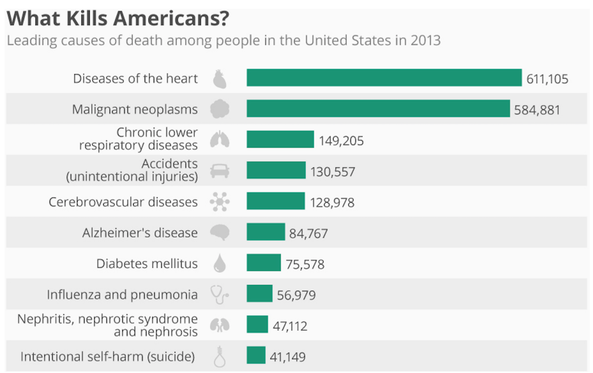
Classification For Heart Disease Diagnosis

Yafeng Wang

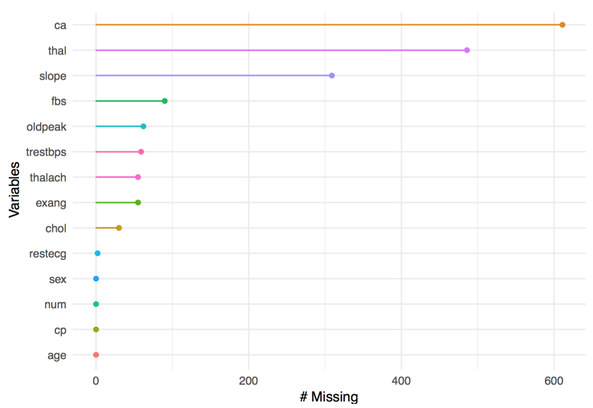
Motivation



- Heart disease is the No.1 cause of death throughout the world each year.
- Automated diagnosis helps reducing diagnosis time, and serves as second opinion for doctors.
- **Goals**: To compare the performances of various diagnostic algorithms, and to select interpretable features that are most predictive of heart disease.

Data Preprocessing

- **Data source**: Merging 4 data sets from UCI ML repository, with a total of 920 patients, 13 features and a categorical outcome 'num' with 5 heart disease categories (0-4).
- **Missing data handling**: 621 patients have missing values (NA). After dropping the 3 features that account for most of the NAs, we obtain 740 patients without NAs:



Task Definition

- **Binary classification**: collapse outcome values to 0 (no heart disease) and 1 (some degree of heart disease).
- **Training-test split**: Randomly split the 740 patients by the 70-30 training-test ratio.
- **Evaluation metric**:
  Accuracy: (TP + TN) / (TP + FP + TN + FN)
  Recall: TP / TP + FN
  Precision: TP / TP + FP
- **Task**:
  Compare classification performance (using 10 features) on the test set.
  Select the most important predictors.

Four Classifiers

1. Logistic regression: **Baseline**.
2. Naive Bayes.
3. Support vector machines: Use 10-fold, 3-repeat cross validation to tune parameters C and σ.
4. Random forests: Use 10-fold, 3-repeat cross validation to tune the parameter mtry (number of candidates at each split).

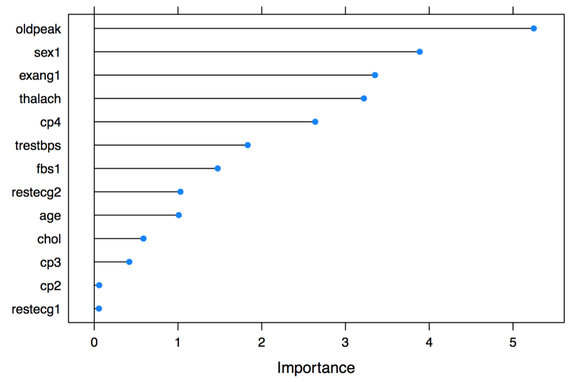
Classification Results

Model \ Metric	Accuracy	Recall	Precision
Logistic Regression	0.7993	0.7569	0.8516
Naive Bayes	0.8132	**0.8000**	0.8210
SVM	**0.8141**	0.7708	**0.8672**
Random Forests	0.7993	0.7643	0.8359

