Predictive Analytics Homework Assignment

Predicting and Explaining Customer Churn

In this homework assignment you are to use a free/open-source data mining tool, called KNIME (http://knime.org), to build predictive models using a relatively small and mostly clean data set related to *customer churn analysis*. You are expected to analyze the given dataset (about the customer churn/attrition behavior for 1000 customers) to develop and compare at least four prediction-type (i.e., classification) data mining models. For example, you can choose to include number-based predictive models like Logistic regression, Discriminant Analysis, *k*-Nearest Neighbor (kNN), Artificial Neural Networks (ANN), Support Vector Machines (SVM) or set-based prediction models like Decision Trees, Naïve Bayes, Random Forest, Boosted Trees type models in your comparative analysis. You are to include at least two of each from the number-based and set-based model types (feel free to do more). Your exercise should also include the following:

- Intimately understand and critically explore the data. You can use tables and charts/graphs to perform this task—remember, Data Explorer node in KNIME is an excellent resource for this task.
- Performs row and column filters, if and when necessary.
- Use the Color Manager for better visualization of rows/tables and tree structures.
- Show your complete final workflow picture for ease of assessment.
- Properly balance the data (you can show the results with and without data balancing).
- Evaluate the output of each model type with both the Scorer and the ROC nodes. Also, at the end you need to have a table where you can compare different models on Accuracy, Sensitivity, Specificity, and ROC value.
- Compare the models' performances by combine the model outputs (using the Joiner nodes) into a single ROC chart.
- Show he first three levels of the decision tree graphical model (the whole tree may be too large to fit into the page with legibility). Based on the decision tree splits, briefly comment on the top variables (i.e., variable importance).
- Produce the Variable Importance graph using RF variable statistics.
- In addition to single-split, also use (on at least one model type, or all model types) the k-fold cross validation, and compare its results to single-split results.
- In the final report, use the steps on CRISP-DM methodology as major headings to structure your professionally organized report.

When you are writing and formatting your homework report, make sure to follow the general reporting guidelines.

Good luck!

D. Delen