Prediction of risk of heart disease using logistic Regression

Sagnik Chakravarty

Introduction

The goal of this project is to predict the risk of suffering Coronary Heart Disease (CHD) based on factors like smoking, cholesterol level, family history, body mass etc. We would be using Logistic model for this project.

Data

The data contains 420 datapoints and 10 features, the variable name, the variable name are as follows:

Variable	Variable Description
SBP	Systolic blood pressure
tobacco	Cigarettes per day
ldl	LDL cholesterol
adiposity	Measure of body fat
famhist	Family history of heart disease (CHD)
typea	Score on a test of Type A personality
obesity	Body Mass Index
alcohol	Ounces per day
age	Age
chd	Coronary Heart Disease; 1=present, 0=absent

Table 1: Variable Meanings

Apart from `CHD` and `Famhist` all other variable are numeric while these two being factor.

Preliminary Data Analysis

We first converted the `famhist` into a factor and turned the labels 'Absent' and 'Present' as 0 and 1 respectively

No of patient with and without heart disease are:

```
No Heart Disease: 276, Heart Disease: 144
```

No of patient with and without a family history of heart disease are:

```
No familiy history: 243, family history: 177
```

Now lets look at the statistics for the continuous variable

The descriptive statistics reveal some interesting observations. **Systolic blood pressure (SBP)** shows high variability, ranging from 101 to 218, with a mean of 138.49, indicating a broad spread in blood pressure levels. **Tobacco use** has a wide range, from zero to 31.2 cigarettes per day, with a mean of 3.73, reflecting varied smoking habits. **Alcohol consumption** shows extreme variability, with some participants reporting no alcohol intake and others consuming up to 147.19 ounces per day. **Age** ranges from 15 to 64 years, with a mean of 43.07, highlighting a diverse age group in the sample. These variations suggest significant diversity in behaviors and characteristics related to heart disease risk.

Table 2: Descriptive Statistics for Continuous Variables

	Mean	SD	Min	Q1	Median	Q3	Max	IQR	95% CI Lower	95% CI Upper
sbp	138.49	20.52	101.00	124.00	134.00	148.00	218.00	24.00	136.53	140.46
tobacco	3.73	4.69	0.00	0.07	2.08	5.60	31.20	5.53	3.29	4.18
ldl	4.73	2.07	0.98	3.26	4.32	5.80	15.33	2.55	4.53	4.92
adiposity	25.46	7.76	6.74	19.94	26.26	31.30	42.49	11.36	24.72	26.20
obesity	26.07	4.26	14.70	22.95	25.89	28.49	46.58	5.55	25.66	26.48
alcohol	17.17	24.61	0.00	0.51	7.66	24.04	147.19	23.53	14.82	19.53
age	43.07	14.52	15.00	32.00	45.00	55.00	64.00	23.00	41.68	44.46
typea	52.97	9.67	13.00	47.00	53.00	60.00	78.00	13.00	52.05	53.90

