SAS Enterprise Miner - Data Mining Model Building Documentation

Introduction

This document outlines the process of building data mining models using SAS Enterprise Miner as part of the hotel booking dataset analysis project. SAS Enterprise Miner played a crucial role in the "Model" stage of the SAS SEMMA (Sample, Explore, Modify, Model, Assess) methodology.

Steps in Model Building

1. **Node Configuration:**
    - Configured parameters for each model, including the number of branches for the decision tree, and the assessment measure (ASE) for gradient boosting and ensemble models.
2. **Model Execution:**
    - Executed each data mining model to analyze the hotel booking dataset.
3. **Evaluation and Comparison:**
    - Evaluated model performance using metrics such as misclassification rate, maximum absolute error, SSE, ASE, and RASE.
    - Utilized the Model Comparison Node to compare and select the final model based on Valid Average Squared Error (VASE).
4. **Challenges Addressed - Overfitting:**
    - Detected overfitting issues due to class imbalance.
    - Introduced oversampling using the Sample node to address overfittin

Models Utilized

Decision Tree Model:

- Configuration:
    - Number of branches: 3
    - Assessment Measure: Average Square Error (ASE)
- Results:

        - Achieved a misclassification rate of 16% on both the training and validation datasets.
        - Maximum absolute error: 0.99 for both datasets.
        - SSE: 4,725.49 for training and 2,016.10 for validation.
        - ASE: 0.11 for both datasets.
        - RASE: 0.34 for training and 0.33 for validation.
        - High accuracy for correctly predicting non-cancelled and cancelled bookings.
        - Challenges observed in distinguishing between non-cancelled and cancelled bookings.

    Gradient Boosting Model:

    - Configuration:
        - Assessment Measure: Average Square Error (ASE)
    - Results:
        - Misclassification rate of 18% for both training and validation datasets.
        - Maximum absolute error: 0.94 for training and 0.92 for validation.
        - SSE: 5,416.54 for training and 2,304.31 for validation.
        - ASE: 0.13 for both training and validation.
        - RASE: 0.36 for both training and validation.
        - Struggled to differentiate between non-cancelled and cancelled bookings.

    Ensemble Model:

    - Configuration:
        - Assessment Measure: Average Square Error (ASE)
    - Results:
        - Achieved an ASE of 0.13, higher compared to other models.

Model Comparison and Selection

- Model Comparison Node:
    - Used to compare the performance of competing models.
    - Selection based on the Valid Average Squared Error (VASE).
- Decision:
    - Decision tree chosen as the final model for the prediction problem.

Challenges Addressed

- Overfitting:
    - Observed overfitting issues due to class imbalance.
    - Addressed through oversampling using the Sample node.

Conclusion

The data mining models built using SAS Enterprise Miner provided valuable insights into the hotel booking dataset. The decision tree, despite challenges, was selected as the final model. The documentation highlights key metrics and configurations for each model, aiding in the assessment of their performance.