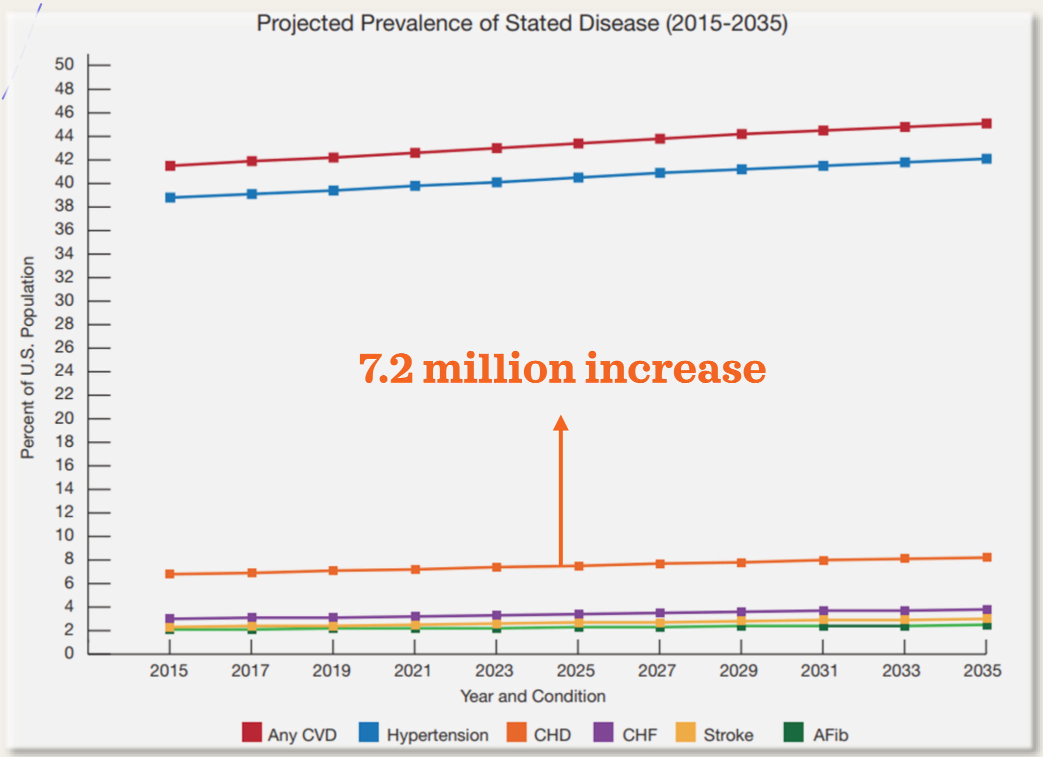
Analysis of 10-Year Coronary Heart Disease Risk

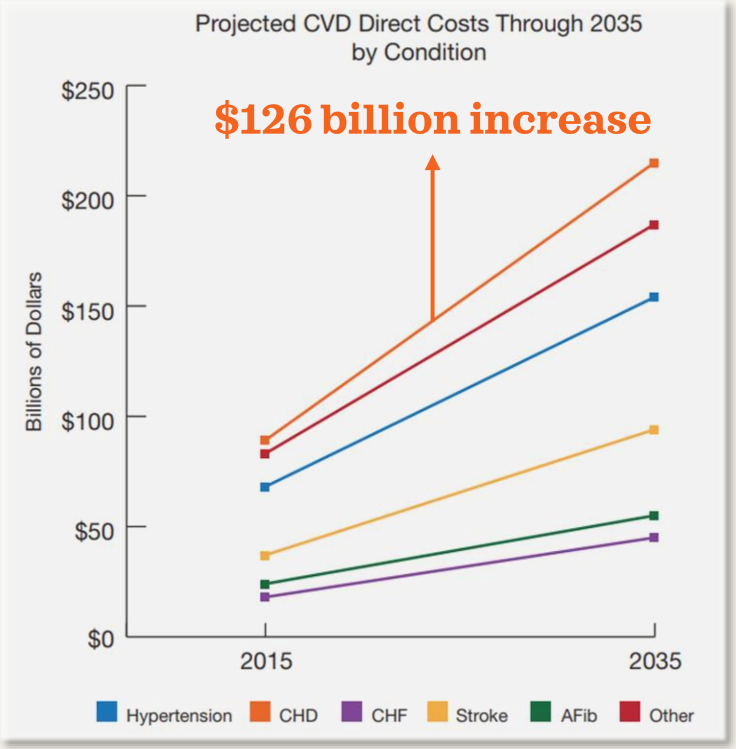
ISYE7406 Project Group 101

JunChang Song
Yuxi Guo
TaeWoo Kim

Introduction



7.2 million increase

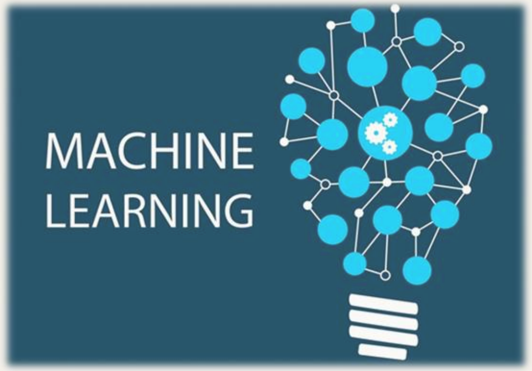


$126 billion increase

"By 2035, **nearly half** of the U.S. population will have some form of cardiovascular disease."

