DATA SCIENCE MODELING

Xin Tang

What is modeling in Data Science

- The modeling is the primary place where the data mining techniques are applied to the data.

- Through supervised learning, modeling builds a simplified representation of reality created to serve a purpose
    - If the target attribute is a category, we call it classification
    - If the target attribute is a number, we call it regression

The types of modeling in Data Science

- The **predictive modeling** is to estimate the unknown value of interest, the target by some known data features.
    - i.e.: Amazon predict what merchandise you like.
    - Judged by its predictive performance

- The **descriptive modeling** is to gain insight into the underlying process or phenomenon.
    - i.e.: What cell phone customers who churn typically look like.
    - Judged by its intelligibility. (how easy to understand)

- The **Optimization modeling** seeks to assess and determine the optimal variable values given an equation.
    - Best dimension to build a fence to improve livestock yield

Build the model, avoid overfitting and under fitting

- Data need to prepared and evaluated.
  - Duplicate? missing data?
  - Data had information interested?
- Data will be split into training and test data set, first set will be used build a model, then the model will be applied to test data to make prediction.
- Many methods can be used to build the model.
  - Decision Tree
  - Linear regression
  - Logic regression
  - Others
- Necessary steps need to taken to avoid overfitting and underfitting
  - Underfitting refers to a model that can neither model the training data
  - Overfitting happens when a model fit the the training data so well that it negatively impacts the performance of the model on test data

What's a good model?

A good model can generalize well to new unseen data based on what it has learned from the training data.

- i.e.: Amazon predicted and recommend merchandised to your based on your browsing history.
- *Do you think Amazon makes recommendations fit your need?*

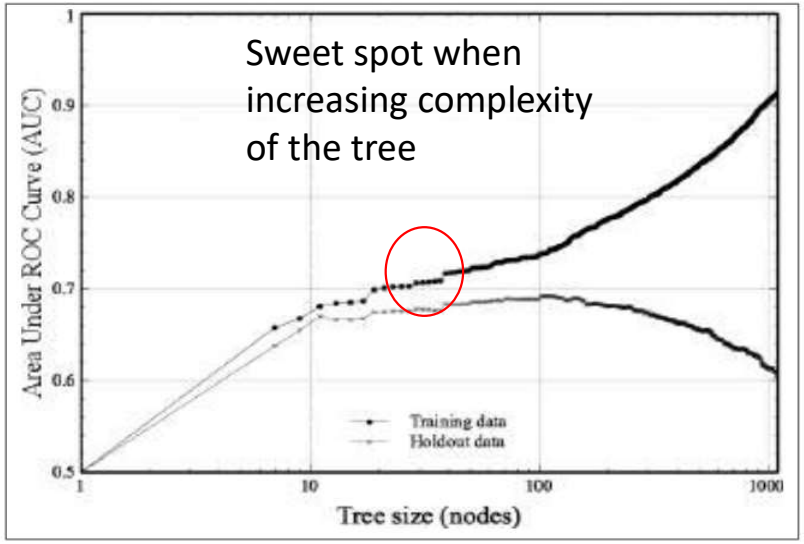
A data scientist will use different metrics to evaluate, visualize various models and compare their performance.

- Accuracy and Precision
- Mean Squared Error
- ROC and AUC: Receiver Operating Characteristic (ROC) curve and Area Under the ROC Curve (AUC)
- Expected value

Visualization

A chart worthy a thousand words.

- Data is too much and too complex to understand directly, same as model.
- A good visualizations helps in comprehension, communication, and decision making



Circled area are tree nodes needed to build model to make best predictions on hold out data.

Acknowledgement and Credits

- Majority statements are quoted from book "*Data-Science-for-Business*" by Foster provost and Tom Fawcett

- Many Thanks to my classmate of DSC500 Winter 2022, they contributed many knowledges through weeks team post/discussion.

- Thanks to Professor Matthew Metzger, his design of course provided many interesting discussion topics, which contribute many contents to this topic.