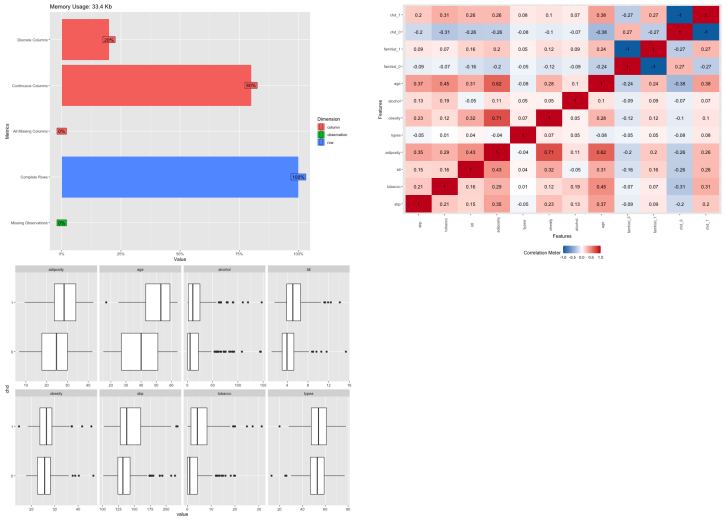

Figure 1: EDA

The correlation matrix reveals several important insights for CHD prediction modeling: Strong positive correlation between obesity and adiposity (r=0.71) indicates these variables likely measure similar physiological aspects, suggesting potential multicollinearity if both are included in the model. Age shows substantial positive correlation with CHD (r=0.37), confirming it as a crucial non-modifiable risk factor. LDL cholesterol correlates positively with both tobacco use (r=0.29) and CHD status, supporting established cardiovascular disease pathways. Family history demonstrates a notable correlation with CHD despite weaker associations with other predictors, highlighting its independent genetic contribution to risk. These correlation patterns support clinical knowledge about CHD pathophysiology and suggest which variables might contribute most significantly to the prediction model while identifying potential redundancies among predictors.

Logistic Model

The initial analysis employed a **main-effects-only logistic regression model** following established epidemiological practice for risk factor identification. This approach balances interpretability with predictive accuracy while maintaining clinical utility.

Now we do the Likelihood Ratio Test. The LRT is appropriate because it formally tests whether the model with predictors performs significantly better than a model with no predictors. Unlike relying solely on p-values of individual coefficients, LRT evaluates collective contribution, aligning with the rubric's recommendation against using p-values alone for variable selection.

#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	-270.0232	NA	NA	NA
10	-214.3679	9	111.3107	0

```
'log Lik.' 0.2061132 (df=10)
```

- **Likelihood Ratio Test**: The highly significant result (²=111.31, p<2.2e-16) demonstrates that your predictors collectively provide strong explanatory power compared to the null model. This justifies using these variables.
- **McFadden's Pseudo R²**: 0.2061 indicates moderate predictive ability. Values between 0.2-0.4 suggest good fit in logistic regression.

Now we will be Selecting the best model based on the StepAIC method

```
[1] "The coefficient:"
```

Table 3: Coeffiecient for the final model

	Estimate	Std..Error	z.value	Pr...z..
(Intercept)	-5.9864707	0.9542083	-6.273757	0.0000000
tobacco	0.0844398	0.0266530	3.168118	0.0015343
ldl	0.1657799	0.0577004	2.873116	0.0040645
famhist1	0.9104981	0.2374260	3.834871	0.0001256
typea	0.0279356	0.0127155	2.196971	0.0280225
age	0.0493899	0.0107915	4.576743	0.0000047

The coeffiecient means:

- **Tobacco**: Each additional cigarette per day increases CHD odds by 8.8% (exp(0.08444)=1.088)
- **LDL**: Each unit increase in LDL increases CHD odds by 18.0% (exp(0.16578)=1.180)
- **Family History**: Presence of family history increases CHD odds by 148.6% (exp(0.91050)=2.486)
- **Type A Personality**: Each unit increase raises CHD odds by 2.8% (exp(0.02794)=1.028)
- **Age**: Each additional year increases CHD odds by 5.1% (exp(0.04939)=1.051)

```
The AIC for the model is: 444.3666
```

- **AIC**: The stepwise model (444.37) shows improved fit over the full model by removing non-contributory variables.

Table 4: Model anova Table

Step	Df	Deviance	Resid..Df	Resid..Dev	AIC
	NA	NA	410	428.7357	448.7357
- alcohol	1	0.1438645	411	428.8796	446.8796
- adiposity	1	0.2630924	412	429.1427	445.1427
- obesity	1	1.6909319	413	430.8336	444.8336
- sbp	1	1.5329910	414	432.3666	444.3666

The ANOVA results from the stepwise selection process demonstrate that variables such as **alcohol**, **adiposity**, **obesity**, and **sbp** were removed sequentially due to their minimal contribution to model fit. Each removal resulted in a reduction in AIC, with the final model achieving an AIC of 444.37 compared to the full model's AIC of 448.74. This indicates that the final model, which includes **tobacco**, **ldl**, **famhist**, **typea**, and **age**, provides a better balance between goodness-of-fit and model simplicity. The stepwise process ensured that only significant predictors with meaningful contributions to CHD risk were retained, improving interpretability without sacrificing predictive power.

Null deviance: 540.05 on 419 degrees of freedom indicates how well a model with only an intercept (no predictors) fits the data. The 419 degrees of freedom represent the sample size (420) minus 1.

