Week 3 Assignment: Predictive Modeling Using Decision Trees

Jonathan Ibifubara Pollyn

College of Science, Engineering and Technology, Grand Canyon University

DSC-530: Predictive Modeling

Filippo Posta

September 22, 2021

Predicts Approval using Debt to Income Ratio, FICO Score, and Request Amount

The decision made on the training data shows that the first split checks whether the FICO score less than or equal to 659.5 is the highest because gini index equal to 0.5 is the highest, which also tells us if we split according to this condition, the sample will be evenly distributed to each class Medium. The sample data of about 150302 is distributed evenly where about 104300 samples are recorded as False, and 46002 are records as True. Next, in the second split, we check to see if the FICO Score greater than 659.5, then check whether the debt-to-income ratio less than or equal to 0.305 and about 104300 samples meet those conditions. These types of splits are done in a way that minimizes the Gini index. The closer the Gini index is to 0, it means that all samples belong to one category. The decision tree diagram for question one should consider the distribution of the data and the decision taken. On the other hand, the test data shows that we have 49698 samples, of which 15481 is True when FICO less than or equal 660.5. Suppose the FICO score was greater than 660.5; it then checks for the debt-to-income ratio of less than or equal to 0.305. The decision trees for the training and test data set have close similarities, but their sample sizes differ.

C5.0 Model Using the Training and Test Dataset

The sample size for the training data is 150302 with an entropy of 1.0, and when the C5.0 model is created, we have 43749 records recorded as true when the FICO score is less than or equal to 656.5. When the FICO score is greater than 656.5, the decision check for the debt-to-income ratio is less than or equal to 0.315 and recorded 106553 samples. The decision tree-based classifiers have some fundamental questions for the tree construction: will the tree be binary or not? Which attribute will be node? How will the missing data be handled? And if the tree becomes too large, how to prune it? So, the difference is the C50 uses technologist Entropy

to choose options with the best information gain as nodes. CART uses Gini Impurity instead. Gini Impurity could be alive of the homogeneity (or "purity") of the nodes. If all data points at one node belong to an identical category, this node is considered "pure." Therefore, by minimizing the Gini Impurity, the decision tree finds the options and then separates the information based on the best criteria.

Random Forest Model using Training and Test Data

Random forests use several decision-making trees' power. It does not depend on a single decision tree's feature significance, considering the significance of characteristics given to various features by different algorithms. During learning, random forest picks random characteristics that depend very much on a specific set of characteristics. Random forests can therefore generalize the data more effectively. Random forests are significantly more accurate than decision trees in this random selection.

Mathematical and Statistical Basis

The inverse statistical method of variance is discriminant performance analysis. The cluster is the variable quantity in MANOVA; hence the variable quantity is the dependent variable. The predictor variable is the independent variable in discriminant function analysis, whereas the group is the dependent variable. As previously stated, discriminant function analysis is frequently used to forecast current group membership. It responds to the question of whether a combination of variables will be utilized to predict group membership. Various factors are often included during a study to ensure that variables result in team prejudice (Wang et al. 2020). The process of discriminant function analysis is similar to that of ballroom dancing: Run a variable test first, and if it's statistically

significant, look for significant differences in variables between teams within the mean. The principal component analysis (PCA) procedure entails extracting the subsequent spinoffs of the variance matrix, running every derivative on the matrix, and removing the explained changes. The eigenvector is the derived practical form: direction denotes the vector's relative direction to the center of gravity in (hyper)space, defined by the orthogonal intersection of all variables. Every eigenvector's eigenvalue is an index of the overall variation it explains. The strategy ends when the tangent is nearly zero (and so explains all the system variation). A chapter in the book offers a basic introduction to PCA and its alpha and validatory issue analysis development. This chapter should be available in the science library or on the science library's website (interlibrary loan).

CHAID is the world's oldest decision tree algorithm. Gordon W. Kass first proposed it in 1980. Then in 1984, CART was discovered, in 1986, ID3 was proposed, and in 1993, C4.5 was published. ChiSquare Automatic Interaction Detection is an acronym for ChiSquare Automatic Interaction Detection. Chisquare is a metric for how important characteristics are in this context. The statistical significance increases as the value rise. CHAID, like the others, uses decision trees to solve categorization problems. In other words, it must contain the categorical target variable's records. In addition, traditional decision tree algorithms tend to branch based on classification criteria. We can still make trees with continuous and numerical properties, though. We're turning continuous functions into categorical functions, which is the trick. We'll talk about numerical qualities with the most informational value (Lin et al., 2017). CART generates binary partitions, one of two possible outputs, but CHAID can generate numerous branches from the same

master/parent node. CHAID is typically used for descriptive analysis, while CART is frequently used for predictive analysis.

The care business collects an excessive quantity of knowledge that's not correctly mined and isn't used optimally. The invention of those hidden patterns and relationships typically remains untapped. Our analysis focuses on this facet of diagnosis by understanding patterns through collected infectious disease data and developing intelligent medical call support systems to support doctors. during this article, we tend to propose to use decision tree C4.5 formula (Buskirk, 2018)., ID3 algorithm and CART algorithm to classify these diseases and compare the potency and repair rate between them. Therefore, the model derived from CART alongside the extended definition to define (diagnose) infectious disease provides a decent model supported classification accuracy. Active learning, domain expert, information mining, dynamic query, ID3 formula, CART algorithm, C5 algorithm.

References List

Buskirk, T. D. (2018). Surveying the forests and sampling the trees: An overview of classification and regression trees and random forests with applications in survey research. *Survey Practice*, *11*(1), 1-13.

Lin, L., Wang, F., Xie, X., & Zhong, S. (2017). Random forests-based extreme learning machine ensemble for multi-regime time series prediction. *Expert Systems with Applications*, *83*, 164-176.

Larose, C. D., Larose, D. T., & Larose, Chantal D., Author. (2019). *Data science using python and r*. John Wiley & Sons,inc,.

Wang, Z., Cao, C., & Zhu, Y. (2020). Entropy and confidence-based undersampling boosting random forests for imbalanced problems. *IEEE transactions on neural networks and learning systems*, *31*(12), 5178-5191.