BAI Data Science Case Study (Heart)

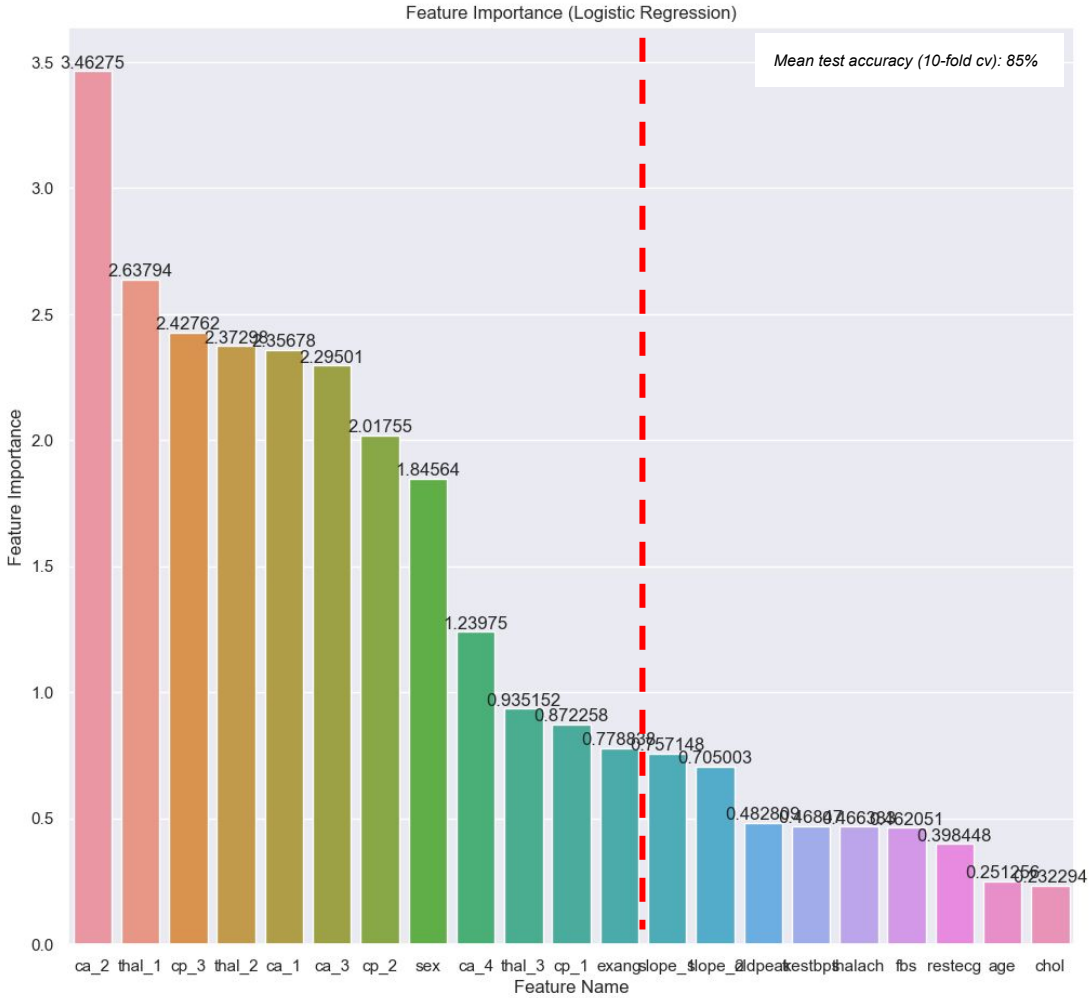
Keith Dowd

Executive Summary

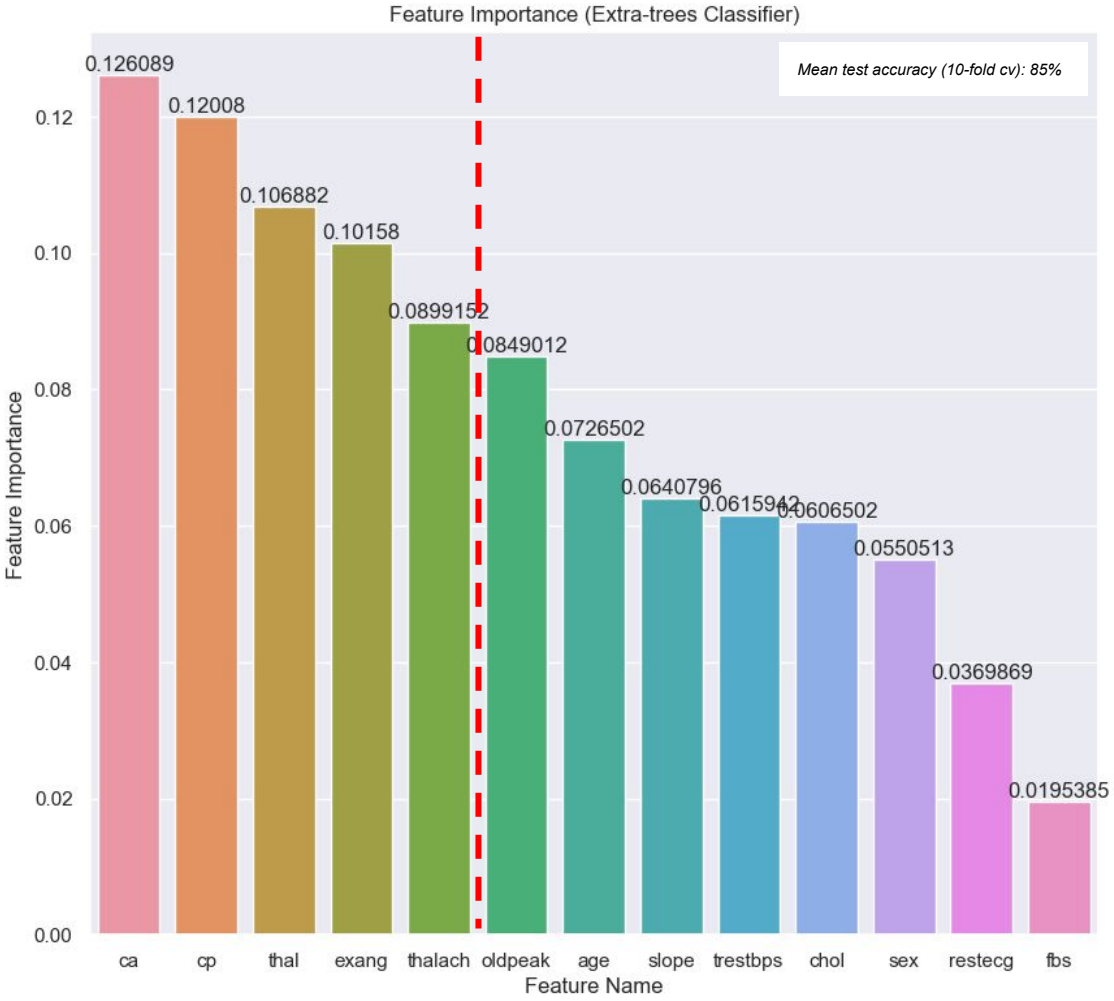
- Across three different modeling approaches, the following contributing factors were consistently identified as the most important factors for predicting the presence/absence of heart disease:
  - Chest pain type (*cp*)
  - Number of major vessels (0-3) colored by fluoroscopy (*ca*)
  - Defect type (*thal*)
  - Exercise induced angina (*exang*)