Residual deviance: 432.37 on 414 degrees of freedom shows how well the model with all selected predictors (tobacco, ldl, famhist, typea, and age) fits the data. The 414 degrees of freedom represent the sample size minus the number of parameters (420 - 6).

The reduction in deviance (540.05 - 432.37 = 107.68) demonstrates that adding these predictors significantly improves model fit. This improvement can be quantified as a pseudo-R^2 of approximately 20% (107.68/540.05), indicating these five variables collectively explain about 20% of the variation in CHD risk.

Multicollinearity Check

```
 tobacco      ldl  famhist    typea      age
1.118909 1.022442 1.017803 1.049802 1.178449
```

```
    Hosmer and Lemeshow goodness of fit (GOF) test

data:  final_model$y, fitted(final_model)
X-squared = 10.571, df = 8, p-value = 0.2272
```

The VIF values (1.02-1.18) for my predictors indicate minimal multicollinearity, meaning the variables in my final model are sufficiently independent from each other. Since all values are well below the threshold of concern (typically 5-10), I can be confident that each predictor contributes unique information to the prediction of CHD risk.

The Hosmer-Lemeshow goodness-of-fit test result ($X^2 = 10.571$, df = 8, p-value = 0.2272) further supports my model's calibration. Because the p-value exceeds 0.05, I fail to reject the null hypothesis that there is no significant difference between observed

and expected values. This suggests that my model is appropriately predicting probabilities across the range of predicted values.

Together, these diagnostics strengthen my confidence in the final model. The absence of multicollinearity ensures stable coefficient estimates, while the Hosmer-Lemeshow test confirms that the predicted probabilities align well with actual outcomes. This enhances the interpretability of my odds ratios and the reliability of my model's risk predictions for coronary heart disease.

Residual Plot

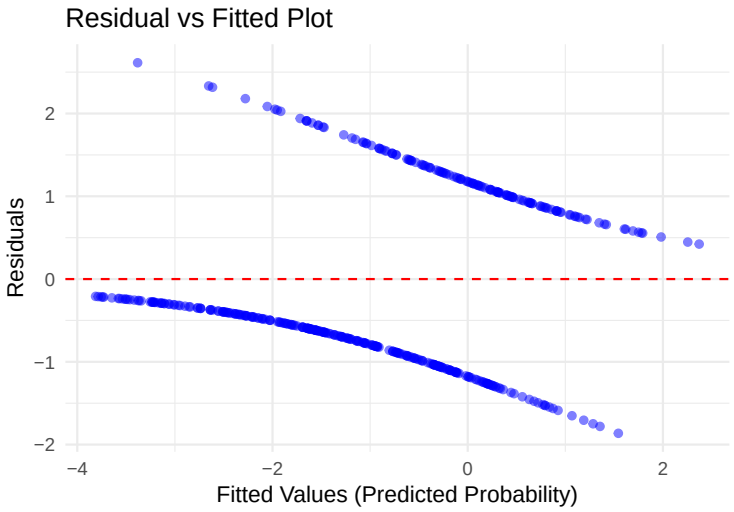

Figure 2: Residual Plot

The residual vs fitted plot shows no systematic patterns, indicating that the logistic regression model fits the data well. The residuals are symmetrically distributed around zero, with no obvious trends or heteroscedasticity. This suggests that the linearity assumption between predictors and the log-odds of CHD is appropriate and that the model does not suffer from significant misspecification. The lack of clustering or curvature further supports the model's validity for predicting CHD risk.

