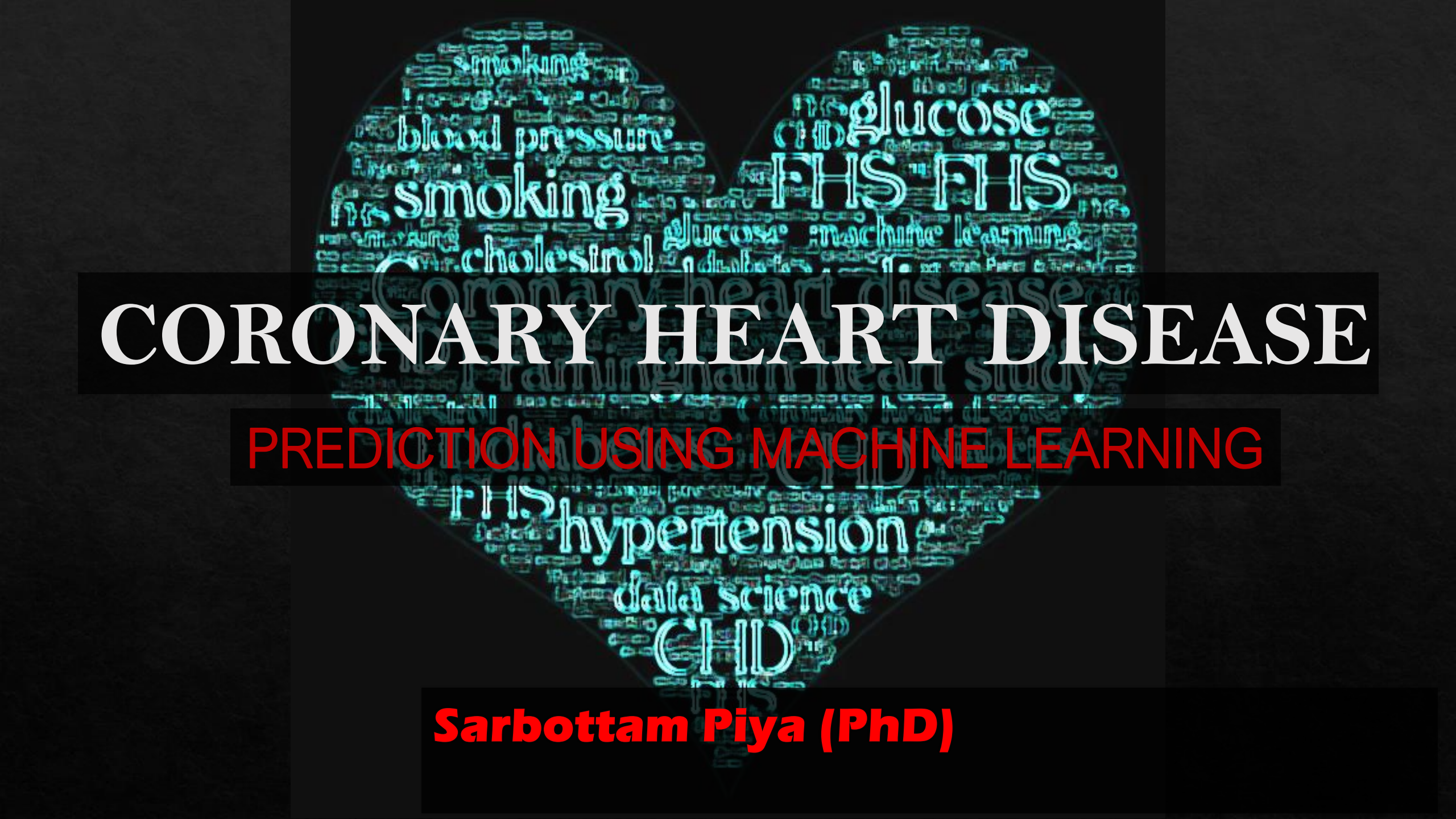


CORONARY HEART DISEASE

PREDICTION USING MACHINE LEARNING

Sarbottam Piya (PhD)

Introduction

Prominent causes of death

17.9 million people die annually of heart related diseases (WHO 2017)

difficult to diagnose

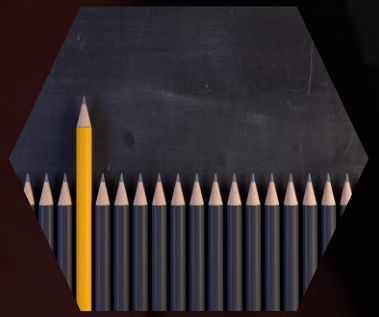
Diagnosis of heart

Expensive

Problem statement

Huge medical records

Difficult to comprehend



Data ignored

Clinical decisions

Objective

- Develop a predictive model of CHD using machine learning based on the demographic, comorbidity, vital sign and some laboratory investigation data

