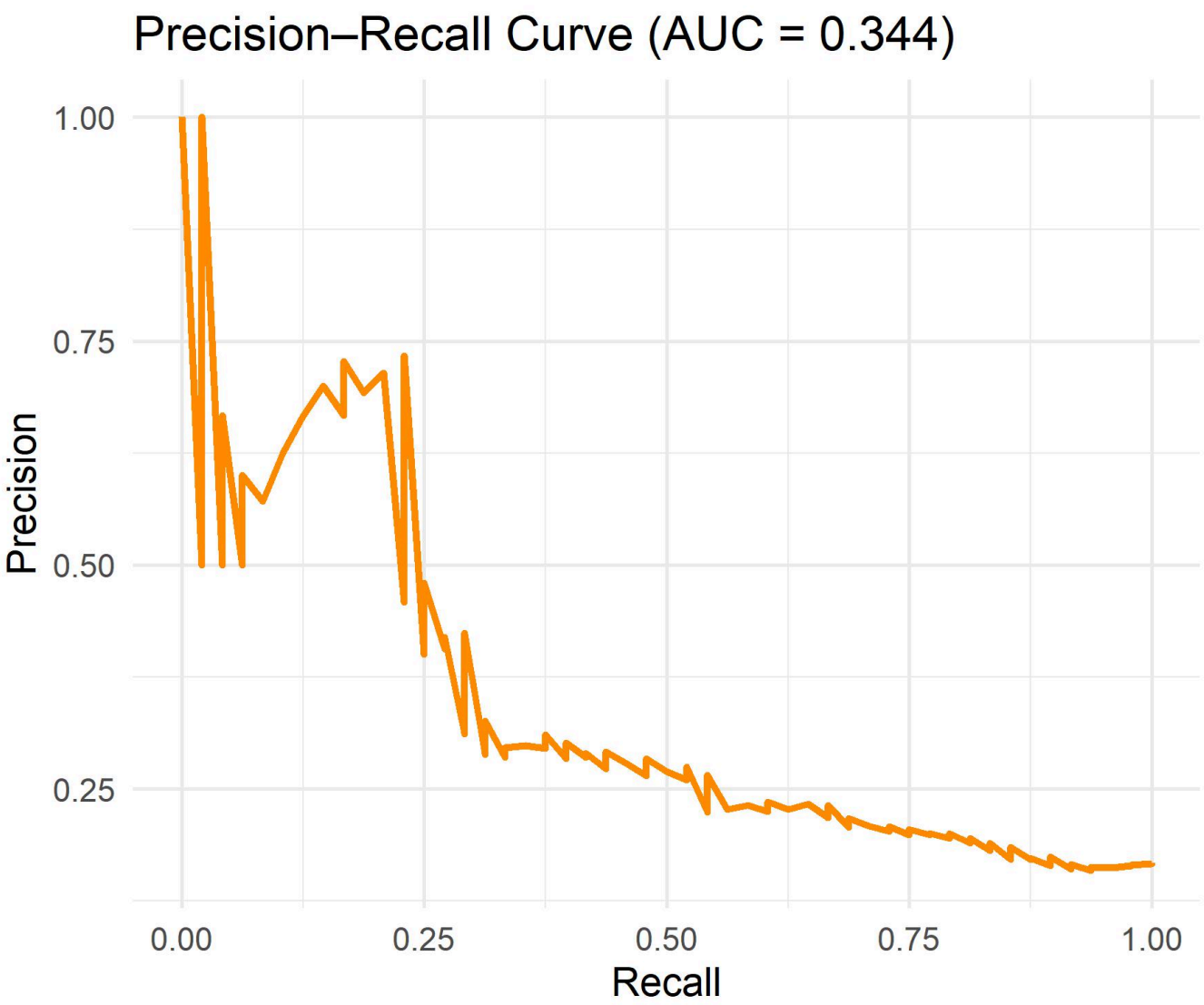
Model Evaluation

- **Two complementary model validation strategies were used:**
  - **1. Leave-One-Out Cross-Validation (LOOCV):**
    - A **95% confidence band for the ROC curve** was computed for each leave-one-out subject prediction using 200 bootstrap replicates.
    - This approach provides:
      - A nearly unbiased estimate of model performance.
      - Maximal use of training data in each iteration.
  - **2. Bootstrap Resampling (B = 200 iterations):**
    - A **95% confidence band for the feature importance** was computed using 200 bootstrap replicates:
      - A bootstrap sample of the dataset was drawn.
      - A random forest model was trained with **no internal resampling**.
      - **Variable importance scores** were extracted using `varImp()`.
    - After all iterations:
      - A distribution of feature importance values was compiled.
      - **95% bootstrap confidence intervals** were computed for each feature's importance.
    - This analyse provides:
      - The **stability of variable rankings**.
      - The **uncertainty around feature contributions** to model predictions.