Reference : CARDIOVASCULAR DISEASE: A COSTLY BURDEN FOR AMERICA PROJECTIONS THROUGH 2035 (link)

Problem Statement & Business Justification

	Potential Problem	Possible Approach	Expected Outcome
Individual	• Increased health risk • Increase in healthcare expenses		• Prevent individual CHD risk • Save potential healthcare expenses
Healthcare business	• Losing market share without further research on CHD	• Exploiting machine learning techniques to identify key factors of CHD risk in 10 years	• Provide personalized healthcare strategies • Increased market share

Data Description

- Data source : Kaggle platform (link). The dataset originates from a continuous cardiovascular study involving the residents of Framingham, Massachusetts
- 4,132 data points with 16 variables
- Target Variable (1):
  - 10-Year CHD Risk (binary)
- Predictor Variables (15):
  - Demographics: Sex, Age
  - Behavioral: Education, Current Smoker, Cigs Per Day
  - Medical History & Current: BP Meds, Prevalent Stroke, Hypertension, Diabetes, Tot Chol, Sys BP, Dia BP, BMI, Heart Rate, Glucose

※ The detail description can be found in Appendix.

Data Analysis Steps

Exploratory Data Analysis

- Data Cleaning
- Distribution
- Correlation
- Potential Outliers
- PCA plots

Feature Selection

- LASSO
- Univariate selection method (SelectKBest)
- Tree-based method (RandomForestClassifier)
- Forward elimination

Oversampling Method

- SMOTE
- SMOTENC

Modeling

- Random Forest
- Gradient Boosting
- Logistic Regression
- SVM
- LDA
- KNN
- MLP

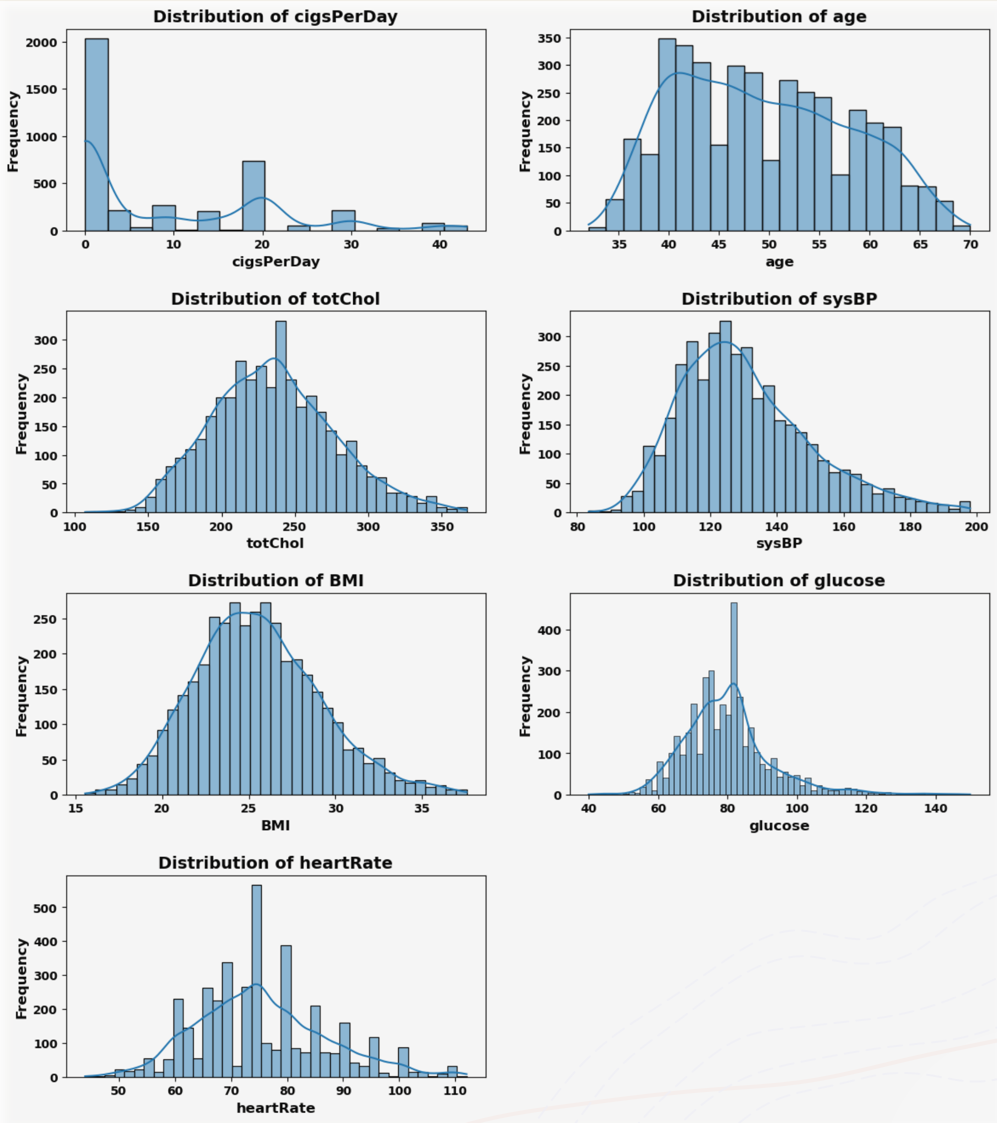
Model Selection & Interpretation

- Performance metrics
- Interpretability

Exploratory Data Analysis
- Distribution

+ Upon examining the distribution of numerical variables, it was observed that while some variables appear to follow a normal distribution, others do not, displaying a skewed distribution.