Audience

- Healthcare professional
- People associated with high risk factor

Dataset

- Original cohort dataset from Framingham heart study (FHS)

Demographic information

- Gender, Age And Education

Comorbidity

- Blood Pressure, Stroke, Hypertension, Diabetes, Smoking Habit

Vital Statistics

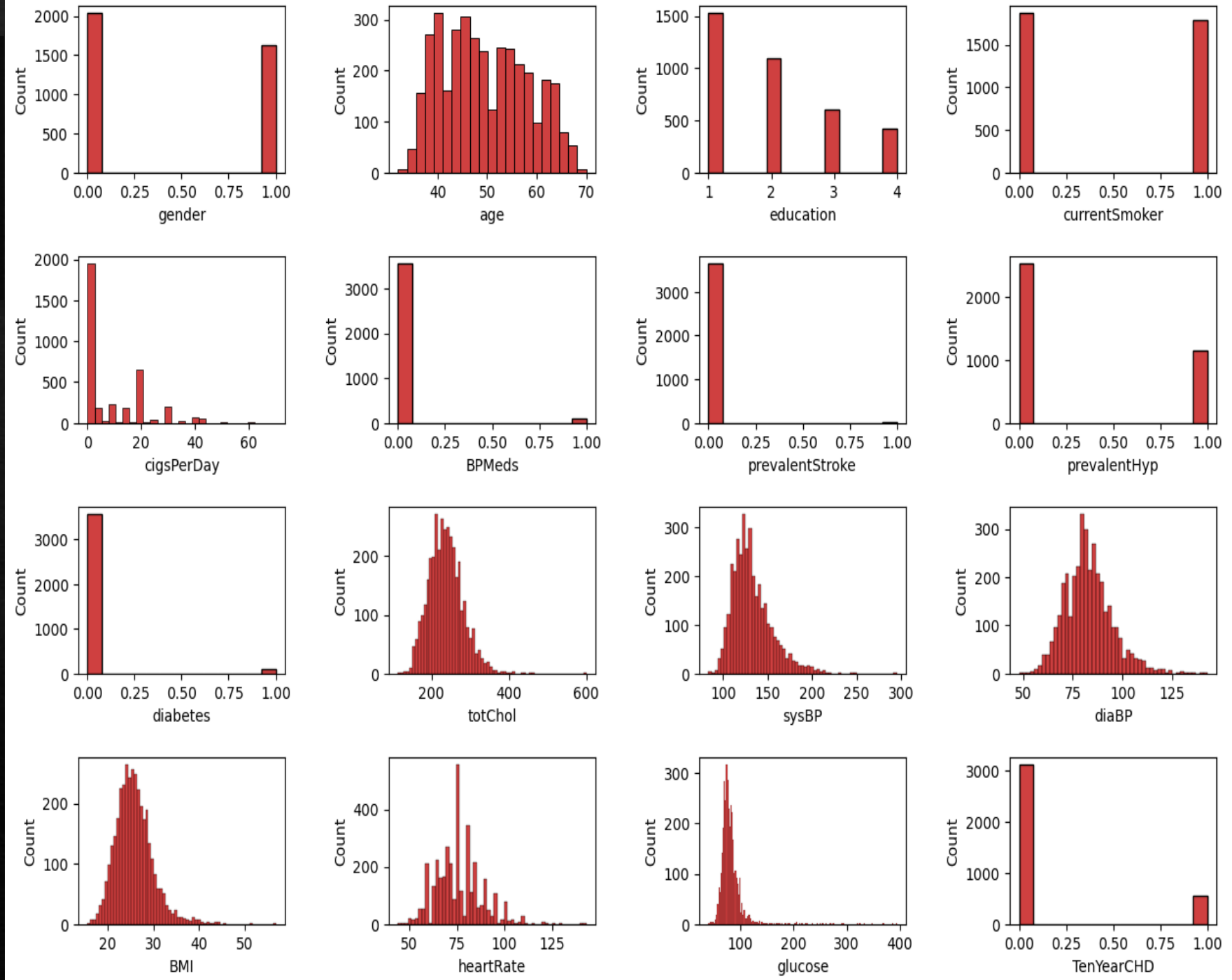
- Systolic Blood Pressure, Diastolic Blood Pressure, Body Mass Index (BMI), Heart Rate

