Talend Data Preparation - Data Cleansing and Transformation Documentation

Introduction

This document elucidates the role of Talend Data Preparation in the hotel booking dataset analysis project. Talend Data Preparation was employed for data cleansing and transformation tasks, ensuring the dataset's quality and compatibility for subsequent modeling stages.

Tasks Accomplished with Talend Data Preparation

Unwanted Variable Removal:

- Identified and removed irrelevant variables, such as 'Booking ID,' using Talend Data Preparation's intuitive interface.

Categorical Variable Encoding:

- Encoded textual variables into numerical formats, a prerequisite for feeding data into machine learning models.
- Variables processed: `room_type_reserved`, `type_of_meal_plan`, `market_segment_type`, `booking_status`.

Feature Engineering:

- Introduced two new features, `length_of_stay` and `total_guests`, to enhance the dataset's predictive power.
    - `length_of_stay`: Represented the duration of the stay by combining weekday and weekend nights.
    - `total_guests`: Indicated the total number of guests by summing up adults and children.

Variable Preparation:

- Loaded the preprocessed data into SAS Enterprise Miner, assigning roles and levels to variables for modeling.

Impute Missing Values:

- Utilized the Impute node in Talend Data Preparation to handle missing values using the mean imputation method.

Data Transformation:

- Addressed skewness in specific variables (`no_of_children`, `no_of_previous_bookings_not_canc`, `no_of_previous_cancellations`) using log transformation.

Variable Selection:

- Applied variable selection using the Chi-Square method to identify relevant features for predicting the target variable (`booking_status`).

Talend Data Preparation Workflow

Task 1: Remove Unwanted Variable

- Identified and removed the 'Booking ID' variable as it was deemed irrelevant for the prediction problem.

Task 2: Encode Categorical Variables

- Converted textual variables into numerical format for compatibility with machine learning models.

Task 3: Feature Engineering

- Introduced new features (`length_of_stay`, `total_guests`) to enrich the dataset with meaningful information.

Task 4: Variable Preparation

- Loaded the preprocessed data into SAS Enterprise Miner, assigning roles and levels for subsequent modeling.

Task 5: Impute Missing Values

- Handled missing values using the Impute node and mean imputation method.

Task 6: Data Transformation

- Addressed skewness in specific variables through log transformation.

Task 7: Variable Selection

- Applied Chi-Square method for variable selection based on relevance to the target variable (`booking_status`).

Conclusion

Talend Data Preparation emerged as a vital tool in the SEMMA methodology, contributing significantly to the exploration and modification stages. This documentation outlines the specific tasks accomplished using Talend Data Preparation, emphasizing its role in ensuring data quality and preparing the dataset for advanced modeling.