Heart Failure Death Prediction

Badger Analysts
Shaonan Wang & Yumian Cui

Content

- Thesis statement
- Motivation for models
- Data processing/modeling
- Model performance evaluation
- Feature Selection/Analysis

Thesis statement

- Found **Random Forest** to be the model with optimal performance applied to this dataset (**Logistic Regression** also works well)

- **Serum_Creatinine, Age, Ejection_Fraction, and Time** most correlated to the Death Event

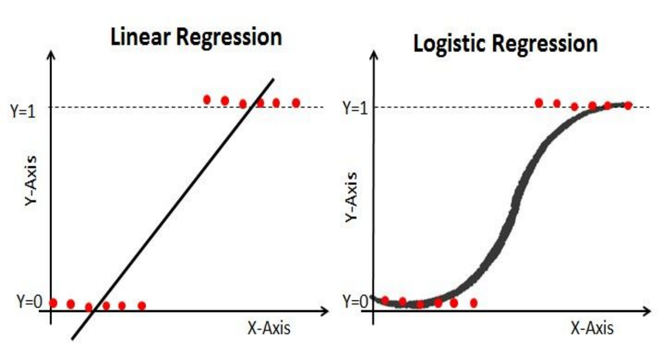
Motivation for Models

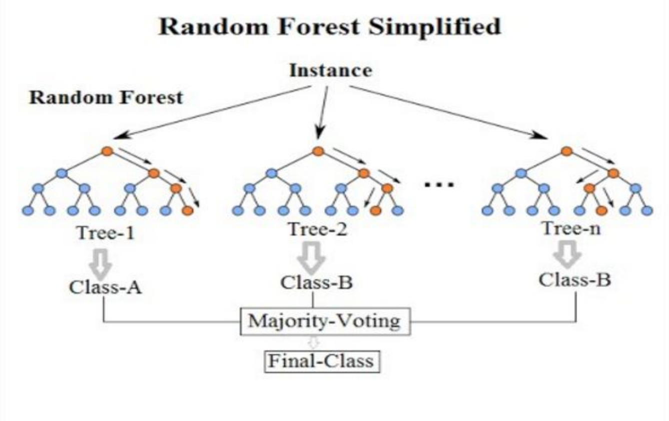
Logistic Regression

- Estimate dependent binary variables
- Provide magnitude & direction of the association
- Limitations: linearity assumption

Random Forest

- Improved upon the decision tree: higher accuracy and lowered variance
- Insensitive to outliers
- Nonlinear nature





Source from medium(top right),Wikipedia(bottom right)

Data processing/modeling

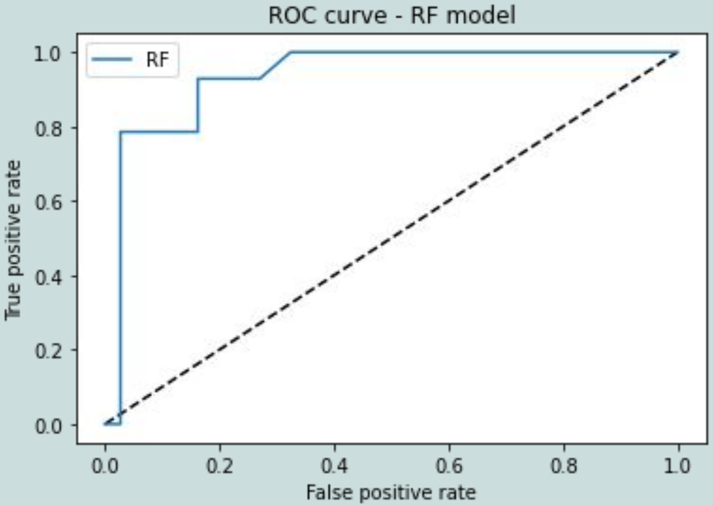
- Data preprocessing
- import the classifiers from sklearn -> get 5-fold cross validation score -> the hyperparameter tuning by GridSearchCV
- In the mid of data modeling , went back to remove outliers -> accuracy increased by 2-3%.