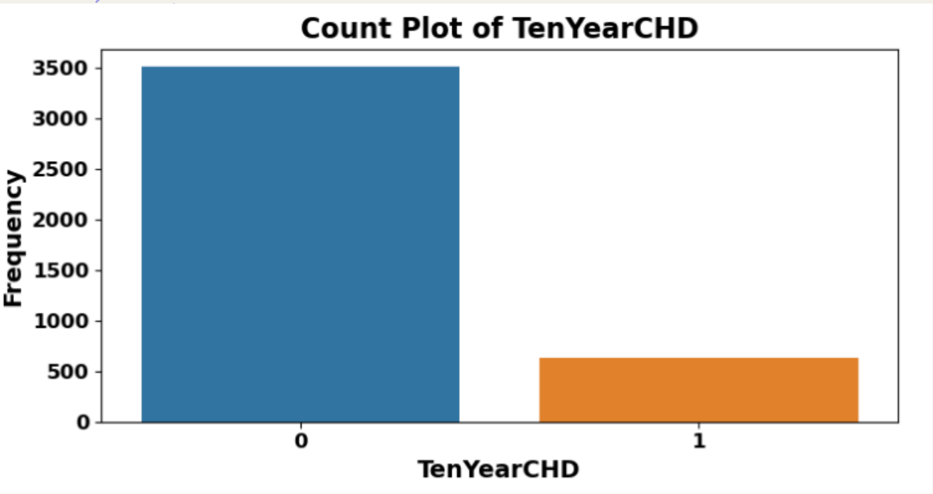
Numerical Predictor Variables



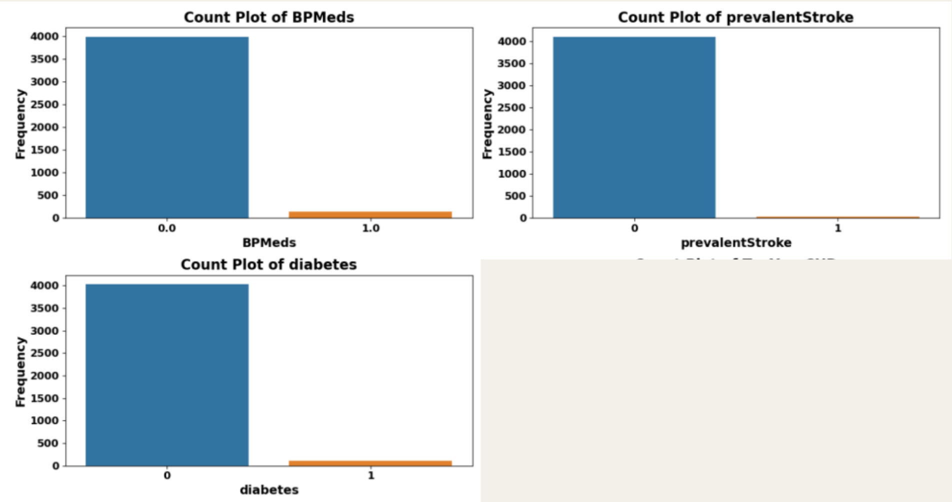
Exploratory Data Analysis

- Distribution

Target Variable



Categorical Predictor Variables



- + Checking our target variable, we can see that it is highly imbalanced. This suggests that we might need to employ resampling methods before modeling.
- + Because of that, categorical variables are highly imbalanced as well, indicating that we should examine them closely.

