Assignment 2: Decision Trees

Noori Selina, Haig Bedros, Julia Ferris, Matthew Roland

For our analysis, we chose to explore a dataset containing factors that may be associated with student performance in Public and Private school settings. For the first leg of our work, we performed EDA to assess important features of our dataset. Specifically, we found that students had previous average scores of 75, but this decreased to 67.24 for the most recent exam. Furthermore, we explored factors that may influence exam scores to assess preliminary correlational relationships. For instance, we found that parental involvement, family income, teacher quality, attendance, hours studied, and tutoring sessions showed some positive relationship with exam scores, but parental involvement, family income, and teacher quality showed the strongest relationships. Using what we know about these factors, we will create a decision tree and random forest models to explore their importance in predicting high or low exam scores.

Decision trees can guide our knowledge regarding the saliency of different combinations of factors that predict categorical outcomes. These models can be very useful in terms of their simplicity and explanatory power. However, it is important to recognize the caveats and limitations associated with these models. For instance, the blog, “The GOOD, The BAD & The UGLY of Using Decision Trees - DeciZone” outlines the disadvantages and problems that may be associated with using decision trees. Importantly, they discuss how decision trees can suffer in situations wherein the goal is to model complicated topics, or where the outcomes are not straightforward. In addition, the blog mentions that people often include obscure features that audiences may be unfamiliar with. Outside of the issues mentioned in the blog, decision trees can also be limited in terms of overfitting, especially in situations where the data are feature dense. For our models, we will address these issues by generating decision trees with different features to assess the predictive accuracy of models with different combinations of salient features. To start, we will first create a random forest model to broadly examine the importance of features within the dataset. Furthermore, we will create two decision tree models based on the random forest output and clearly label and define the features we are using in terms of what they represent, as well as their importance in the models.

Our random forest model incorporated a multitude of categorical and continuous features to predict the likelihood of receiving a high versus low score. Overall, the model derived adequate performance, with an R-squared value of 0.67 and an MSE of 4.6. We also determined that the trees within the random forest model were stable by performing cross-validation. Also, the tree depth was high, which means the model could be prone to overfitting when used for new data. In terms of individual features, we determined that Attendance played the most important role in predicting exam performance, followed by. In addition, we found that previous exam scores, hours of sleep, physical activity, resource access, and parental involvement play important roles in performance as well. Thus, we will implement combinations of these features into our decision tree models.

Our first decision tree model incorporated features related to study habits, such as the number of hours studied, attendance, and students' previous scores. