Review of Basic Analytics and Software

BC2407 ANALYTICS II SESSION 2

Pre-Class Learning Activities for S2

- Refer to the PDF document.
- Complete the learning activities before class.

Purpose of Session 2

- To review the 3 key techniques covered in BC2406 Analytics I as BC2407 Analytics II will build on top of those techniques, and thus assumed familiarity:
  1. Linear Regression
  2. Logistic Regression
  3. Classification and Regression Tree (CART)
- To compare the code and outputs from R and Python.

Some R Learning Resources

- BC2406 Analytics I: Visual and Predictive Analytics [Course Materials]
- R Tutorial Blog: https://www.statmethods.net/r-tutorial/index.html
- [eTextbook] Hands-On Programming with R: https://rstudio-education.github.io/hopr/
- [eTextbook] R for Data Science: https://r4ds.had.co.nz/index.html
- List of eTextbooks by RStudio: https://rstudio.com/resources/books/

Distributions of Python

- Anaconda
  - https://www.anaconda.com/distribution/
- Winpython
  - https://winpython.github.io/#overview
  - Portable
    - Can be used without installation. Good if you are not permitted to install software.
    - Can be run from a thumb drive too.
  - Only available in Windows OS.
  - A few variants, depending on your software bundling needs.
- Several Other Distributions
  - https://wiki.python.org/moin/PythonDistributions
- You are free to use your preferred distributions.

Some Resources for Learning Python

- AB0403 Decision Making with Programming and Data Analytics
- BC3407 Programming for Business Transformation
- Online Python Tutorial:
  - https://www.programiz.com/python-programming/tutorial
- Latest Python Beginner Documentation and Tutorial
  - https://docs.python.org/3/tutorial/index.html
- Think Python: How to Think Like a Computer Scientist 2ed. https://open.umn.edu/opentextbooks/textbooks/43
- List of eTextbooks:
  - https://www.java67.com/2017/05/top-7-free-python-programming-books-pdf-online-download.html
  - https://realpython.com/best-python-books/

Python as another Software Tool for Analytics

- BC2406 and BC2407 are about Analytics, not SAS, R or Python.
- Learn just enough R or/and Python to execute selected Analytics techniques in BC2407.
- Not enough and not meant to train you to be a R or Python Programmer.
  - Reflected in BC2407 Course Title.
  - Focus on the concepts and how to apply them to solve a business problem.

Quiz

Ungraded. Check your understanding of this Session Content.
Use your real name (not nickname) in the quiz.

Review of Basic Analytics Techniques and Code with 2 Datasets

R OR PYTHON (YOUR CHOICE)

Dataset: resale-flat-prices-2019.csv

- Data Source: Housing and Development Board (HDB)
- Background:

The dataset was compiled by HDB and contains resale HDB flat prices in Singapore 2019 along with some flat characteristics. All flats start with a lease of 99 years, at the end of which they “expire” and are returned to the Singapore government. Thus, the older the flat, the shorter the remaining life, and has less market value.

Activity 1

Linear Regression on Flat prices

Est. Duration: 20 mins

1. Create a derived variable "remaining_lease_years" defined as the remaining lease (in years) of the flat as in 2019 and save it as a new column.
2. Set the Baseline Reference level for "town" to "Yishun".
3. Build a Linear Regression model using floor_area_sqm, remaining_lease_years, town, and storey_range to estimate resale_price.
4. What are the model coefficients, R^2 and RMSE?

Notes:

- Not necessary to create train-test split. Focus on the model creation and outputs.
- You may use either R or Python. Volunteer to explain your code and analysis.

Dataset: default.csv

- Data Source: Fictional
- Background:

A list of loan applicants profile and their default status.

Default:

- No: Paid back their loan and interest in full.
- Yes: Did not pay back their loan and interest in full.

AvgBal:

Average amount of unpaid credit card balance over the last 3 months.

Activity 2

Logistic Regression on Default

Est. Duration: 20 mins

1. Build a Logistic Regression model to predict Default status.
2. What are the model coefficients and confusion matrix?
3. Which cases has P(Default = Yes) > 90%?

Notes:

- Not necessary to create train-test split. Focus on the model creation and outputs.
- You may use either R or Python. Volunteer to explain your code and analysis.

Activity 3

CART on Default

Est. Duration: 20 mins

1. Build an optimal CART model to predict Default status.
2. How many decision rules are in the optimal CART model and what is the confusion matrix?
3. Which cases has P(Default = Yes) > 90%?

Notes:

- Not necessary to create train-test split. Focus on the model creation and outputs.
- You may use either R or Python. Volunteer to explain your code and analysis.

