PURDUE UNIVERSITY

MGMT 57100 Data Mining
Team Project

PTSD Team

Pooja Udayanjali Kannuri

Shubhankar Bhajekar

Dounia Theoni

Model 1

The provided diagram delineates a structured data analysis and model development workflow, as detailed below:

Train Node: This initial stage represents the entry point of the dataset, earmarked for model training.

Filter Node: Data preprocessing is conducted at this juncture, utilizing the standard deviation from the mean as the default method for filtering interval variables, facilitating outlier detection or data normalization.

Data Partition Nodes: The dataset undergoes a bifurcation into two subsets: 70% for training and 30% for validation, with a random seed parameter set to 54321 to ensure reproducibility. The partitioning follows a default method, unspecified in the diagram.

Regression Node: A logistic regression model is applied, employing a logit link function. Logistic regression is typically utilized for binary classification tasks and is characterized by its probabilistic approach.

Gradient Boosting Node: The workflow incorporates a gradient boosting algorithm characterized by 50 iterations, a learning rate of 0.05, and a seed of 12345. The model is trained on 60% of the data, with a maximum tree depth of 10. The assessment of the model's performance is based on decision metrics.

Ensemble Node: An ensemble technique is adopted, presumably integrating the outputs of the logistic regression and gradient boosting models to enhance predictive accuracy.

Model Comparison Node: This stage involves the evaluation of the models' performances based on the Receiver Operating Characteristic (ROC) curve—a tool for visualizing the trade-off between sensitivity and specificity in binary classifiers. The models are compared using a validation dataset.

Test Node: The preferred model undergoes further evaluation against a test dataset.

Score Node: Post-testing, the model's predictive power is quantified, typically through accuracy or another relevant performance metric.

Save Data Node: The final action in the pipeline entails the preservation of the model output or the model itself for subsequent utilization or examination.

Ensemble with ROC Validation: The ensemble model emerges as the selected model, demonstrating an ROC validation score of 0.959, indicative of its robustness.

Kaggle Score: This field is marked with an 0.94319, indicating that the Kaggle competition score, corresponding to the model's performance, is either pending or has not been included in the diagram.

The flowchart encapsulates a comprehensive approach to machine learning model development, from preprocessing and partitioning to training, validation, and testing, culminating in an ensemble model with a high ROC score, indicative of its superior performance on the given task.

Model 2

The updated data analysis workflow introduces a "Transform Variables Node" before the logistic regression model. This node applies a log transformation to interval variables, a method used to normalize data distributions and reduce the influence of extreme values. This step is essential for preparing the data for the logistic regression model, which utilizes a logit link function. The rest of the workflow, including data partitioning, application of a gradient boosting model, ensemble creation, model comparison, and final output storage, remains unchanged. The ensemble model, which combines the logistic regression and gradient boosting models, continues to be the selected model, evidenced by its ROC validation score of 0.965. The Kaggle score is noted as 0.93419 indicating it is unspecified.