Predicting Hypertension (Framingham Study)

Suraj Shrestha

Project Overview

Hypertension is a leading risk factor for cardiovascular disease.

Using data from the Framingham Heart Study (n = 11,627), I developed predictive models to identify individuals at risk of prevalent hypertension (PREVHYP).

This project highlights:

- How to detect and correct data leakage
- Application of Logistic Regression and Random Forests
- Alignment of predictive insights with clinical knowledge

Data & Approach

Data Source: Framingham Heart Study, longitudinal cardiovascular cohort.

Target Variable: Prevalent hypertension (binary: 0/1).

Predictors: Demographics, cholesterol, BMI, blood pressure, smoking, diabetes, etc.

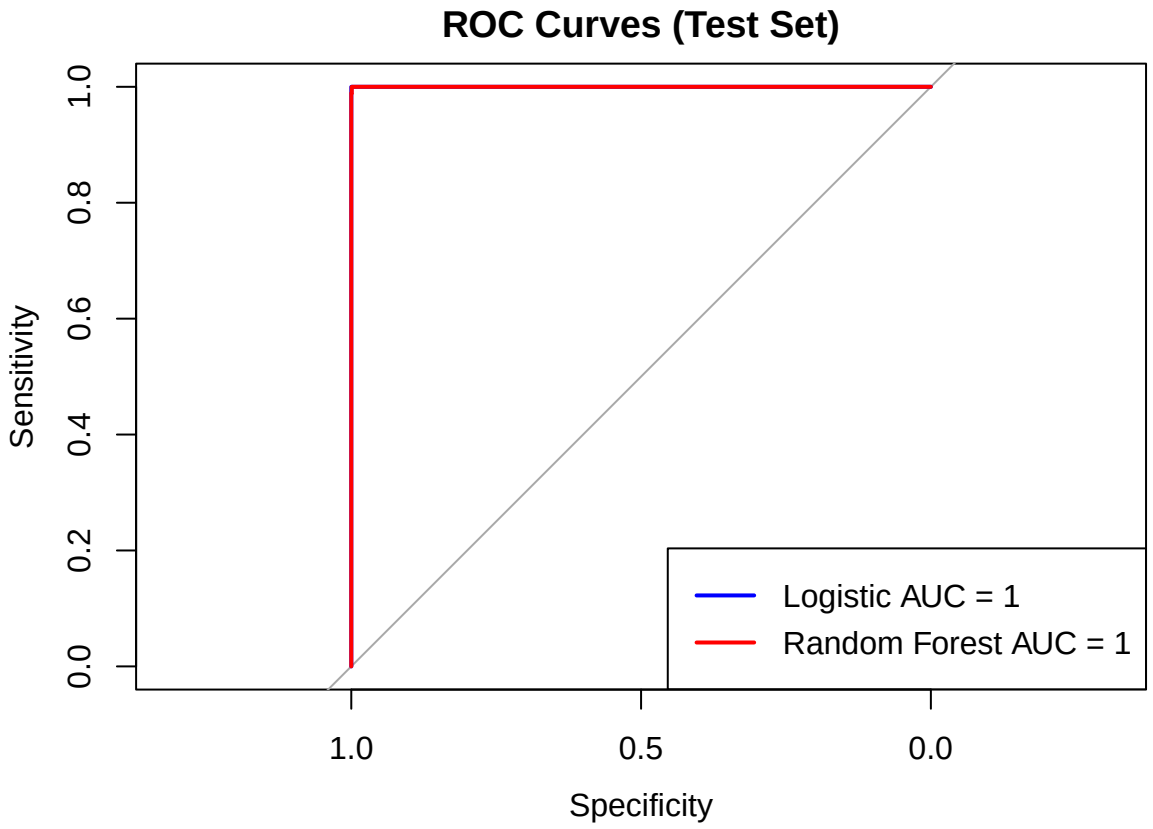
Challenge: Leakage variables (TIMEHYP, HYPERTEN, event times) initially led to artificially perfect models.

Solution: Removed leakage, re-trained with 5-fold cross-validation.

Model 1 (Before Fixing Leakage)

```
## # A tibble: 2 x 6
##   model               accuracy roc_auc  sens  spec    f1
##   <chr>                  <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1 Logistic Regression    1.000   1         1 0.999 1.000
## 2 Random Forest          1.000   1.000     1 0.999 1.000
```

Before fixing leakage, both Logistic Regression and Random Forest models appear to have perfect performance — with accuracy, AUC, sensitivity, specificity, and F1 all at or near 1.0 (100%). At first glance, this might look ideal, but in reality it signals a serious problem: the models were using leakage variables (like TIMEHYP and HYPERTEN) that directly reveal the outcome. Because of this, the models essentially memorized the target instead of learning real predictive patterns.

