The Importance of Context in Model Selection

Predicting Heart Disease: A Comparison of Machine Learning Algorithms and Transformation Techniques

Khushnur Binte Jahangir, Adam Broniewski

What is the Challenge?

Predict if someone has heart disease based on 13 measures of health.

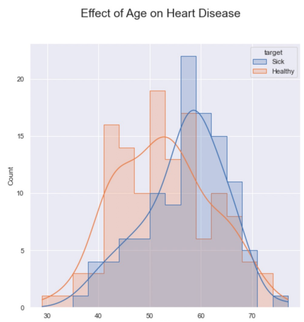
Apply various data preparation, data transformation, and machine learning techniques. Choose the best methods given the **data context.**

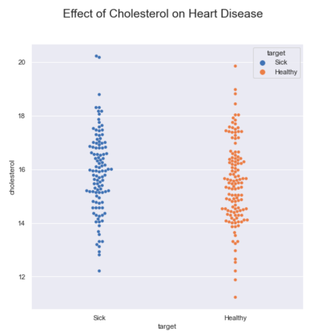
The dataset:

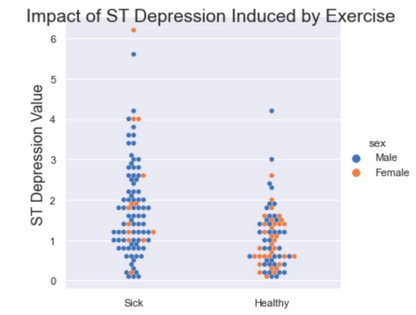
- 303 instances:
- 13 attributes:
  - Age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting electrocardiographic results, max heart rate after exercise, exercise induced angina, ST depression induced by exercise, # of major vessels visible, thalium stress
- Target: 139 with and 164 without heart disease.

Exploring the Data

- No correlation:
  - fasting blood sugar, cholesterol and resting blood pressure
- Positive correlations:
  - Age (especially in men), ST depression seen after exercise
- Negative correlations:
  - maximum heart rate achieved after exercise







Model Building

Four models were tested:

- Logistic Regression
- Decision Tree
- Random Forest
- Neural Network

Model tuning was done applying different combinations of parameters using a grid search. Computationally expensive models used **a random gird search for breadth** of selection, and smaller ranges in **regular grid search for fine tuning.**

Each model was tested using:

1. Full training dataset
2. Feature selected dataset

Validation: Using Context to Identify the Best Approach

Model	F-score	Precision	Recall	Accuracy
Neural Network	0.831	0.839	0.829	0.834
Logistic Regression	0.826	0.836	0.826	0.830

We compare our **two top performers** using four metrics. F-score showed average of precision and recall, but each metric was considered independently. False negatives can have life altering consequences.

Logistic regression chosen for interpretability. Small increase in performance of neural network not deemed worthy of loss of contextual information.

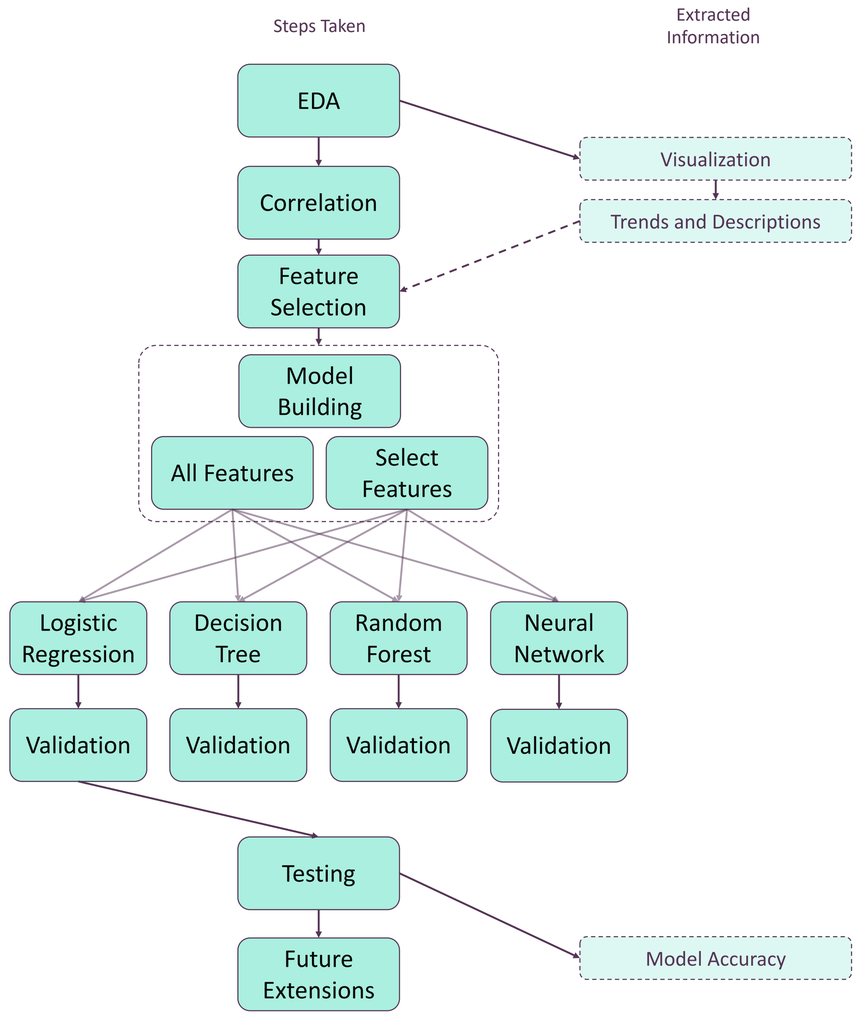
Testing the Best: Results

Logistic regression predications on test data generally had similar results to validation. Precision decreased and recall increased.

Model	F-score	Precision	Recall	Accuracy
Test Results	0.836	0.781	0.893	0.836
Validation Results	0.826	0.836	0.826	0.830

Evaluating F-score or accuracy alone would not have revealed change in performance.

Testing Different Approaches



Trying Again: Future Extensions and Issues

- **Amount of Data:** Limited dataset made multiple layers of stratification impractical.
- **Different Folds:** 5-fold (instead of 10-fold) means fewer instances in model building. Try different validation methods for neural network and random forest.
- **Feature Selection:** Try one-hot encoding ordinal features, as severity change between steps is unknown. Use random forest to programmatically select features.
- **Regularization:** Given success of feature selection, apply lasso or ridge method to vary the weight of features.