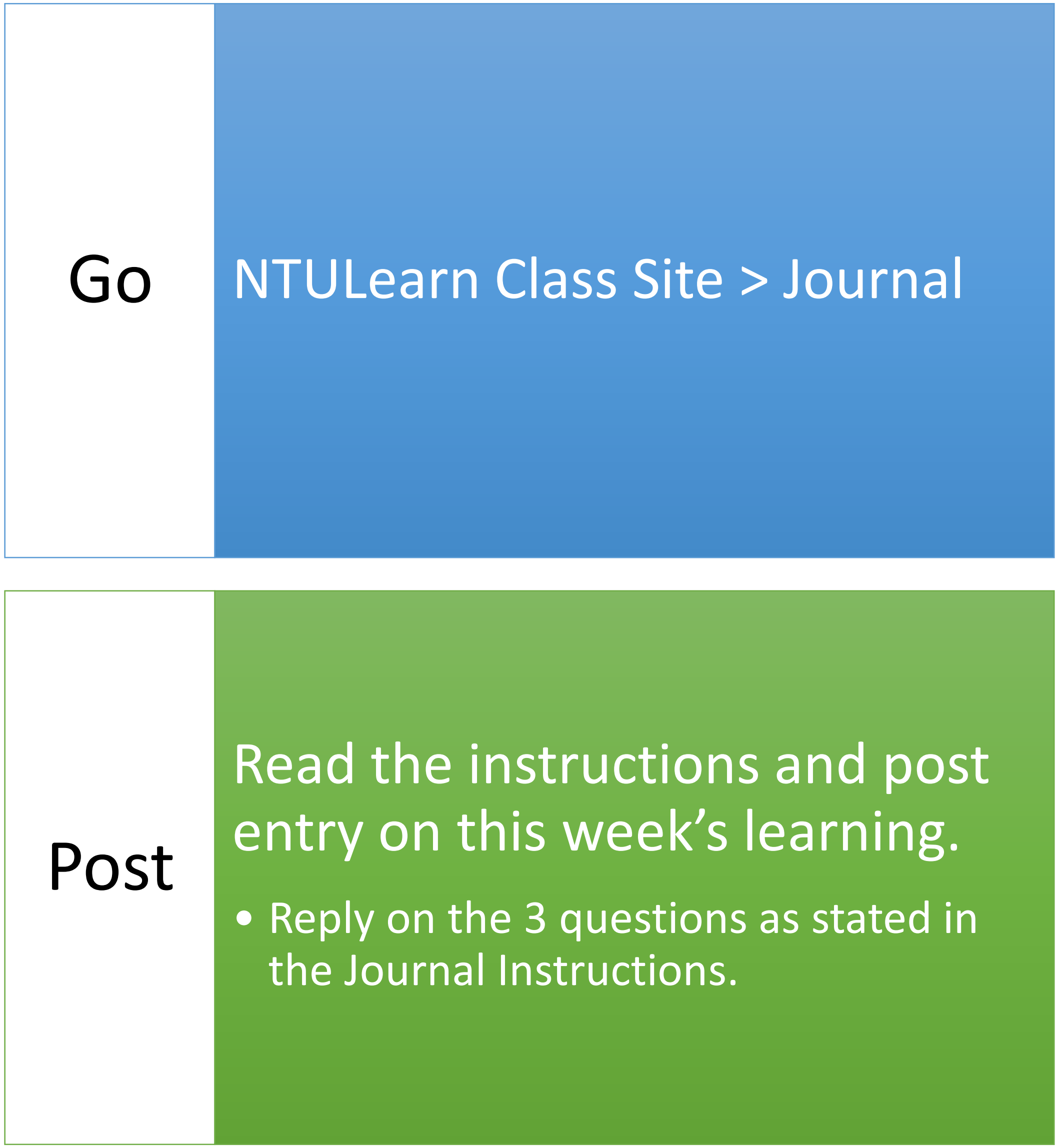
Summary

- Linear Regression review
  - The first requirement.
  - Model Assumptions and Model diagnostics.
- Logistic Regression review
  - The first requirement.
  - Logistic Function.
  - The main weakness.
- CART review
  - Best Splits.
  - Decision Rules.
  - Extends to Random Forest and XGBoost.

Reminder

Please complete the Pre-Class Learning Activities before next class.

Reflection on your Learning



Go

NTULearn Class Site > Journal

Post

Read the instructions and post entry on this week’s learning.

- Reply on the 3 questions as stated in the Journal Instructions.