Lab Investigation

- Total cholesterol level
- Glucose level

Data wrangling

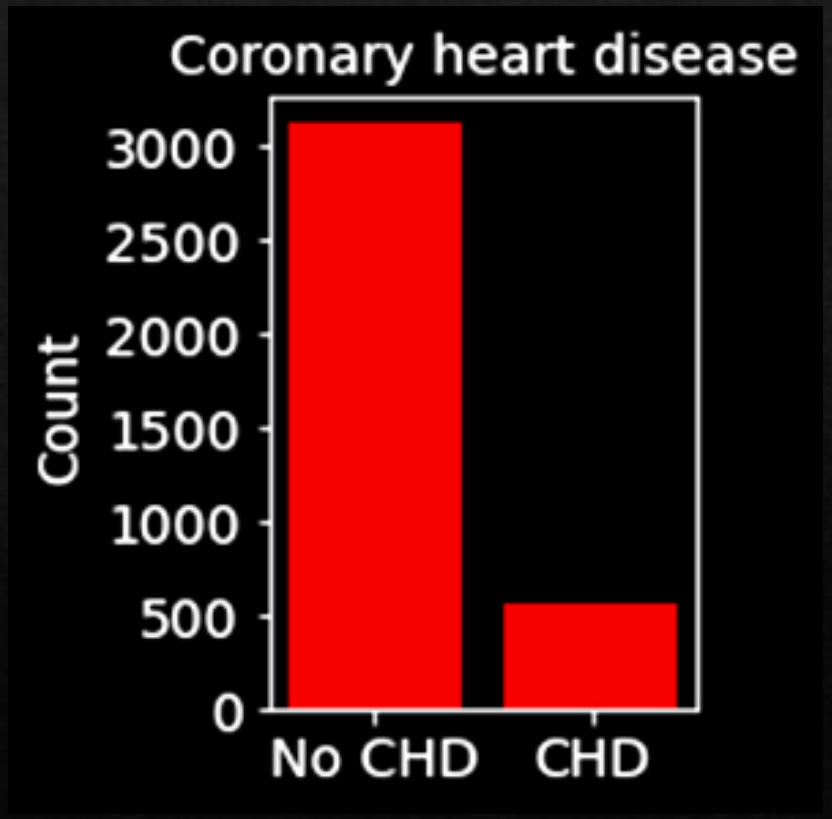
- No duplicate rows
- 4240 rows
- 3658 rows left after removing missing data



