Team Meeting

02 SEPTEMBER 2022 / 15:15 AM / CONFERENCE ROOM

Attendees

Ni, Stefan, Xavier, Ate, Zexi, Jiadong, Shuai

Last Meeting Follow-up

- Variable Selection by Random Forest

Random Forest	Original - 833	Round 1 - 319	Round 2 - 121
E-Risk	0.8115	0.7787	0.8286
BSGS	0.8158	0.8014	0.7724

New Business

2. Experiment design for application
    - Plan 1
        - Use modelling to select variables from all 5 datasets, then use the selected feature to train E-Risk
    - Plan 2
        - Use 80% from each of the 5 datasets to do variable selection and training, then the rest 20% on testing
    - Plan 3
        - Use 4 datasets to do the model training and variable selection, then the test 1 on testing
    - Plan 4
        - Use all 5 datasets to do cross-validation for training and variable selection (5C4 times)
    - **Plan 5**
        - Use 4 datasets to do variable selections and training, and the remaining one for testing, repeat this process for 5 times (do the cross validation on the 4 datasets).

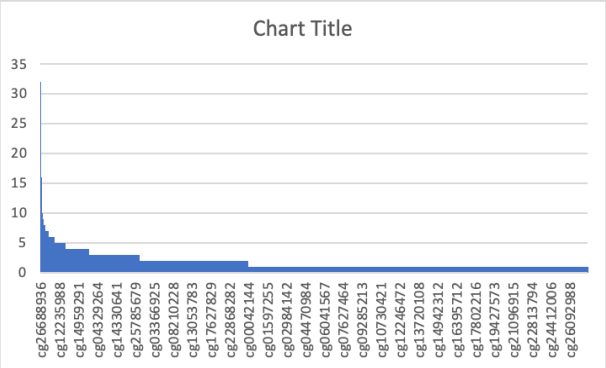
Replace the missing values with mean

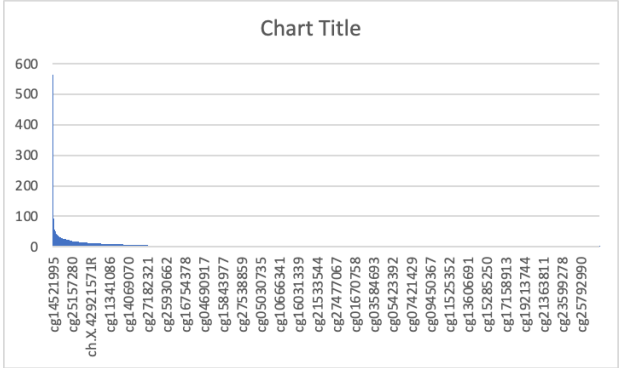
3. Dataset size

Dataset Name	Number of Data For training	Number of null	Number of family members
E-Risk	1464	194	
BSGS	130	228	256
Denmark	180		
E-MTAB	625	23	
AMDTSS	246		215

4. Missing value for 450K dataset

Dataset Name	Column With Missing Value	Total Missing Value	Average Missing Value Per Column	Max Missing Value Per Column	Total row number
E-Risk	0	0	0	0	
BSGS	183201	740468	4.0	564	614
Denmark	0	0	0	0	
E-MTAB	18339	32283	1.8	32	648
AMDTSS	0	0	0	0	





Notes

Action Items

1.