**Evaluate a Decision Tree and Classification Model**

David Yedlin Broom

Northcentral University

TIM\_8520: Inferential and Predictive Analytics

Dr. Brian Holbert

December 18, 2022

**Evaluate a Decision Tree and Classification Model**

An important and effective tool for data mining is the decision tree model.  The decision tree model provides classification of data by a systematic method.  The decision tree model can be tested for accuracy and is one of the most useful tools for data mining.  The decisions tree model will be used to classify groups by their income on a data set with many categories.

Song and Lu (2015) describes the decision tree model as a method that classifies a population into branch-like segments of an inverted tree.  The parts of the tree are the root node which is at the top of the model and internal nodes with terminal nodes being the leafs.  This decision tree model was introduced in the 1960’s.  One of the most effective data mining techniques in the decision tree model.

The part of the decision tree model is the selection of variables to be used.  Song and Lu (2015) states the variables chosen should be relevant to a study.  The decision tree model is a great way to determine the relevance of the model.

The data may contain missing variables.  A decision tree can handle missing variables so it will reduce bias and be more efficient in data mining.  There are two ways that the decision tree method deals with missing data.  One way a decision tree model deals with missing data is to classify it into a separate category.  The other way for dealing with missing data in a decision tree model is to make the variable with lots of missing data the target variable (Song & Lu, 2015).

Decision tree models can be used for prediction.  Predicting future data off results from a decision tree model is very easy.  One of the most important aspects of the decision tree model is that it can be used to predict future data.

The ROC curve is an important tool for measuring accuracy.  ROC stands for receiver operating characteristic.  It can be used for different types of models and is a simple graphical tool.  Two main factors are involved in an ROC curve which are sensitivity and specificity.  Sensitivity is the true positive rate while specificity is the true negative rate.

Zou et al. (2007)  states that an ROC curve is a plot of sensitivity as a dependent variable and specificity as an independent variable.  There is a diagonal line with a slope of one which represents a random chance.  The best ROC curve is a line from the origin moving up by one unit and the left by one unit.  The area under the ROC curve is a measure of accuracy across all test levels.

There are several ways of developing an ROC curve, two of which Zou et al. (2007) labels as nonparametric and parametric methods.  For the nonparametric approach there is an advantage that no structural assumptions are made about the form of the plot.  Parametric methods are an alternative approach that uses a binormal model.

The summary measures for the ROC curves include confidence intervals, area under the curve, comparison of different curves, partial area, and optimal threshold.  Zou et al. (2007) states that the comparison of curves derived from two different diagnostic tests may use the Pearson coefficient to estimate a correlation between the two different curves.  When evaluating an optimal threshold for an ROC curve the sum of the sensitivity and specificity will be maximized.

The data set that will be studied and analyzed contains social, education and occupational facts for an individual.  The Knime platform will be used to sort the data and perform calculations.  A decision tree will be created from the data set and classification results will be found.  The accuracy of the calculations will be measured in several different ways to maintain the quality of the study.

The data set contains sixteen columns and 32549 rows with various educational, professional, social, and physical information related to a certain individual. Age, race, sex, are the physical characteristics of of an individual in the dataset.  Employment information is partitioned as work class, final weight, occupation, capital gain and loss, hours per week, and income.  The social and education factors are native country, education level, along with marital and family status.  Also each row has an individual id attached to it and all id numbers end in seventy seven.

The age variable has data ranging from seventeen to ninety years old.  The race variable has possible answers of, Black, White, Asian-Pacific-Islander,American-Indian-Eskimo, and Other.   The sex variable contains only male and female as answers.

The first variable related to employment is work class and has the possible values of State-gov, Self-emp-not-inc, Private, Federal-gov, Local-gov, Self-emp-inc, Without-Pay, and Never-worked.  The final weight variable contains a range of discrete numerical values ranging from 12285 to 1484705.  There fourteen different possible values for occupation are specific in description to that individual’s job.  Capital gain and loss are two separate fields with numerical discrete data with a range of zero and zero to 99999 and 4356.  The number of hours worked ranges from one to forty.  Income data is listed as below or above fifty thousand.

The variable Marital status for data related to an individual’s social life has seven possible values specific to describe an individual’s marriage.  The field for relationship has six values such as, husband, wife, unmarried, only child, other relative and not in family.  Two variables describe an individual’s education.  One field for education has a discrete number ranging from one to sixteen.  The other field describes the highest education completed by a certain individual.

Much can be learned from this data set since it contains a good amount of important information about an individual.  A decision tree model  can be helpful in classifying certain patterns that occur in the data set.  The Knime platform will provide an organized view of the data processing and calculations needed to build a decision tree model.

In the Knime workflow process data will flow from the original file to the decision tree model.  From the design tree model the data will enter into a node for viewing and and a node to test for accuracy.  The process is organized and an easy to visualize map that shows the flow of data in connecting nodes.