Exploratory data analysis

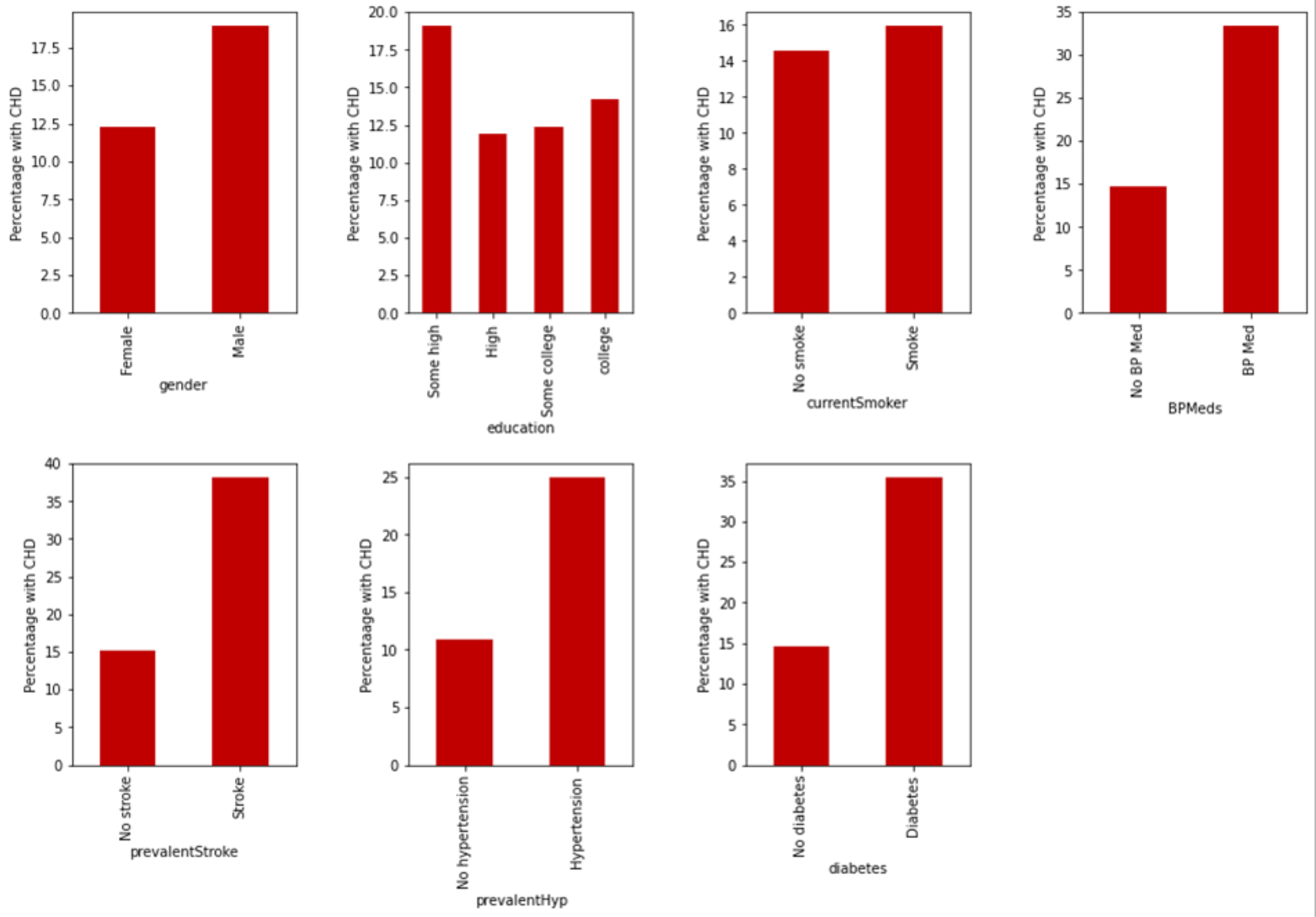
Target variables

- Class imbalance in the target variables
- Approximately, 5.5-fold higher dataset for no CHD compared to CHD

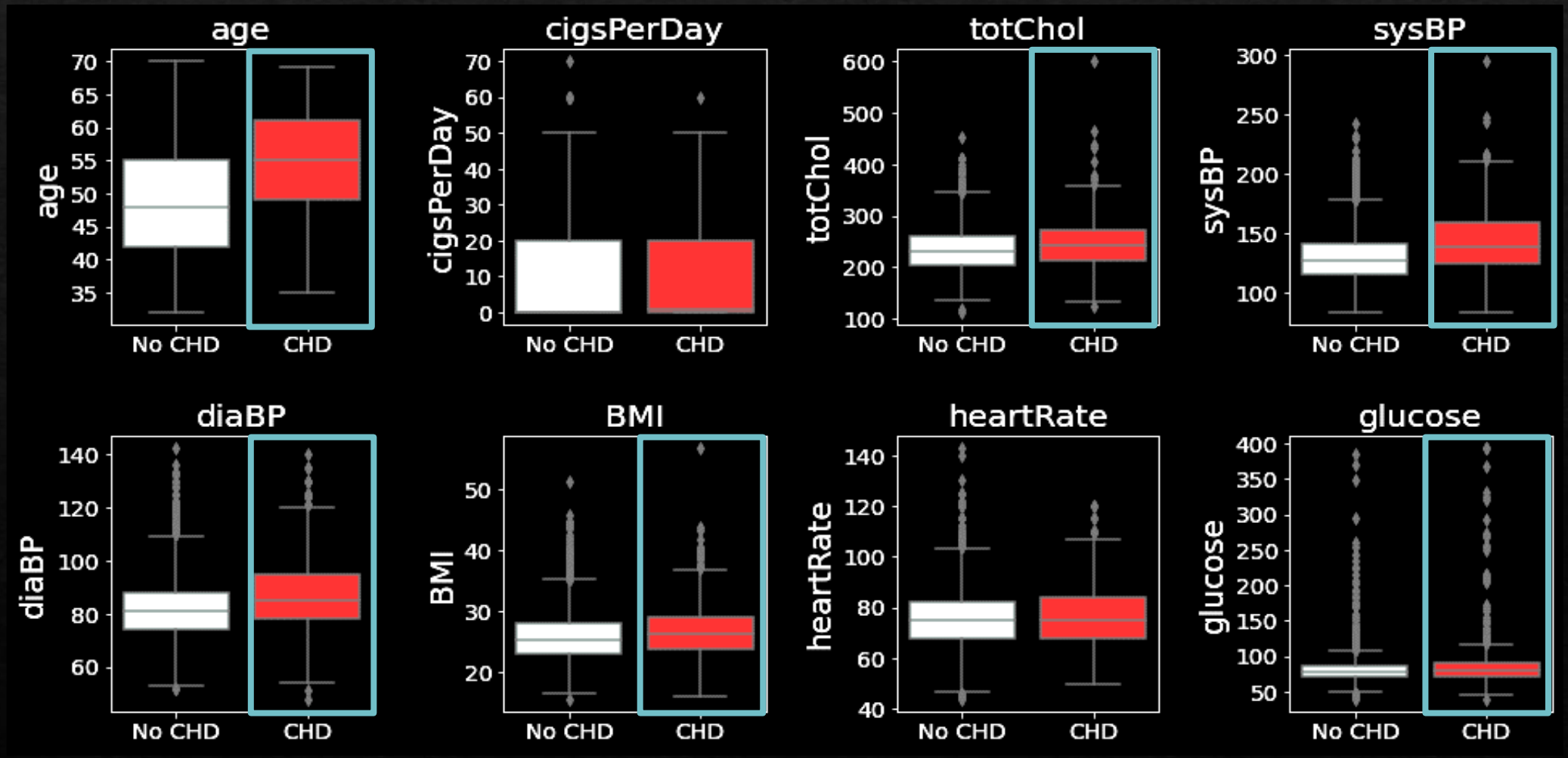


Categorical features

- Chi-square test showed that all the categorical features except ‘currentSmoker’ were significantly associated with CHD