Exploratory Data Analysis

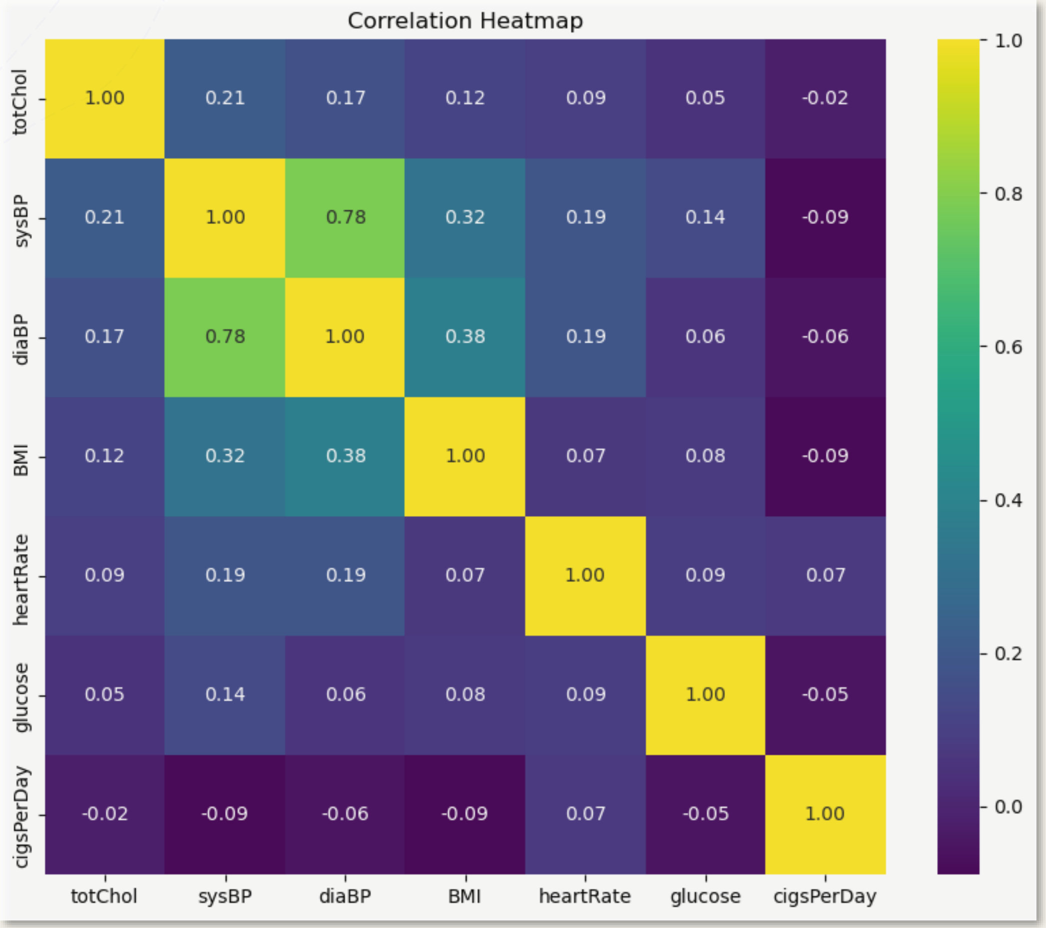
- Chi Square Test

- + male: Significant association, p-value = 8.45e-08
- + education: No significant association, p-value = 0.087
- + currentSmoker: No significant association, p-value = 0.308
- + BPMeds: Significant association, p-value = 3.04e-09
- + prevalentStroke: Significant association, p-value = 0.00018
- + prevalentHyp: Significant association, p-value = 0.00103
- + diabetes: p-value = 8.34e-10

The chi-square tests reveal that male, BPMeds, diabetes, prevalentStroke, and prevalentHyp have a statistically significant association with the target variable, whereas education and currentSmoker do not.

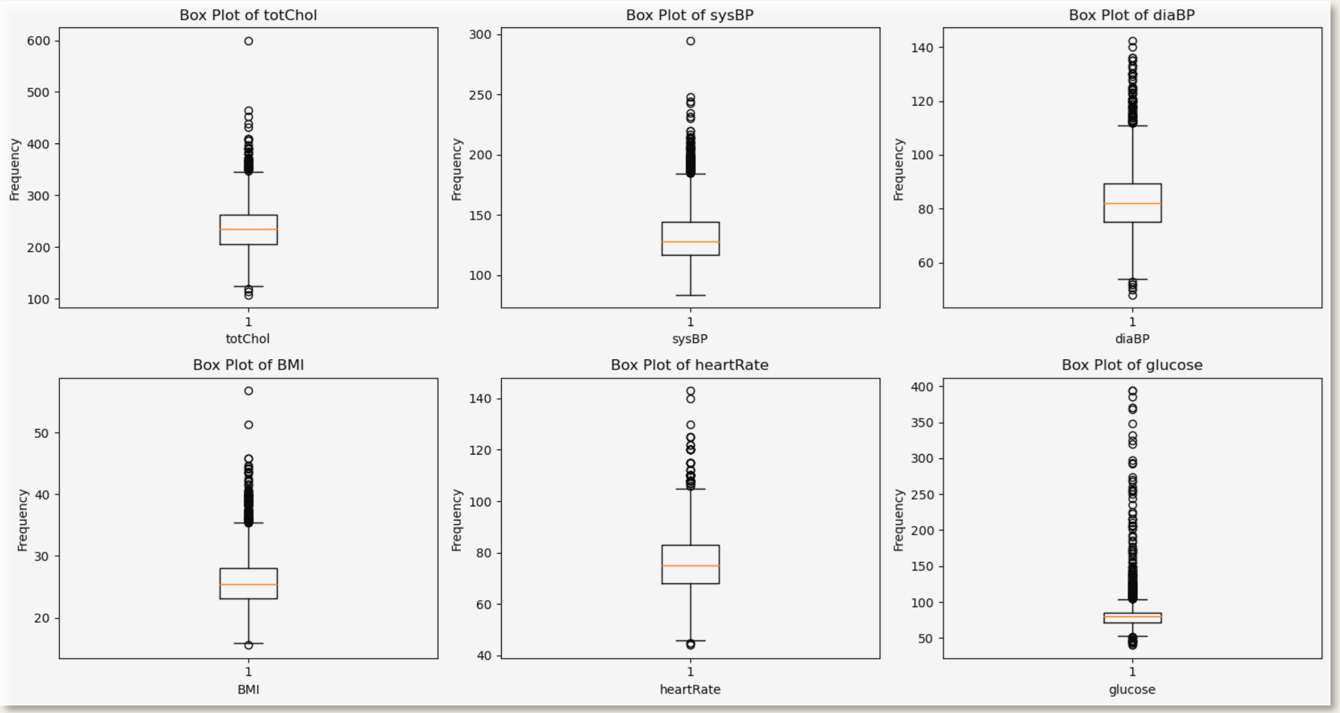
Exploratory Data Analysis

- Correlation



- Upon reviewing the correlation matrix, it appears that there is not a significant linear relationship between the predictor variables and the target variable.
- Examining the VIF result table below, all the values are below 10, indicating that no multicollinearity exists.