Methodology

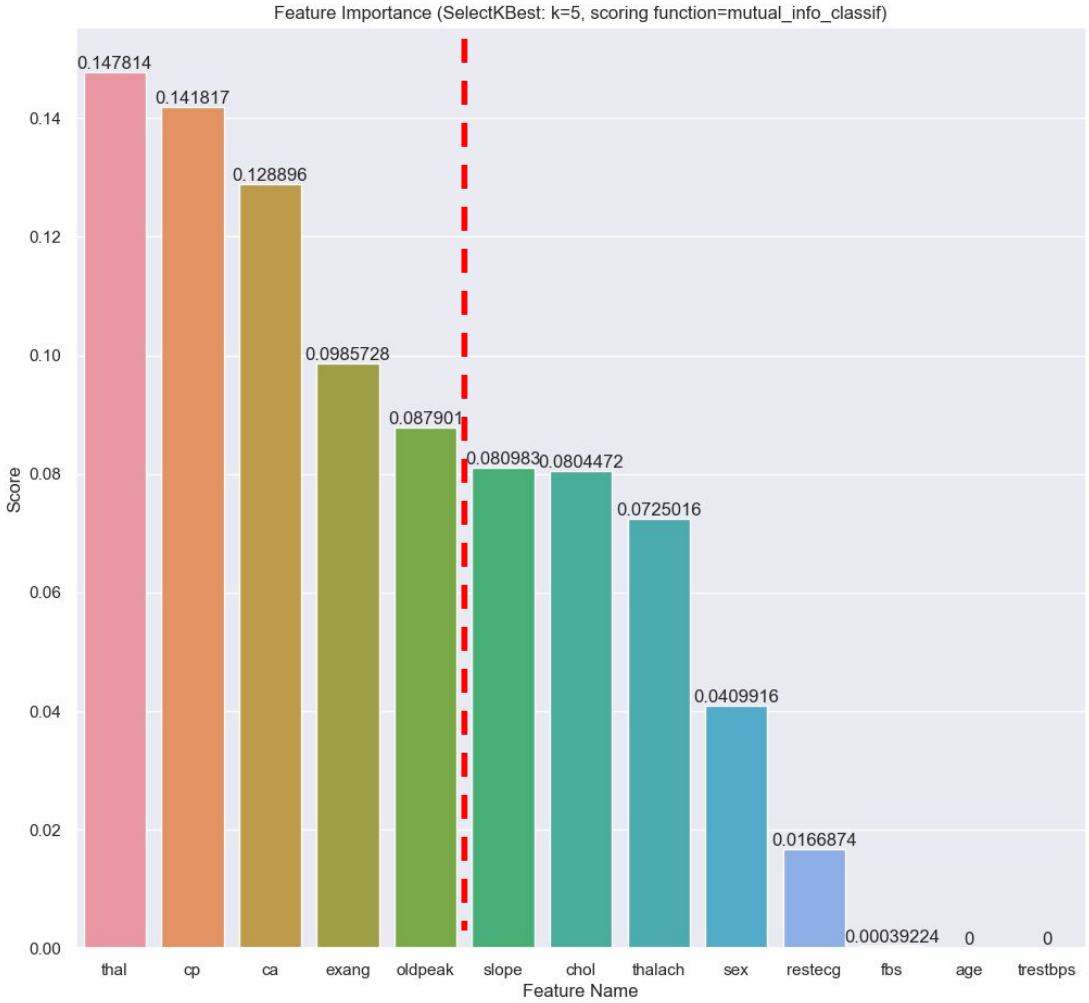
1. Review descriptive statistics for the data:
    - Of particular interest are the correlations between the predictors and target.
2. Fit data to a variety of models:
    - Use results to select two modeling approaches.
3. Select two different modeling approaches for further investigation:
    - In this case, logistic regression and extra-trees classifier.
    - Both report similar model fit scores.
    - Each uses a different approach (regression vs. trees) for modeling.
4. Fit data using 10-fold cross-validation:
    - Confirm model fit scores on train vs. test.
5. Fit all data and extract feature importance:
    - Coefficients for logistic regression and feature importance scores for extra-trees classifier.
6. Compare feature importance across two modeling approaches:
    - Rank order feature importance.
7. Bonus: Univariate feature selection!



- Top 5 most important contributing factors (predictors) for predicting heart disease (target) identified by the **logistic regression** model:
  - # of major colored vessels (*ca*)
  - Defect type (*thal*)
  - Chest pain type (*cp*)
  - Sex (*sex*)
  - Exercise induced angina (*exang*)



- Top 5 most important contributing factors (predictors) for predicting heart disease (target) identified by the **extra-trees classifier** model:
  - # of major colored vessels (*ca*)
  - Chest pain type (*cp*)
  - Defect type (*thal*)
  - Exercise induced angina (*exang*)
  - Maximum heart rate (*thalach*)



- Top 5 most important contributing factors (predictors) for predicting heart disease (target) identified by the **SelectKBest (k=*5*, score function=*mutual_info_classif*)** model:
  - Defect type (*thal*)
  - Chest pain type (*cp*)
  - # of major colored vessels (*ca*)
  - Exercise induced angina (*exang*)
  - ST depression (*oldpeak*)

Results

- The top 5 most important contributing factors (predictors) for predicting the presence/absence of heart disease (target) identified by the **logistic regression** model: *ca, thal, cp, sex, exang*
- The top 5 most important contributing factors (predictors) for predicting the presence/absence of heart disease (target) identified by the **extra-trees classifier** model: *ca*, *cp*, *thal*, *exang*, *thalach*
- The top 5 most important contribution factors (predictors) for predicting the presence/absence of heart disease (target) identify by the **univariate feature selection** model: *thal*, *cp*, *ca*, *exang*, *oldpeak*
- In general, the three modeling approaches are relatively consistent in the contributing factors identified as the most important for predicting the presence/absence of heart disease.
    - **Specifically, *cp*, *ca*, *thal*, and *exang* were consistently identified as the most important contributing factors across all three modeling approaches.**