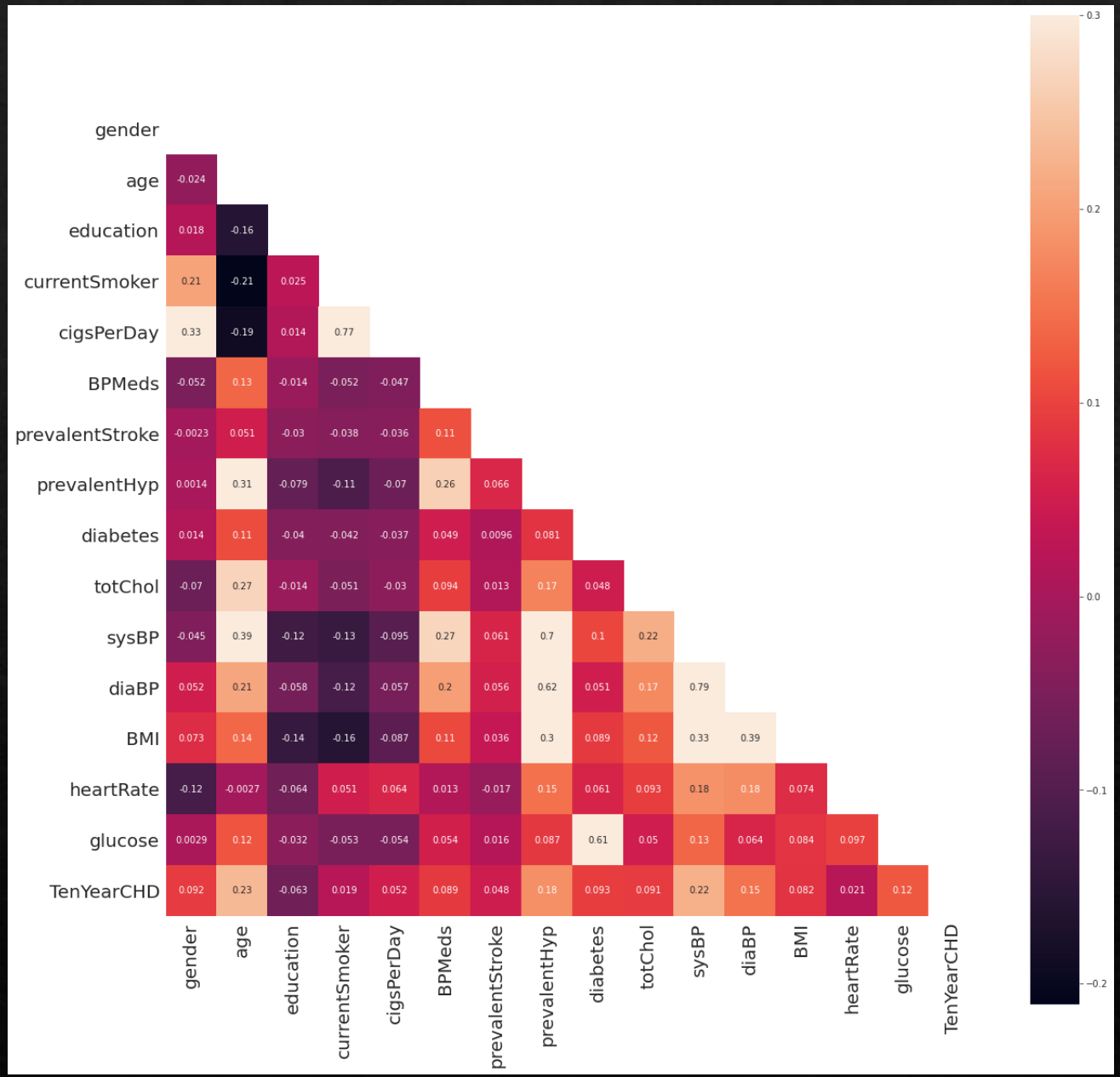


Numerical features

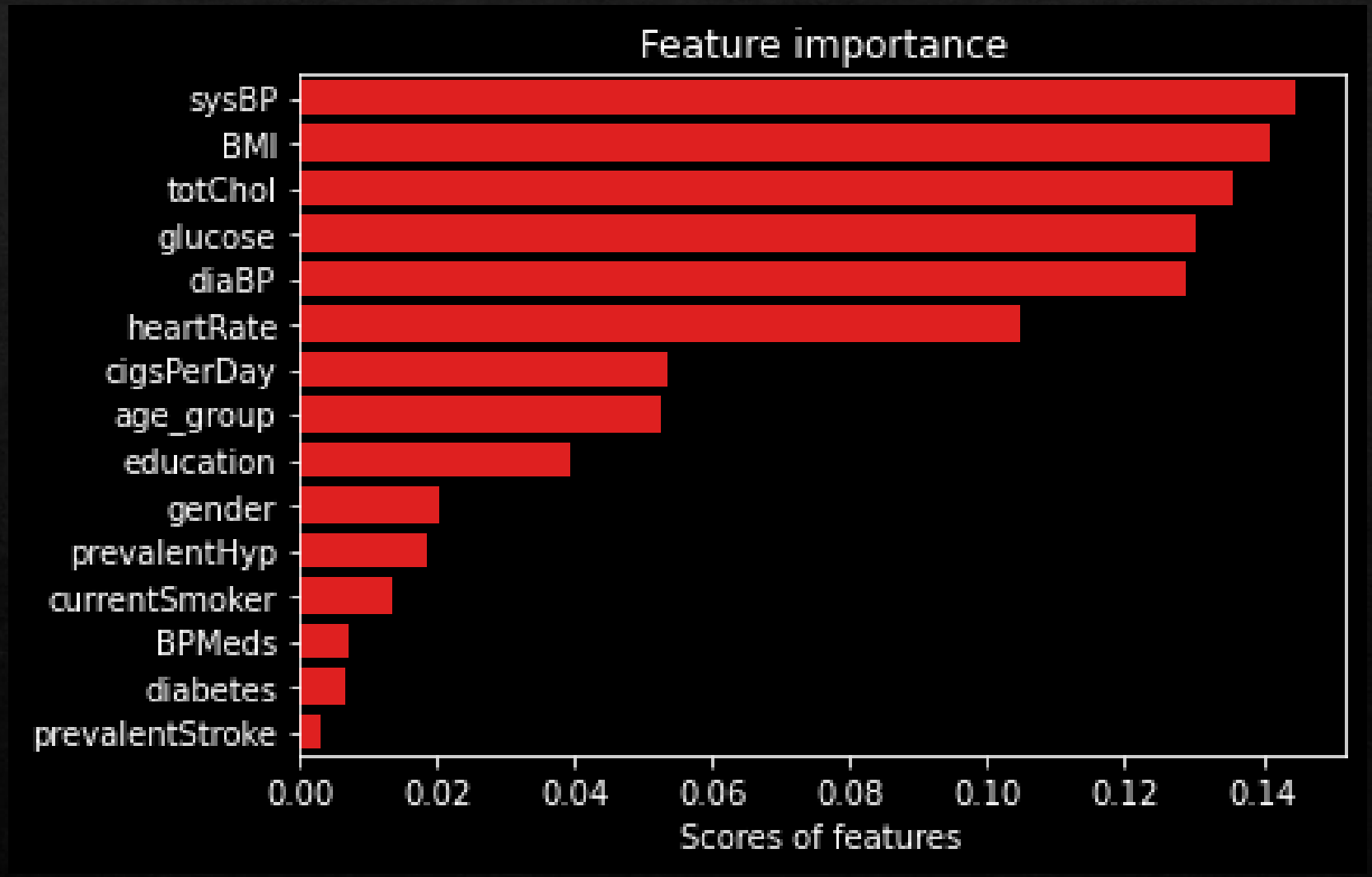


Correlation among variables

- **None of the features were strongly correlated to CHD**
- **Some of the features were strongly correlated among themselves**



Feature selection



Machine learning

- Various machine learning models were evaluated
- Grid search was done for hyperparameter tuning
- ROC-AUC was used as a scoring metrics

Selection of best model

Algorithm	ROC-AUC score
Logistic Regression	0.732449
Naive Bayes	0.718285
Random Forest	0.717543
Gradient Boost	0.713462
KNN	0.696594
SVM	0.667709