male	age	education	currentSmoker	cigsPerDay
1.198498	1.356255	1.02634	2.561237	2.699904
BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol
1.115104	1.020592	2.058312	1.582344	1.111583
sysBP	diaBP	BMI	heartRate	glucose
3.741963	2.971938	1.225777	1.096229	1.60437



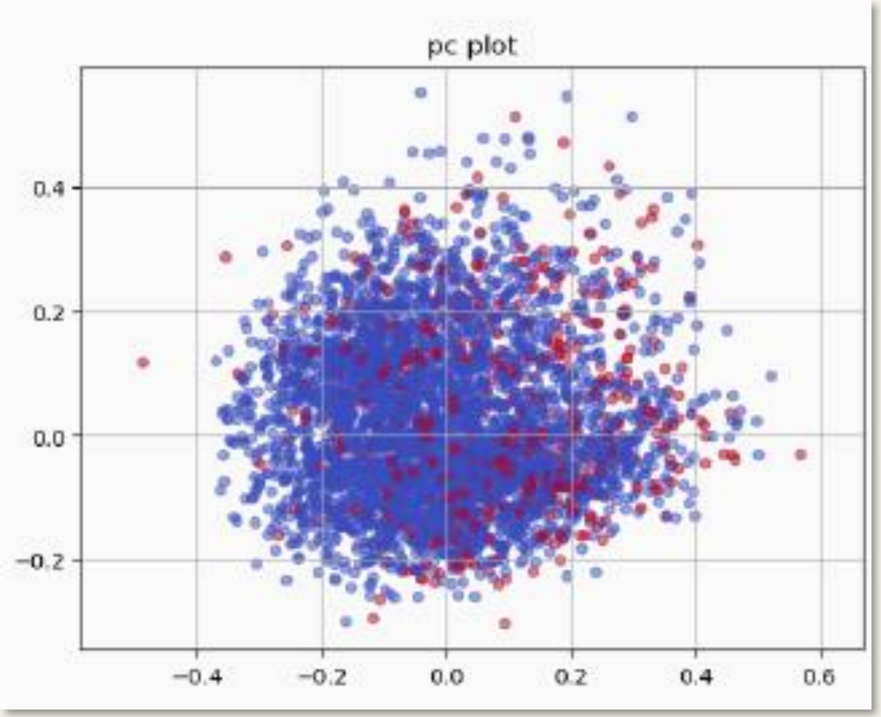
Exploratory Data Analysis

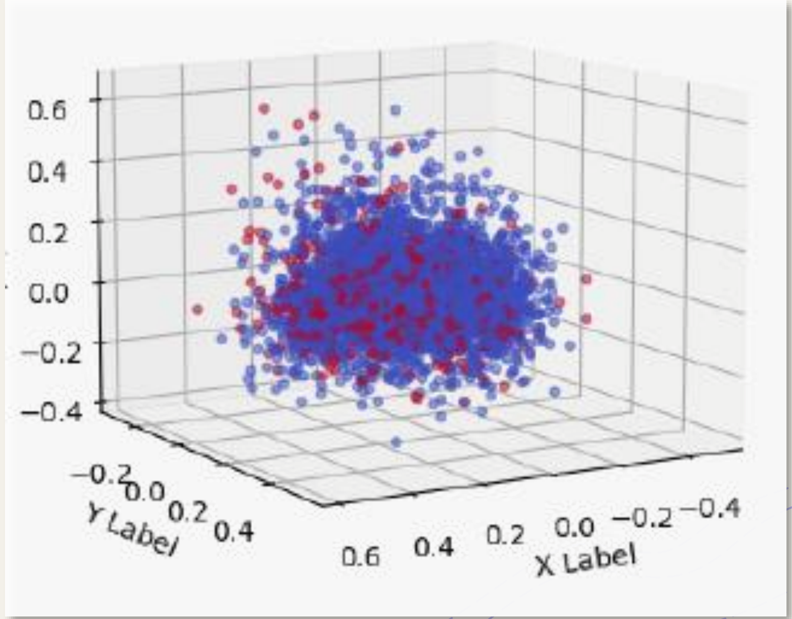
- Potential Outliers

- Looking at boxplots of numerical predictor variables, we can observe potential outliers.
- Using the z-score, 219 data points with a score over 3 and below -3 are removed, leaving 3913 data points.

Exploratory Data Analysis
- PCA Plot

- The PCA plots below display the data points projected onto principal components. The red color indicates individuals with a 10-Year CHD Risk.
- This suggests that our classification problem could be a little tricky to solve.
- 3 PCs explain 40% of variance in data





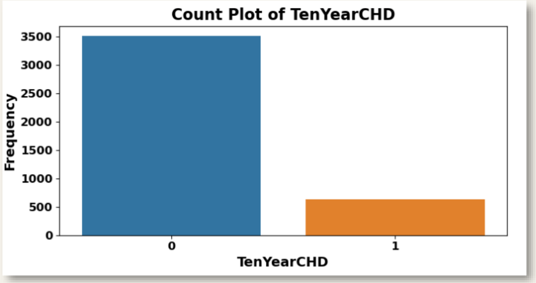
Feature Selection

+ Considering the earlier Chi-Square Test result, and the result of feature selection methods, the **final unselected variables are heartRate, education and currentSmoker**

