IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

PERFORMANCE METRICS Part-5

LECTURE 20

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

PERFORMANCE METRICS

- Oversampling of rare class members
  - Simple random sampling vs. stratified sampling
- Oversampling approach
  1. Sample more rare class observations (equivalent of oversampling without replacement)
     - Lack of adequate no. of rare class observations
     - Ratio of costs is difficult to determine
  2. Replicate existing rare class observations (equivalent of oversampling with replacement)

PERFORMANCE METRICS

- Typical solution adopted by analysts
  - Sample equal no. of members from both the classes
- Oversampling adjustment for performance evaluation
  - Score
    1. Validation partition without oversampling
    2. Oversampled validation partition and then remove the oversampling effects by adjusting weights

PERFORMANCE METRICS

- Typical steps in rare class scenario
  1. Build the candidate models on training partition with 50% class 1 observations and 50% class 0 observations
  2. Validate the models with the validation partition drawn using simple random sample taken from original dataset
- Detailed steps
  1. Separate the class 1 and class 0 observations into two strata (distinct sets)
  2. Half the records from class 1 stratum are randomly selected into training partition

PERFORMANCE METRICS

- Detailed steps
  3. Remaining class 1 records are reserved for validation partition
  4. Randomly select class 0 records for training partition equal to no. of class 1 records in step 2
  5. Randomly select class 0 records to maintain the original ratio of class 0 to class 1 records for validation partition
  6. For test partition, a random sample can be taken from validation partition

PERFORMANCE METRICS

- When ‘Validation partition without oversampling’ is not useful
  - Due to very few class 1 records
  - Second approach of

  ‘Using oversampled validation partition for evaluation as well and adjusting the weights to get rid of oversampling effects’

  is taken
    - Adjustment of validation partition classification matrix and lift curve is performed to get reliable accuracy measures

PERFORMANCE METRICS

- Lift Curve on oversampled validation partition
  - Multiply the net value of a record with proportion of class 1 records in original data
- In a two-class scenario, records which are difficult to classify by the model, can be labeled with a third class option
  - ‘cannot say’
  - Expert judgment can be used for such cases

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…