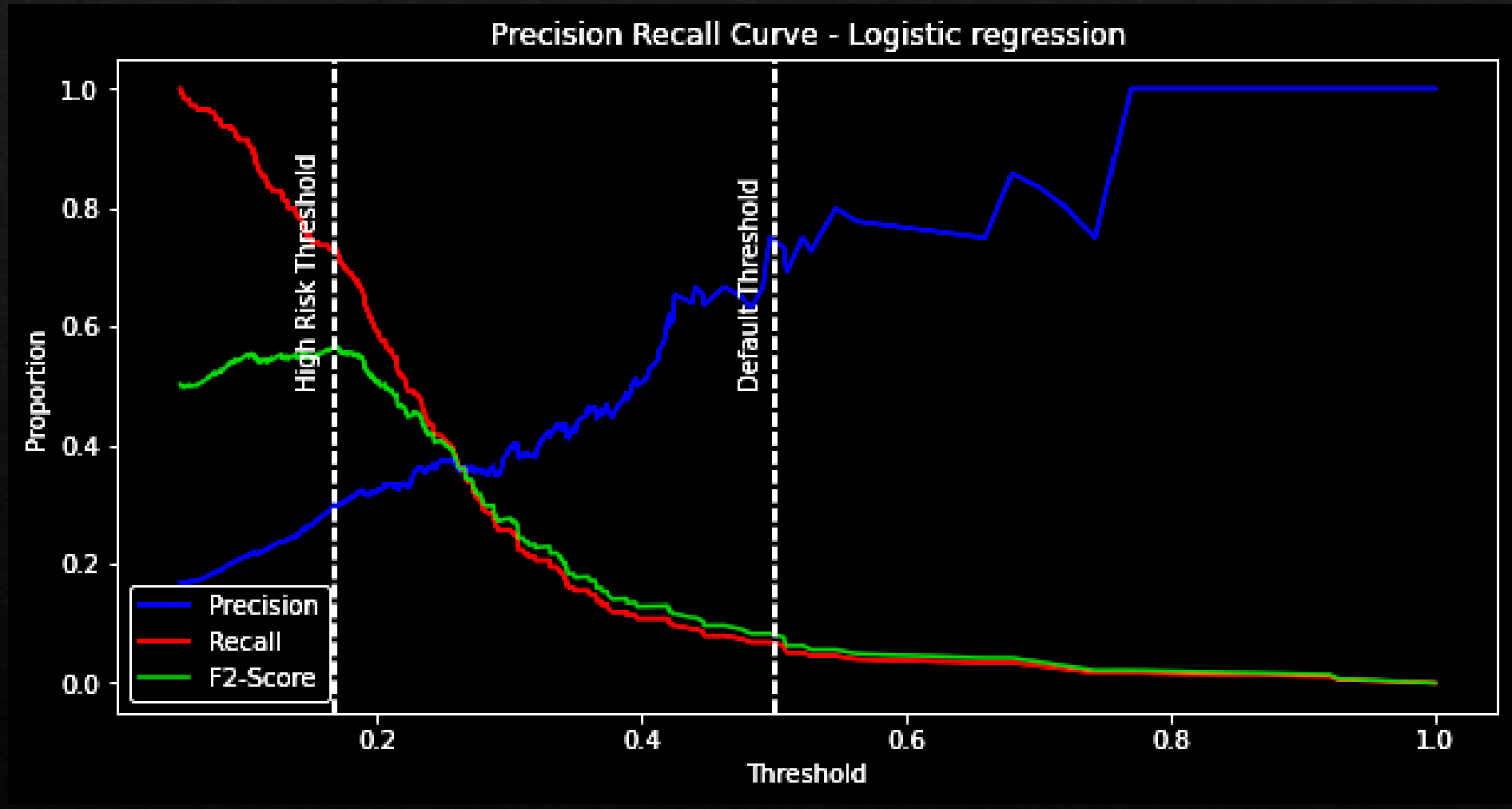
Adjusting the Probability Threshold

	Precision	Recall	F1-Score	Support
No CHD	0.85	1.00	0.92	923
CHD	0.73	0.06	0.12	175
Accuracy			0.85	1098
Macro avg	0.79	0.53	0.52	1098
Weighted avg	0.83	0.85	0.79	1098

beta =2,, threshold=0.162

	Precision	Recall	F1-Score	Support
No CHD	0.93	0.67	0.78	923
CHD	0.30	0.73	0.42	175
Accuracy			0.68	1098
Macro avg	0.61	0. 70	0.60	1098
Weighted avg	0.83	0.68	0.72	1098