Category	LASSO	Univariate selection (Select K Best)	Tree-based (RandomForest)	Backward Elimination	All variables
Number of selected variables	7	14	12	8	15
Unselected variable	education, currentSmoker, BPMeds, prevalentStroke, diabetes, totChol, BMI, heartRate	heartRate	BPMeds, diabetes, prevalentStroke	BMI, BPMeds, currentSmoker, diaBP, diabetes, education, totChol	-
Logistic Regression Performance	0.857854	0.862248	0.861635	0.861941	0.862069

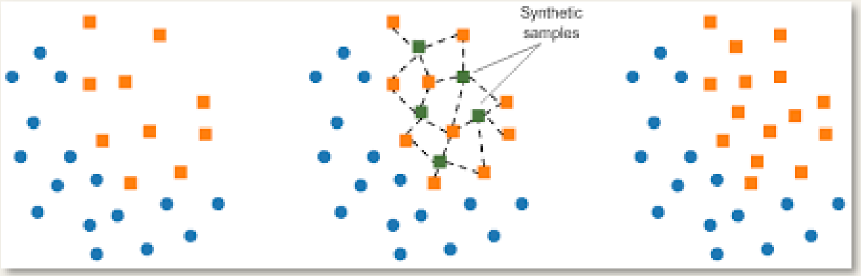
Oversampling Method

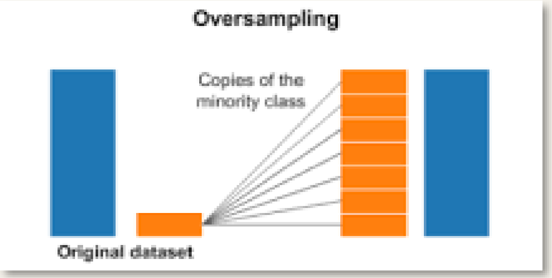
Imbalanced Data



- Earlier, as we observed, our target variable is imbalanced. This imbalance can lead to biases in our future models, such as favoring the majority class, or overfitting to the minority class due to insufficient data, etc.

SMOTE(Synthetic Minority Oversampling Technique)





- This issue can be addressed using resampling methods like undersampling and oversampling. We opted for the oversampling method, particularly SMOTE, to tackle this.
- SMOTE generates synthetic samples for the minority class to address imbalanced datasets.
- We apply SMOTE before every modeling step.

Method

1. Random Forest

- The Random Forest algorithm embodies an ensemble learning strategy, particularly leveraging the bagging technique. This method involves the generation of numerous decision trees at the time of training, which collectively determine the outcome by identifying the most frequently occurring class among the individual trees.

2. Gradient Boosting

- Gradient Boosting builds a series of decision trees over time. Instead of relying on a single tree's prediction, Gradient Boosting combines the predictions from all the individual trees it has created. Each tree's vote is weighted based on its accuracy, and the combined result is used to make the final classification. This process is a bit like asking a group of experts to weigh in on a decision and then taking an average of their opinions, with more weight given to the more knowledgeable experts.

3. Logistic Regression

- Logistic Regression is a statistical model used to solve classification problems. It predicts the probability of a sample belonging to a specific category based on given input variables. The logistic function, also known as the sigmoid function, produces an S-shaped curve and transforms input values into the range [0, 1], representing probabilities. However, the decision boundary it generates is linear. Due to this characteristic, Logistic regression outputs probabilities in binary classification and classifies samples based on a threshold.

Method

4. Support Vector Machine (SVM)

- The main objective of SVM is to find the optimal decision boundary, known as the hyperplane, with the maximum margin between classes in a high dimensional space. SVM achieves this by identifying support vectors, which are data points closest to the decision boundary. The appropriate kernel is decided by GridsearchCV.

5. Linear Discriminant Analysis (LDA)

- LDA assumes each class follows a Gaussian distribution and shares the same covariance, simplifying computation and enabling effective classification even with limited data. It uses Bayes' rule to find posteriors for each class from data points and decides the label by comparing these posteriors.

6. K-Nearest Neighbors (KNN)

- KNN examines the 'K' nearest labeled data points to the new data point requiring classification and allocates to this new point the most frequently occurring class among those neighbors. Through cross-validation, the optimal value of 'K' determined for classifying 10-Year CHD Risk is 26.

7. Multilayer Perceptron (MLP)

- In an MLP, each neuron receives inputs, computes a weighted sum along with biases, and applies an activation function to produce an output, allowing the network to capture complex patterns in the data. MLPs use the backpropagation algorithm to adjust the weights and biases to minimize the error