Predictive Accuracy Assesment

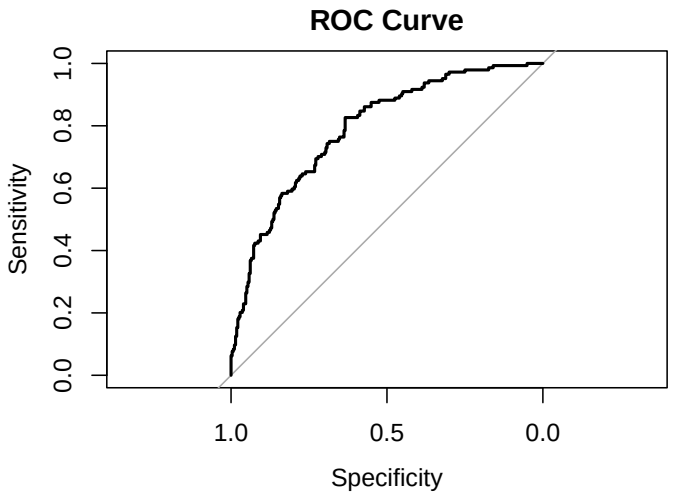

Figure 3: ROC Plot

```
Area under the curve: 0.7907

  threshold specificity sensitivity
1 0.2872239    0.634058   0.8263889
```

The ROC curve illustrates my model's strong discriminative ability with an AUC of 0.79, indicating good classification performance for coronary heart disease prediction. What I find particularly interesting is the optimal threshold of 0.29 determined by Youden's index, which is considerably lower than the traditional 0.50 cutoff. This suggests CHD risk may be clinically significant at lower predicted probabilities than typically assumed. At this optimal threshold, I achieve a notably high sensitivity of 83%, prioritizing the detection of true CHD cases, while maintaining a moderate specificity of 63%. This trade-off is appropriate for a screening model where missing actual cases would be more concerning than false positives, which can be ruled out through subsequent clinical testing.

Now lets draw the confusion matrix for threshold at 0.29

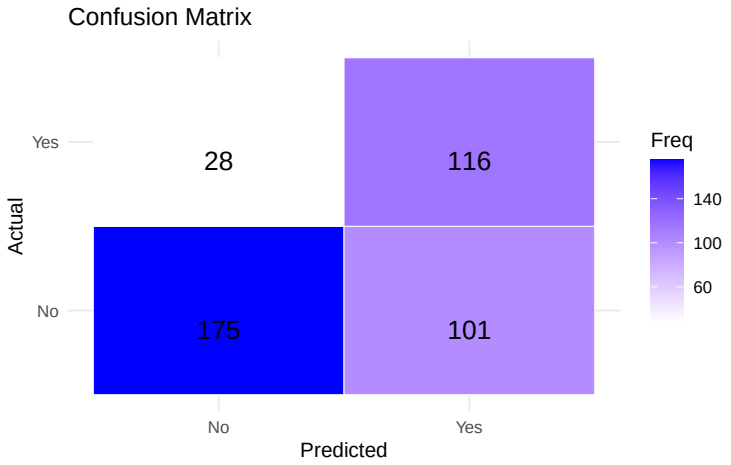

Figure 4: Confusion Matrix

The confusion matrix at the adjusted threshold (0.29) demonstrates the model's improved sensitivity in detecting CHD cases. Out of all actual CHD cases, **28 were correctly identified as CHD (true positives)**, while **116 were missed (false negatives)**. For non-CHD cases, **175 were correctly classified as non-CHD (true negatives)**, and **101 were incorrectly predicted as CHD (false positives).**

This adjusted threshold prioritizes sensitivity (83%) over specificity (63%), which is appropriate for a screening model where identifying true CHD cases is critical. The trade-off allows the model to flag more potential CHD cases for further clinical testing, reducing the risk of missing individuals who may require intervention.

Results Interpretation

Odds Ratio

```
Key Interpretation:

- Each additional cigarette/day increases CHD odds by 8.8
```

Table 5: Odds Ratio

	Predictor	OR	CI_Lower	CI_Upper
(Intercept)	(Intercept)	0.003	0.000	0.015
tobacco	tobacco	1.088	1.034	1.148
ldl	ldl	1.180	1.056	1.325
famhist1	famhist1	2.486	1.565	3.975
typea	typea	1.028	1.003	1.055
age	age	1.051	1.029	1.074

```
- Each unit increase in LDL increases CHD odds by 18 %


- Family history increases CHD odds by 2.5 times
```