Plot demonstrating F2 score, precision and recall at different thresholds



Undersampling improved the model

Beta =2, threshold = 0.458

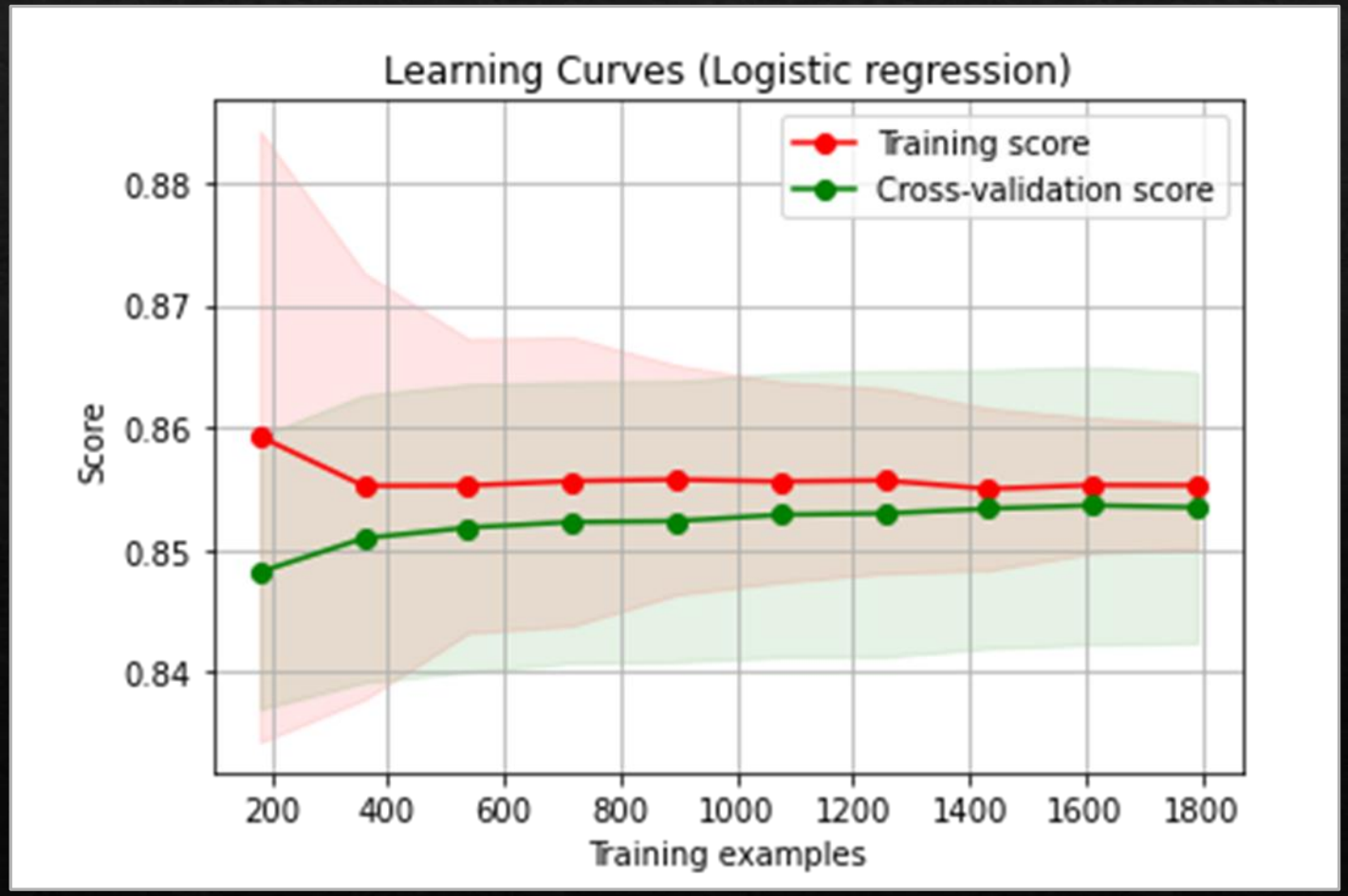
	Precision	Recall	F1-Score	Support
0	0.94	0.58	0.72	923
1	0.26	0.79	0.39	175
Accuracy			0.61	1098
Macro Avg	0.60	0.68	0.55	1098
Weighted Avg	0.83	0.61	0.66	1098

Oversampling with SMOTE further improved the model

Beta =2, threshold = 0.392

	Precision	Recall	F1-Score	Support
0	0.95	0.47	0.63	916
1	0.25	0.88	0.39	182
Accuracy			0.53	1098
Macro Avg	0.60	0.67	0.51	1098
Weighted Avg	0.83	0.53	0.59	1098

Will additional data improve the model?



Major findings

- Blood pressure, BMI, cholesterol, and glucose level are the key predictive features
- Logistic regression with oversampling performed the best (recall=0.8 and precision=0.28)
- Scope of future improvement

Future improvements

- Second (children of original cohort) and third generation (grandchildren of original cohort)
- Omni cohort that reflects more diverse samples

Acknowledgements

- **I would like to thank Ben Bell for providing very helpful feedback to improve this project.**

REFERENCES

CDC (2020). Heart Disease in the United States. https://www.cdc.gov/heartdisease/facts.htm. Retrieved on 1/30/2021.

World Health Organization (2017) Cardiovascular diseases (CVDs). https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Retrieved on 1/30/2021.

Thank You