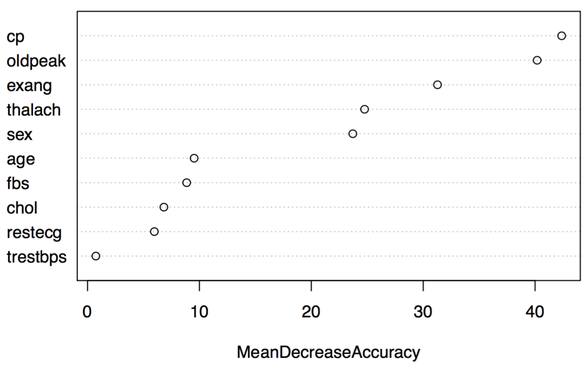
- These are the results on the test set, which contains a total of 269 patients, 128 of whom are healthy (num = 0) and 141 of whom have some level of heart disease (num = 1).
- The four methods have very similar performances, with SVM having the best accuracy and precision and Naive Bayes have the best recall.
- In disease diagnosis, false negatives is arguably more costly than false positives. Hence it is important to prioritize improving recall rate.

Variable Importance

- From the 10 features used to train the models, we would like to select the features that are most predictive of heart disease.
- For logistic regression model, the **absolute value of the t-statistics** for each model parameter is used as importance measure.

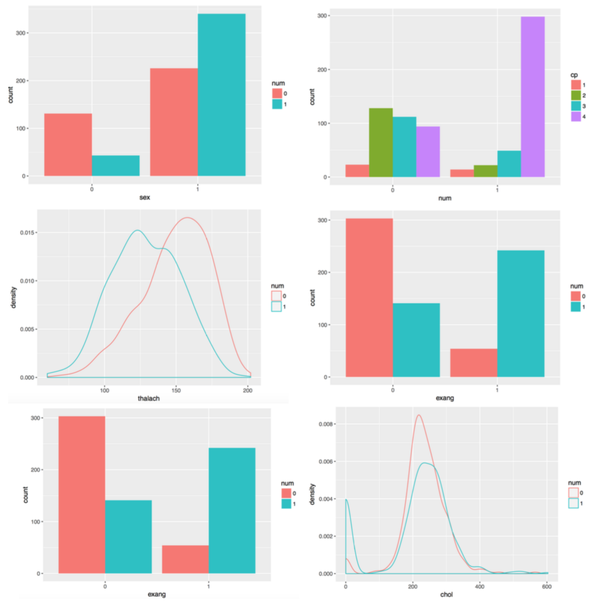


- For random forests, the **mean decrease in accuracy** for each variable (e.g, decrease in accuracy due to permutation the variable) is used as the importance measure.



- The two measures of variable importance converge to the conclusion that five variables are the most important:
  - sex: gender
  - cp: chest pain type
  - thalach: maximum heart rate achieved
  - exang: exercise induced agnia
  - oldpeak: exercise induced ST depression

Visualizing variables



- Data Visualization suggests that some of the important predictors (e.g., sex) may be artifacts of this particular data set.

Results Revisited

- We retrain the models using only the five most important variables, together with the variable 'col' (serum cholestoral level) based on background knowledge.
- The performances of the four models based on these six variables are nearly identical with the performances of the full models.

Model \ Metric	Accuracy	Recall	Precision
Logistic Regression	0.8104	0.7770	0.8438
Naive Bayes	**0.8178**	**0.7883**	0.8438
SVM	0.8104	0.7692	**0.8571**
Random Forests	0.7993	0.7643	0.8359

Future Work

- Experiment with more classifiers to see if classification performance can be improved.
- Experiment with other methods of feature selection to see if the variable importance selection is robust.