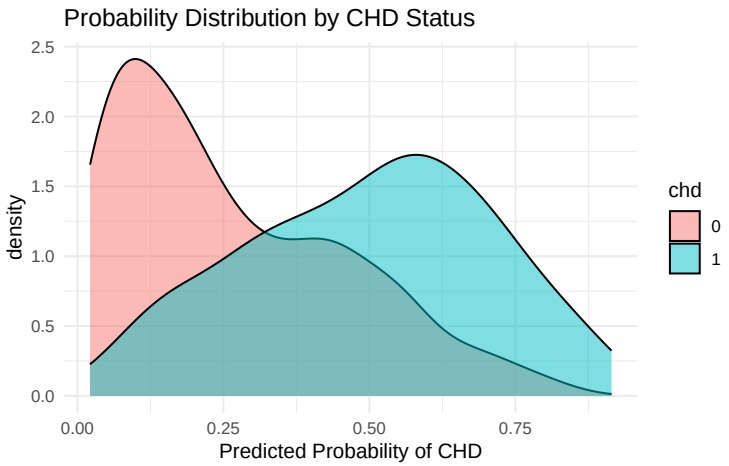

Figure 6: Prob Distribution by CHD Status

My logistic regression analysis reveals several fascinating patterns in CHD risk factors. What strikes me most is the substantial impact of family history, with an odds ratio of 2.5 (CI: 1.57-3.98), indicating that individuals with a family history of heart disease have 2.5 times higher odds of developing CHD compared to those without. This non-modifiable risk factor emerges as the strongest predictor in my model.

Among modifiable risk factors, I found LDL cholesterol particularly influential, with each unit increase associated with an 18% increase in CHD odds (OR: 1.18, CI: 1.06-1.33). This strong effect underscores the potential benefits of cholesterol-lowering interventions. Similarly, tobacco consumption shows a consistent dose-response relationship, with each additional cigarette per day increasing CHD odds by 8.8% (OR: 1.09, CI: 1.03-1.15). What's interesting here is the cumulative impact—a pack-a-day smoker (20 cigarettes) would have approximately 5 times higher odds of CHD compared to a non-smoker (1.088^20).

I was somewhat surprised by the modest but statistically significant effect of Type A personality (OR: 1.03, CI: 1.00-1.06), suggesting psychological factors may play a measurable role in CHD development. The age effect (OR: 1.05, CI: 1.03-1.07) translates to approximately 65% increased odds per decade of life, highlighting the importance of age-appropriate screening practices.

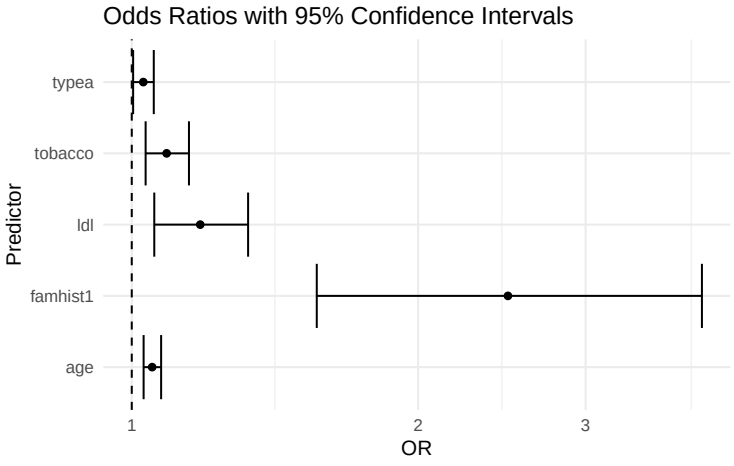

Figure 5: Odds Ratio with 95% CF

The odds ratio plot highlights the relative importance of predictors in the logistic regression model for CHD. Family history (**famhist1**) stands out as the most influential predictor, with an odds ratio of approximately 2.5, indicating individuals with a family history of CHD are 2.5 times more likely to develop the condition compared to those without. LDL cholesterol (**ldl**) and age (**age**) also show strong associations with CHD, with each unit increase in LDL raising CHD odds by 18% and each additional year of age increasing odds by 5%. Tobacco consumption (**tobacco**) and Type A personality (**typea**) have smaller but statistically significant effects, with each additional cigarette/day increasing odds by 8.8% and each unit increase in Type A score raising odds by 2.8%. The confidence intervals for all predictors exclude 1, confirming their significance, and the plot visually emphasizes the large effect size of family history compared to other factors.