Final Project Presentation

1052. Data Science in Practice

Yang-Ming Lin

OUTLINE

Input -

Modeling -

Your Goal/Output -

Demo -

Input

- Data source
  - Kaggle
    - IBM HR Analytics Employee Attrition & Performance
- Input format
  - CSV
- Any preprocessing?
  - No missing data

https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

Modeling

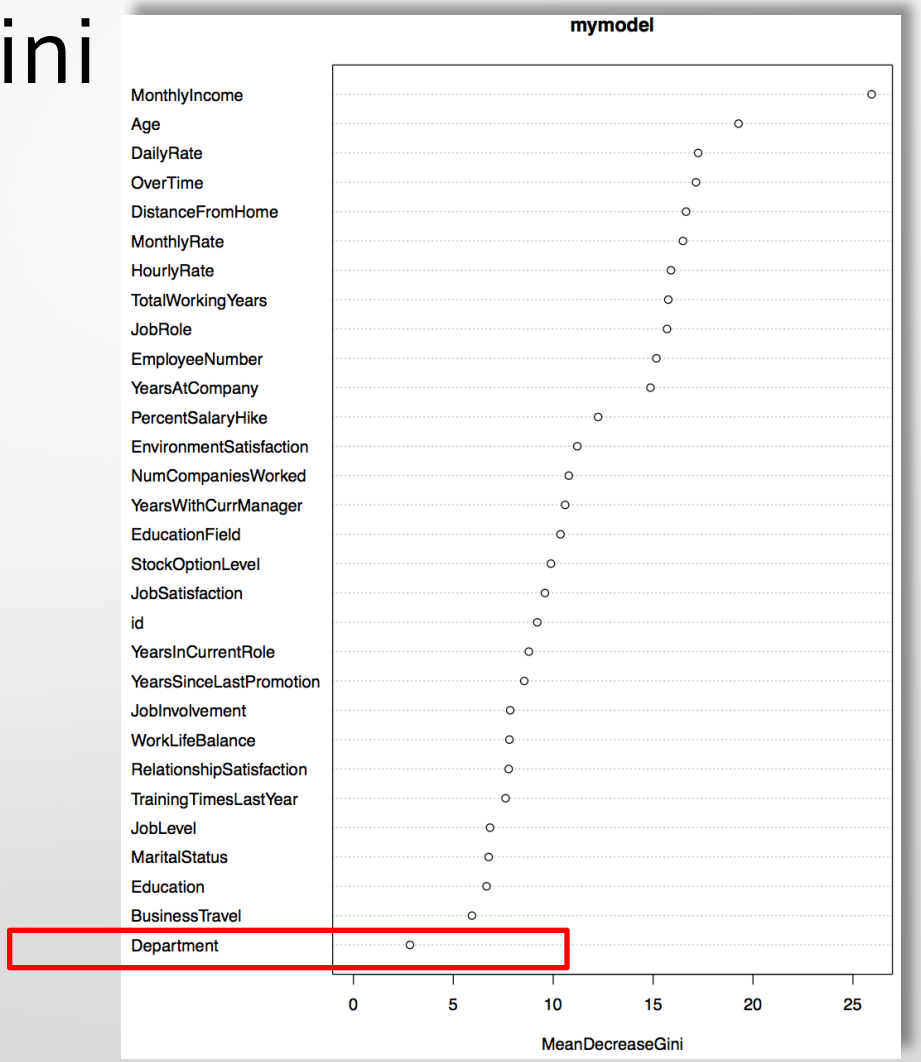
- Which method do you use?
  - random forest r package
- What is a null model for comparison?
  - AUC = 0.5

Modeling

- How do your perform evaluation?
  - Cross Validation
  - mean decrease gini
- Is your improvement significant?
  - Before : All variable
  - After : All except Department

Modeling

- mean decrease gini



Output

- How about performance?

set	accuracy
trainning	0.86
calibration	0.85
test	0.85

set	accuracy
trainning	0.86
calibration	0.86
test	0.86

Output

- How about performance?

method	sensitivity	specificity	F1	AUC
dTrainAll_1	0.8	1	0.89	0.98
dTrainAll_2	0.78	1	0.88	0.98
dTrainAll_3	0.8	1	0.89	0.98
dTrainAll_4	0.81	1	0.9	0.98
dTrainAll_5	0.79	1	0.88	0.98
dTrainAll_6	0.8	1	0.89	0.98
dTrainAll_7	0.81	1	0.9	0.98
dTrainAll_8	0.78	1	0.88	0.98
dTrainAll_9	0.79	1	0.88	0.98
dTrainAll_10	0.87	1	0.93	0.99
highest	dTrainAll_10	dTrainAll_1	dTrainAll_10	dTrainAll_10

Output

- Demo
  - On-line visualization : Shiny
    - https://yangminglin.shinyapps.io/final_project/
  - How do you document your project?
    - Github : https://github.com/Komegaga/1052DataScience/

Output

- Demo
  - How to reproduce your result?
    - Download from my github link
    - Follow readme file
  - What is the challenge part of your project?
    - Decide the dataset
    - Integrate my homework code
    - Visualization

Thanks for Listening!

1052. Data Science in Practice

Yang-Ming Lin