IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

PERFORMANCE METRICS Part-4

LECTURE 19

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

PERFORMANCE METRICS

- Open RStudio
- Asymmetric Misclassification Costs
  - When misclassification error for a class of interest is more costly than for the other class
  - Example, misclassifying a customer as false positive who is actually likely to respond to the promotional offering
    - Opportunity cost of foregone sale vs. costs of making an offer (profit of ₹20 for a ₹ 100 item vs. ₹1 scenario)
  - Misclassification rate is not appropriate metric in this case

PERFORMANCE METRICS

- Asymmetric Misclassification Costs
  - Other considerations
    - Costs of analyzing data
    - Actual net value impact per record
    - New Goal : minimization of costs or maximization of profits
- Open Excel
- How to improve actual classifications by incorporating asymmetric misclassification costs?
    - Change the rules of classification e.g. cutoff value

PERFORMANCE METRICS

- Performance Metrics based on asymmetric misclassification costs

$$\text{average misclassification cost} = \frac{c_0 n_{0,1} + c_1 n_{1,0}}{n}$$

  - Measures average cost of misclassification per observation
  - Where c_i is cost of misclassifying a class i observation

PERFORMANCE METRICS

- Ratio of costs (c_0/c_1)
- Future misclassification costs
  - Prior Probabilities (p_0/p_1)
  - (p_0/p_1)* (c_0/c_1)
- Lift curve incorporating costs
- Open RStudio
- Lift vs.
  - No. of records or cutoff value?

PERFORMANCE METRICS

- Asymmetric misclassification costs for m classes (m>2)
  - Classification matrix will be 'm×m'
  - m prior probabilities
  - m(m-1) misclassification costs
  - Matrix for misclassification costs becomes complicated
  - Lift chart not usable for multiclass scenario

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

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...