IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

PARTITIONING PROCESS

LECTURE 6

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

DATA MINING PROCESS

- Partitioning
  - Using same data for model building and model evaluation introduces bias
  - Selection of best model from several candidate models could be due to
    - Genuine superiority of the final model over other candidate models
    - Chance occurrence leading to better match between final model and data
  - Many data-driven techniques can end up producing the latter situation due to overfitting

DATA MINING PROCESS

- Partitioning
  - Partitioning of dataset into two or three parts can solve this problem
  - Typically three partitions- training, validation, and test sets are created following a predetermined proportions for each set and records are randomly assigned to different partitions
  - Sometimes records are assigned based on a relevant variable

DATA MINING PROCESS

- Partitioning
  - Training Partition
    - Usually largest
    - To build the candidate models
  - Validation Partition
    - To evaluate the candidate models
    - Or to fine-tune and improve the model
  - Test Partition
    - To evaluate the final model

DATA MINING PROCESS

- Types of Datasets
  - Cross-Sectional Data
    - Observations on variables related to many subjects (individuals, firms, industries, or countries)
    - Observed at same point of time (snapshot)
    - Unit of analysis is specified
    - Each observation represents a distinct subject
    - Main idea is to compare differences among the subjects

DATA MINING PROCESS

- Types of Datasets
  - Time Series Data
    - Observations on a variable related to one subject
    - Observed over a successive equally spaced points in time
    - Each observation represents a distinct time period
    - Main idea is to examine changes in the subject over time

DATA MINING PROCESS

- Types of Datasets
  - Panel Data or Longitudinal Data
    - Observations on variables related to same subjects over a successive equally spaced points in time
    - Main idea is to compare differences among the subjects and to examine changes in the subjects over time
    - Cross-sections with time order

DATA MINING PROCESS

- Types of Datasets
  - Pooled Cross-Sectional Data
    - Observations on variables related to subjects at different time periods
    - Main idea is to examine the impact on subjects due to environmental changes caused by certain events or policies
    - Independent cross-sections from different time periods

DATA MINING PROCESS

- Model Building
  - An example with Linear Regression
  - Open RStudio

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…