Modeling – Performance

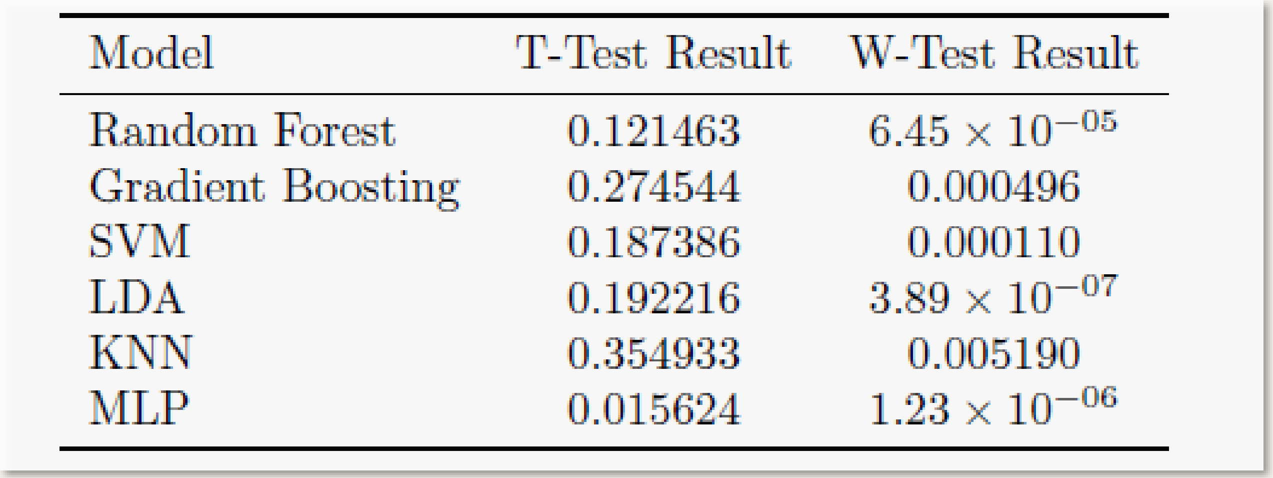
- Variables selected from the feature selection process were employed.
- Numerical variables were scaled before being fitted into models.
- SMOTE was applied before fitting into models.
- Parameters were tuned using GridSearchCV.
- Results may vary depending on performance metrics and different number of cross-validation

(Performance metric : accuracy)

Category	Random Forest	Gradient Boosting	Logistic Regression	SVM	LDA	KNN	MLP
Test Accuracy (w/o CV)	0.872232	0.873935	0.875639	0.869676	0.878194	0.869676	0.869676
Test Accuracy (w/ CV = 50)	0.860204	0.860902	0.862964	0.860596	0.860562	0.861345	0.858415

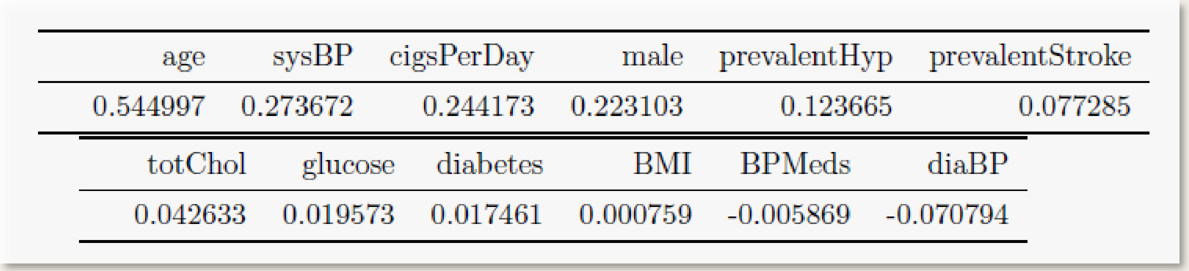
Modeling – T-Test & W-Test

- Two-sample test such as T-Test and W-Test were utilized to validate the performance of the best model, and the result indicates that the Logistic Regression shows reasonable significance compared to all other models.
- From the result, we reject the null hypothesis of two models having similar performance because at least the result from W-Test clearly reject null hypothesis.



Model	T-Test Result	W-Test Result
Random Forest	0.121463	6.45×10^{-05}
Gradient Boosting	0.274544	0.000496
SVM	0.187386	0.000110
LDA	0.192216	3.89×10^{-07}
KNN	0.354933	0.005190
MLP	0.015624	1.23×10^{-06}

Model Selection & Interpretation



age	sysBP	cigsPerDay	male	prevalentHyp	prevalentStroke	totChol	glucose	diabetes	BMI	BPMeds	diaBP
0.544997	0.273672	0.244173	0.223103	0.123665	0.077285	0.042633	0.019573	0.017461	0.000759	-0.005869	-0.070794

- Logistic Regression yielded the best performance among all models.
- The findings highlight ‘age’ as the predominant factor influencing the 10-year risk of Coronary Heart Disease (CHD), with ‘sysBP’, ‘cigsPerDay’, ‘male’, and ‘prevalentHyp’ also playing significant roles.
- Notably, ‘cigsPerDay’ is a factor within an individual’s control, suggesting that reducing cigarette consumption or quitting smoking altogether could effectively lower the risk of 10-years risk of CHD.
- Additionally, understanding that lowering ‘sysBP’ can also reduce the 10-year risk of CHD, individuals can benefit from physical activities aimed at decreasing systolic blood pressure, potentially further reducing the risk of CHD.
- From the perspective of healthcare businesses, this presents an opportunity to elevate awareness regarding the link between smoking, systolic blood pressure and CHD. This could allow them to enhance offerings in anti-smoking campaigns and physical activities.
- Additionally, focusing efforts more towards male demographics could be beneficial, underlining a potential association between gender and CHD risk.

Challenges

+ Imbalanced Data

: As the proportion of our minority class was significantly low, we decided to use oversampling method. We chose SMOTE because unlike other oversampling techniques SMOTE does not simply duplicate or repeat existing samples of the minority class. This helps to learn a more generalized decision boundary. After using SMOTE the performance increased around 5%.

+ Relatively low performance in other metrics

: Although SMOTE increased our performance metric 'accuracy' in general, the other performance metrics gives low score. The number of true positive from confusion matrix was low, implying that possibly models high accuracy comes from true negative. We could consider other metrics focused on true positive such as F1-score or Recall.