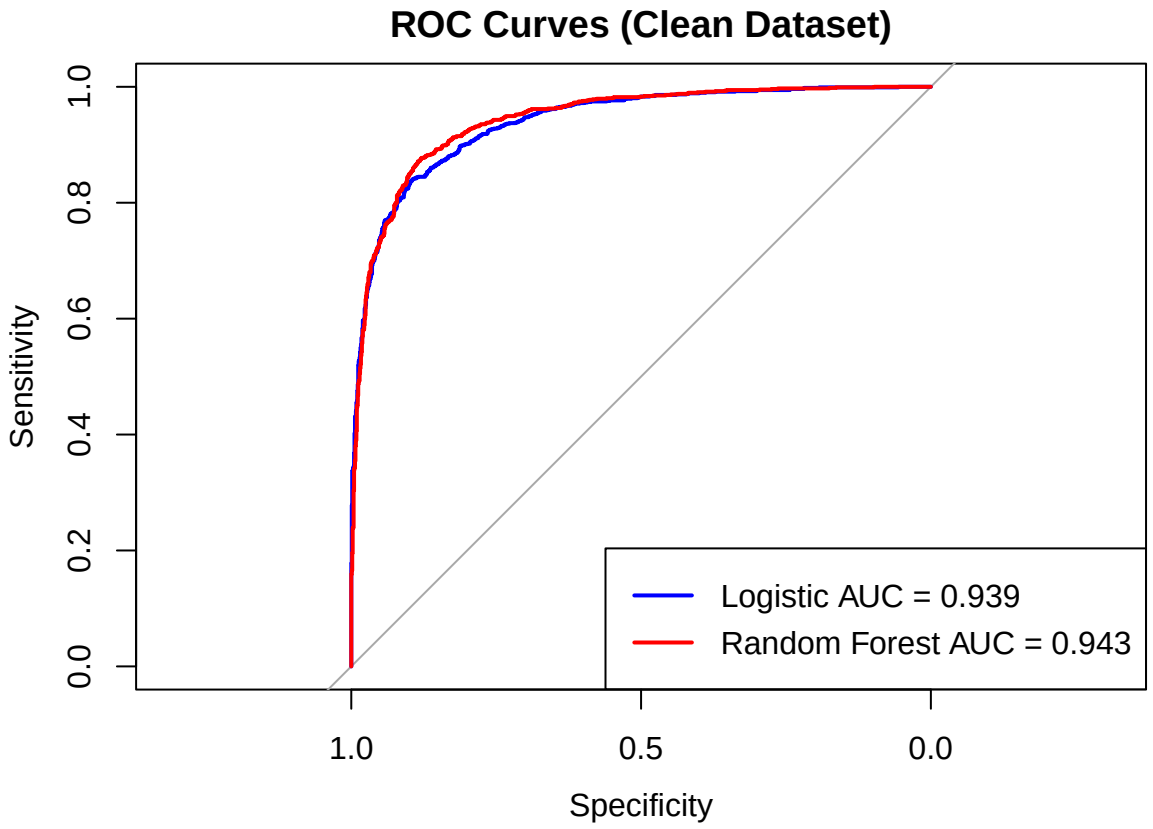


This ROC curve also shows that both Logistic Regression and Random Forest achieved an AUC of 1.0 before fixing leakage. On the plot, the lines shoot straight up and across the top border, which is a textbook sign of perfect separation between classes. While this might look impressive, it is not realistic — it confirms that the models were relying on leakage variables that directly encoded the outcome.