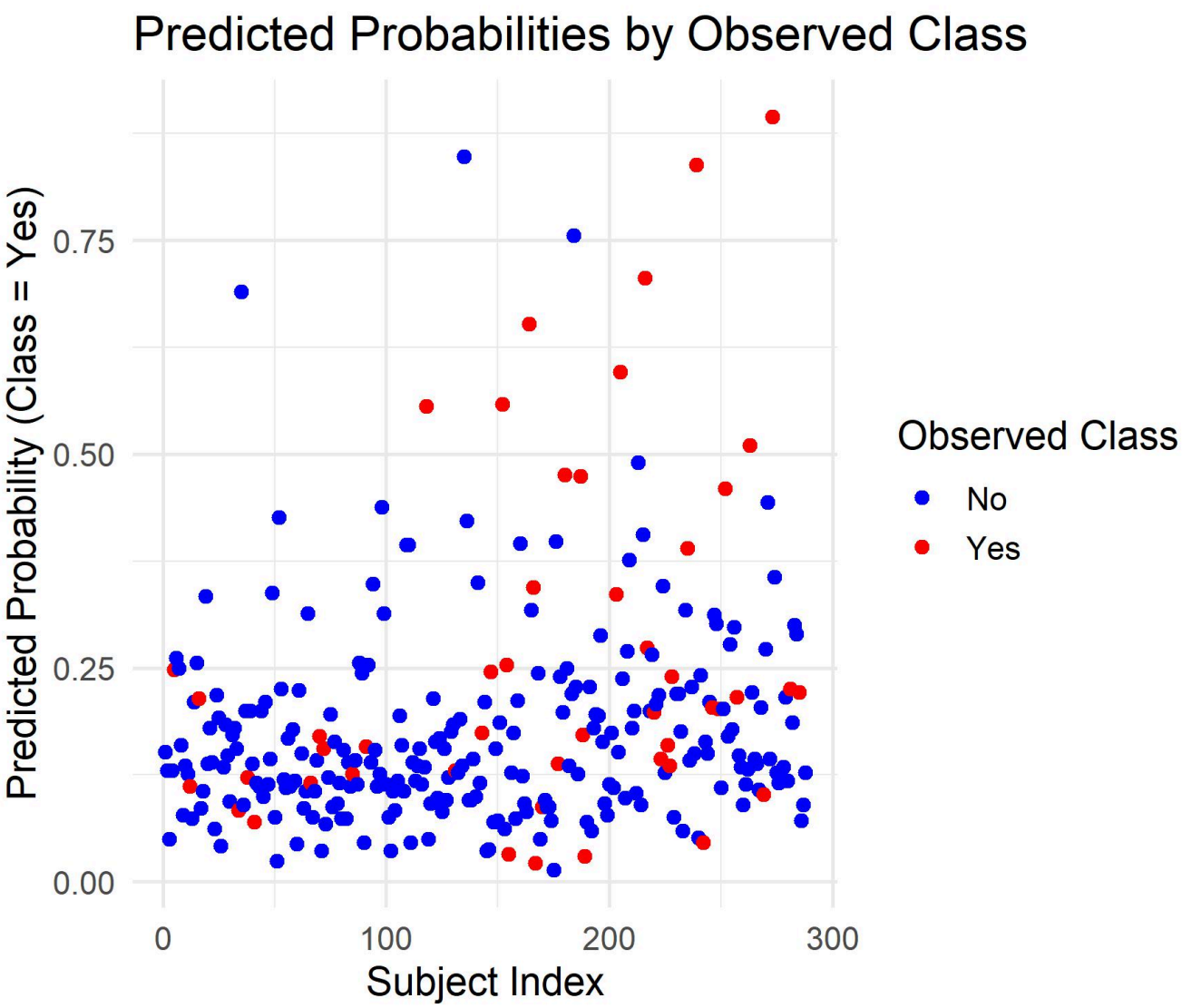
Conclusion

1. **Model Performance:**
   - **Random Forest** achieved the highest performance among all methods:


ROC Curve with 95% CI Band (AUC = 0.639)
True Positive Rate (Sensitivity)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
False Positive Rate (1 - Specificity)