	5 fold cross validation accuracy score	After Hyperparameter Tuning
RF	0.8746 $\xrightarrow{N =55}$	0.8748
LR	0.8357 $\xrightarrow{C:1\ L1}$	0.8619
~~KNN~~	~~0.7505~~	~~N/A~~

Model evaluation

- Confusion Matrix + ROC curve
- Both models do well, but Random Forest is slightly better.
- The limitation may be...not sure model performance after removing outliers

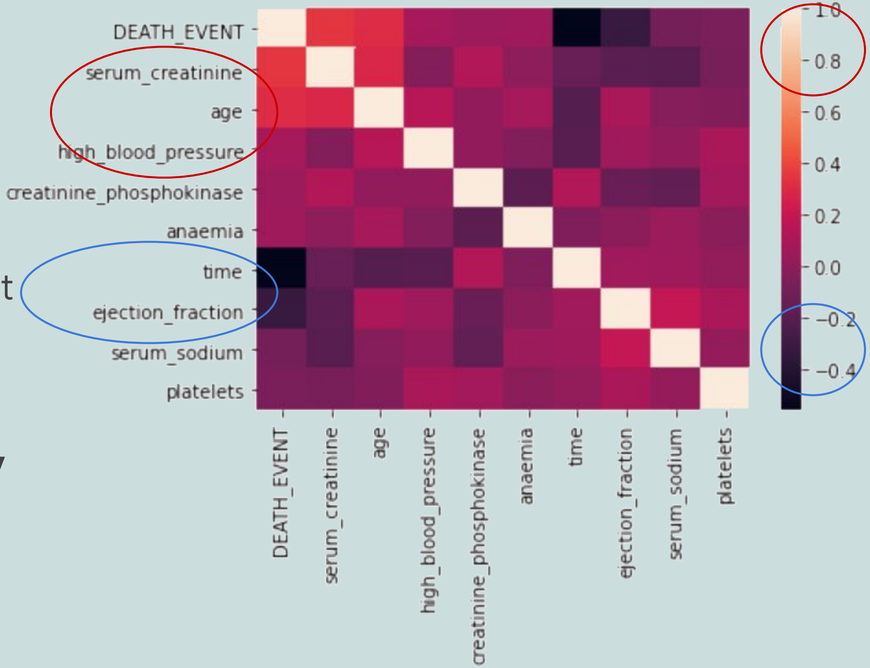
	RF***	LR
Accuracy	**0.8823**	0.8627
Precision	**0.7857**	0,7692
Recall	**0.7857**	0.7142


ROC curve - RF model

Feature Selection/Analysis

General Procedures

- **Identify the correlations** between DEATH_EVENT and other features
- **Choose highly correlated features** as different independent variable (x) combinations
    - ○ Match with Random Forest feature importance
- **Evaluate based on the accuracy scores** of the confusion matrix in testing and training sets



Feature selection/Analysis

- Choose features with high positive correlations
- Divide the testing and training sets into 50/50
    - Should increase the sample size
- Choose features with high negative correlations
- Combine features
- More accurately predict factors preventing the heart failure death, and identify features highly correlated to the death

Test	Training sets accuracy	Diff of accuracy btw training and testing sets
1&2	75%	5%~7%
3	76%	1%
4	82%	8%
5&6	83%	12%~13%

Conclusion

Related Features

- Older people with higher levels of serum creatinine in the blood
  - more likely to die of heart failure
- People with a higher percentage of blood leaving the heart at each contraction and follow-up more frequently
  - less likely to induce heart failure death
- Other features like high blood pressure and smoking
  - likely to cause heart-failure death

Health Suggestion

- Older people keep the levels of serum creatinine and blood pressure low
- Exercise more to strengthen the heart muscles to push more blood out of the heart

Thanks for listening

Questions?