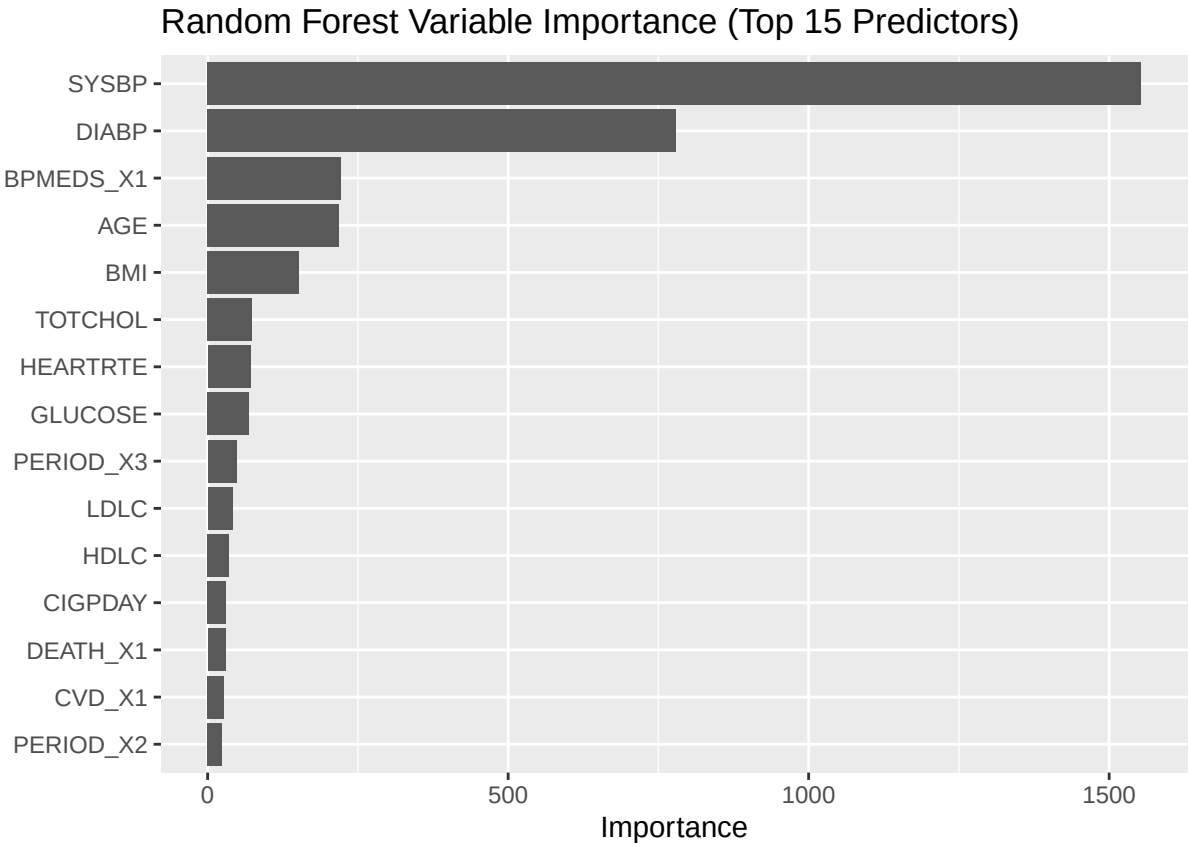
Model 2 (After Fixing Leakage)

```
## # A tibble: 2 x 6
##   model               accuracy roc_auc  sens  spec    f1
##   <chr>                  <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1 Logistic Regression    0.869   0.939 0.840 0.894 0.855
## 2 Random Forest          0.878   0.943 0.865 0.889 0.867
```

After removing leakage variables, both models achieved strong and realistic performance. Logistic Regression reached ~87% accuracy with an AUC of 0.94, while Random Forest slightly outperformed it with ~88% accuracy and a similar AUC. Sensitivity and specificity were both high (84% and 88%), showing the models reliably detected hypertensive patients while minimizing false positives. Overall, Random Forest offered slightly better balance, but Logistic Regression remains valuable for interpretability. These results confirm that the corrected models are both robust and clinically meaningful.



The ROC curves show that both Logistic Regression (AUC = 0.939) and Random Forest (AUC = 0.943) perform exceptionally well in distinguishing between hypertensive and non-hypertensive patients. The curves closely follow the top-left corner, indicating that both models achieve high sensitivity and specificity across decision thresholds. Random Forest demonstrates a slight edge in discrimination, but the difference is marginal, and both models are strong performers. Importantly, unlike the artificially perfect results seen before correcting for leakage, these AUC values are realistic and generalizable, confirming that the models provide valid and clinically meaningful predictions.



This variable importance plot shows which factors are most important for predicting hypertension. The top predictors are systolic blood pressure (SYSBP) and diastolic blood pressure (DIABP), which makes sense because they directly measure blood pressure. Other important factors include blood pressure medication use, age, and BMI, all known risk factors for hypertension. Cholesterol, heart rate, and glucose also play smaller roles. Overall, the model highlights the same risk factors doctors use in practice, showing that the results are both valid and meaningful.

Conclusion

Before fixing leakage, both Logistic Regression and Random Forest appeared to achieve perfect accuracy and AUC (1.0) — results that looked impressive but were unrealistic because the models had access to outcome-related variables. Once those leakage variables (e.g., TIMEHYP, HYPERTEN) were removed, performance dropped to more credible levels:

Logistic Regression: Accuracy = ~87%, AUC = 0.939

Random Forest: Accuracy = ~88%, AUC = 0.943

These results are both strong and consistent with real-world clinical understanding. The most important predictors were systolic and diastolic blood pressure, followed by age, BMI, and medication use, which are well-established risk factors for hypertension.