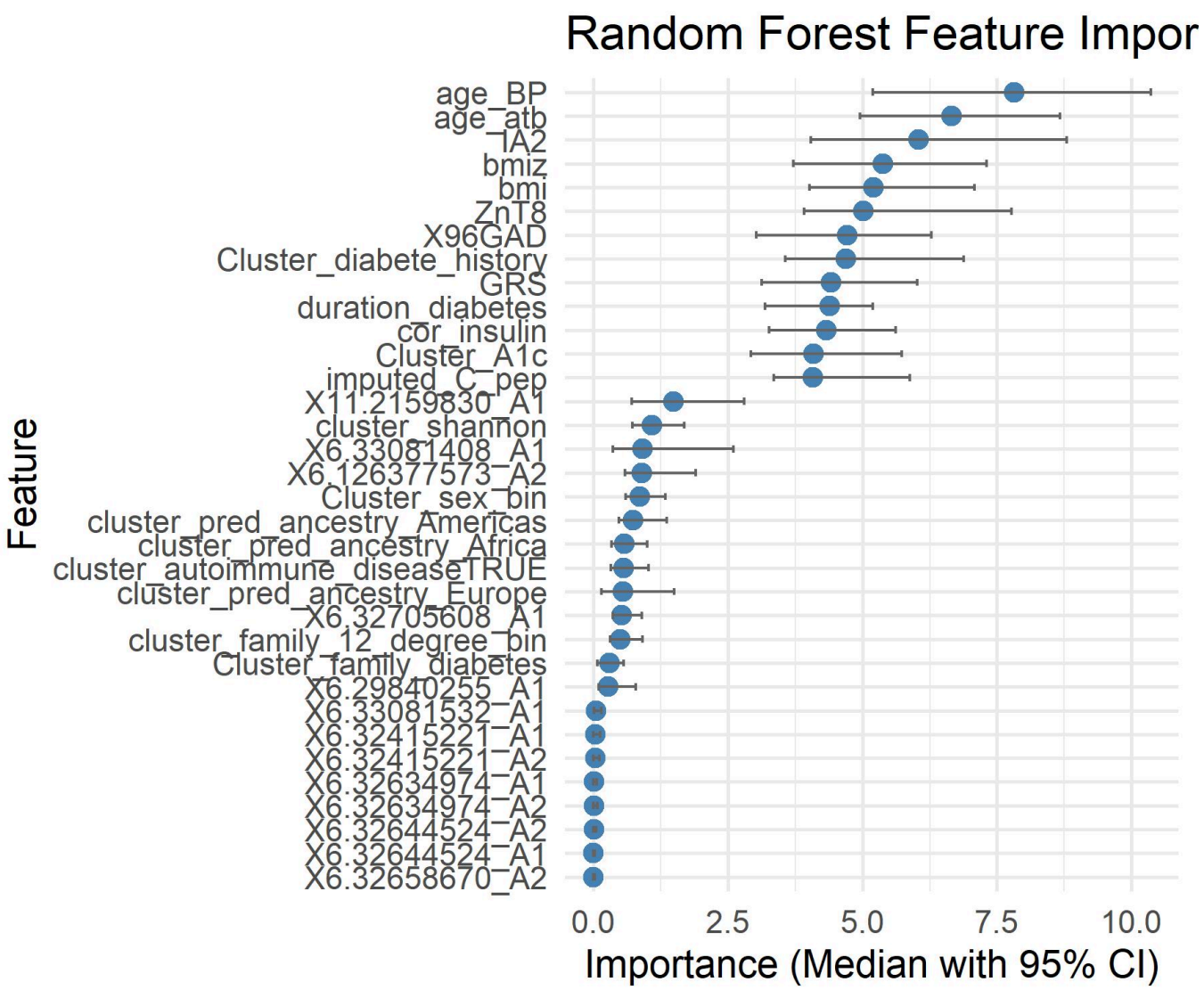


- The leave-one-out predicted probabilities showed that the model performance have relative low type-1 error (false positive rate) and high type-2 error (false negative rate).
- Specifically, if the model predicted probability > 0.4, then it is very likely that the subject has elevated BP.



1. **Key Predictors:**

- Across bootstrap samples, **age**, **autoantibody titers (IA2, GAD, ZnT8)**, **BMI and BMI-zscore** emerged as the top 5 most important features (Figure 2).


Random Forest Feature Impor
Feature
age_BP
age_atb
IA2
bmiz
bmi
ZnT8
X96GAD
Cluster_diabete_history
GRS
duration_diabetes
cor_insulin
Cluster_A1c
imputed_C_pep
X11.2159830_A1
cluster_shannon
X6.33081408_A1
X6.126377573_A2
Cluster_sex_bin
cluster_pred_ancestry_Americas
cluster_pred_ancestry_Africa
cluster_autoimmune_diseaseTRUE
cluster_pred_ancestry_Europe
X6.32705608_A1
cluster_family_12_degree_bin
Cluster_family_diabetes
X6.29840255_A1
X6.33081532_A1
X6.32415221_A1
X6.32415221_A2
X6.32634974_A1
X6.32634974_A2
X6.32644524_A2
X6.32644524_A1
X6.32658670_A2
0.0
2.5
5.0
7.5
10.0
Importance (Median with 95% CI)