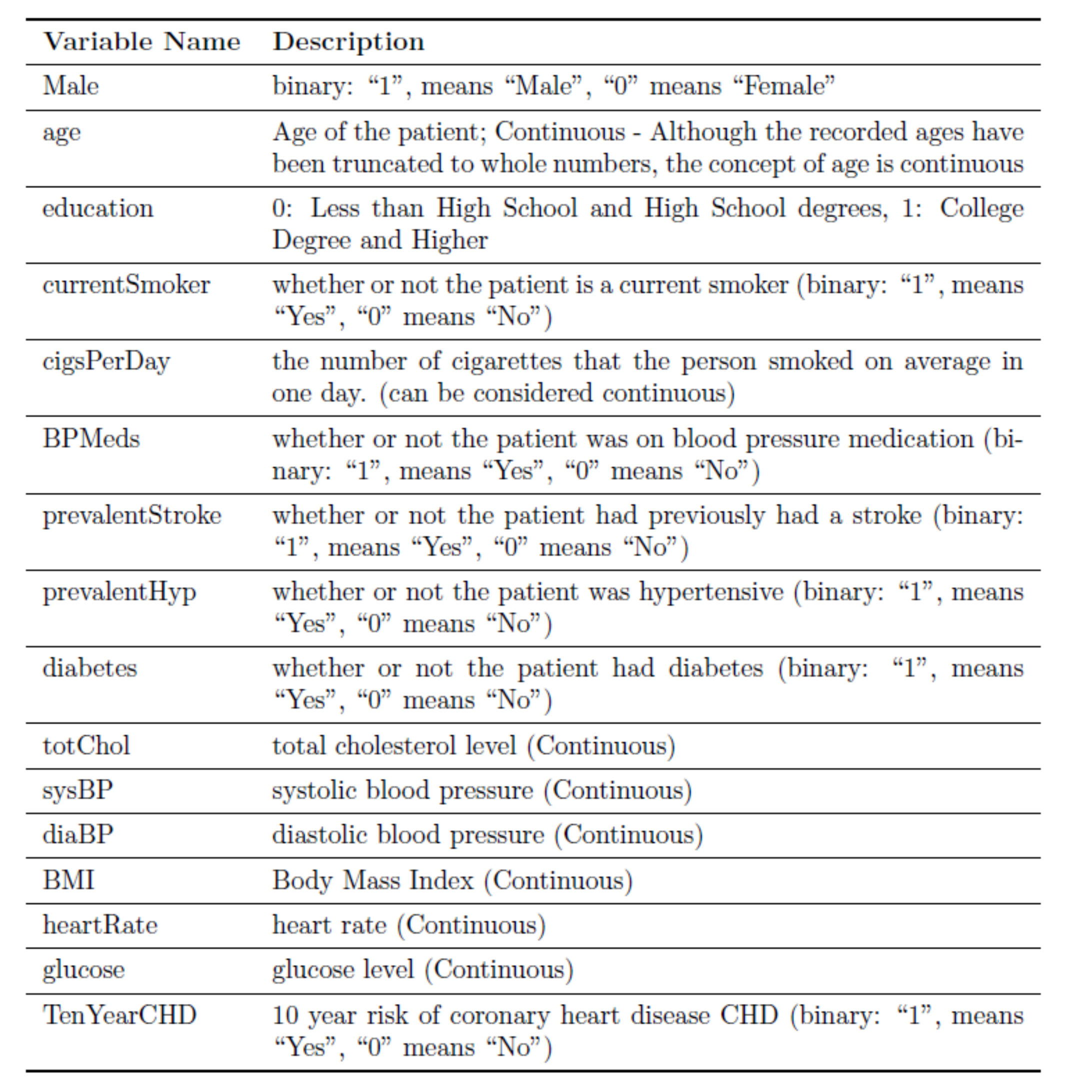
Conclusion

+ With high accuracy achieved, our models possibly can help healthcare providers, insurance companies, and patients make more informed decisions about managing and reducing the risk of CHD. This can ultimately lead. to reduced healthcare costs and improved patient health management

+ However, with the limited number of attributes, our model still has its limitations in classifying patients with a 10-year risk of CHD. Additionally, based on the results of coefficient degrees, we can assume that not one single factor is influencing the 10-year risk of CHD. Therefore, further research will be needed to explore various attributes that can potentially impact CHD risk to build a more robust model

+ In future work, with the possible addition of new attributes and more data, we can explore ways to leverage data and enhance machine learning techniques to further advance our goals.

Appendix.



Variable Name	Description
Male	binary: “1”, means “Male”, “0” means “Female”
age	Age of the patient; Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous
education	0: Less than High School and High School degrees, 1: College Degree and Higher
currentSmoker	whether or not the patient is a current smoker (binary: “1”, means “Yes”, “0” means “No”)
cigsPerDay	the number of cigarettes that the person smoked on average in one day. (can be considered continuous)
BPMeds	whether or not the patient was on blood pressure medication (binary: “1”, means “Yes”, “0” means “No”)
prevalentStroke	whether or not the patient had previously had a stroke (binary: “1”, means “Yes”, “0” means “No”)
prevalentHyp	whether or not the patient was hypertensive (binary: “1”, means “Yes”, “0” means “No”)
diabetes	whether or not the patient had diabetes (binary: “1”, means “Yes”, “0” means “No”)
totChol	total cholesterol level (Continuous)
sysBP	systolic blood pressure (Continuous)
diaBP	diastolic blood pressure (Continuous)
BMI	Body Mass Index (Continuous)
heartRate	heart rate (Continuous)
glucose	glucose level (Continuous)
TenYearCHD	10 year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)