The first part in the Knime workflow process is the able reader node. A file name ADULT\_JOINED.TABLE was configured into the table reader node.  From the table reader node that data flowed to the partitioning node.

The partitioning node separates the data into a seventy five twenty five split.  The first training set will use 75% of the data while a second test set will retain the leftover 25%.  The data will then flow to the Decision Tree Learner node and the Decision Tree Predictor node.

The decision tree learner node processed the data into a decision tree model.  The node was configured to have the class column set to the income variable.  The variable is of binary value representing less and equal or more than fifty thousand dollars of income.  The data then flows to the decision tree viewer node and the decision tree predictor node.

The decision tree predictor node receives data for two separate nodes which are the partitioning node and the decision tree learner node.  The 25% of partition data is what is received from the partitioning node.  Data also flows in from the decision tree learner node.  The decision tree predictor node uses the unprocessed data to compare and contrast results.

             From the decision tree predictor node data flows to the scorer node.  The scorer node is used to show and check the accuracy of the model.  A Confusion Matrix is produced by the Scorer node and the results for classified and non classified data is reported for the node’s output.

             The data also flows for the decision tree learner node to a decision tree viewer node. The decision tree viewer node produces an in depth view of the tree model with attributes of each leaf and branch of the tree.  The class category of income is the class node while each category of relationship branched from the root node.

The data from the decision tree predictor also flows toward an ROC curve node which is another way to check for the accuracy of the model.  A view of different ROC curves can be seen for each category in respect to true and false positive rates from the node.  The ROC curve node and decision tree nodes are both terminal nodes for the flow of data in the model.

             Data was processed in the Knime platform.  A decision tree model was created using various types of nodes in the Knime platform.  Results were found and their accuracy was measured.

             The results from the decision tree model give insight about classification of income to individuals gathered in certain groups.  The root node is relationship categories by the binary value of income.  From the relation category of wife the incomes of different occupations are categorized by the binary value of income.

The first results from the model can be seen in the decision tree viewer node.  The root node lists a split of 75.9%  and 24.1% respectively to less than or equal to 50k and greater than 50k respectively.  The class node branched to the six different categories of relationships.  The relation is equal to the wife node showed a 48.1% chance of having income greater than 50k.

The occupations with labels Adm-Clerical, Exec-managerial, Handlers-Clears, Prof-specialty, and Other-service had incomes greater than 50k being 46.7%, 62.2%, 28.5%, 72.1% and 18.5% respectively.  For occupations labeled as Sales, Craft-repair, Transport-moving, Farming-fishing, Machine-op-inspct, Tech-support, and Protective-serv, and Priv-house-serv had percentages above 50K respective to 34.8%, 48.4%, 37.7%, 2.0% 21.8%, 53.6%, 63.9%, and 2.0%.  These nodes were branched from the relationship with the category of wife node.

The scorer node produced a Confusion Matrix.  The calculations from this matrix are that 82.856% of the data was accurately classified.  The percent error of the matrix is the complaint of the accuracy at 17.144%.

The ROC curves were calculated for age, hours-per-week, and education-num.  The curve for age rises the highest after about .4 for false positive rate.  The curve for education-num rises the fastest at the beginning of the cart.  After about 0.3 for a false positive rate, hours per week is below the other two curves.

From the results it can be seen that from the occupations of a wife the Prof-specialty has the greatest chance to be over 50k for income.  The least chance for the income of a wife to be over 50k is Farming-fishing.  The confusion matrix shows that the model has good accuracy.

The decision model was optimized for accuracy by changing a few parameters in the decision tree learner node. Brach pruning was selected. The accuracy measure was changed to the gain ratio from the Gini index. The option for binary optimal splits was selected. The nodes were executed and the results for the ROC curve and Confusion Matrix showed a significant improvement in accuracy.

The data set for individuals and their incomes was studied and classified with the decision tree model.  The decision tree model was built on the Knime with a system of various nodes.  The model was tested for accuracy with two methods.  Useful information was found for the classification of individuals based on a binary income value.

**References**

Song, Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044

Zou, K. H., O’Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, *115*(5), 654–657. https://doi.org/10.1161/CIRCULATIONAHA.105.594929

**Appendix A**

Screenshots of Work from the Knime Platform

Diagram

Description automatically generatedDiagram

Description automatically generatedThe figures are screenshots of the graphs and charts from the Knime Platform. A decision tree was modeled from a data set using various different nodes in the Knime workflow.

Figure 2 The Output from The Decision Tree Viewer Node

Figure 1 The Knime Work Flow

Chart, line chart

Description automatically generatedA picture containing chart

Description automatically generatedGraphical user interface, application

Description automatically generated

Figure 4 The First ROC Curves

Figure 5 The Confusion Matrix After Optimization

Figure 3 The First Confusion Matrix

Graphical user interface, application

Description automatically generated

Figure 6 The node configuration for a more accurate model.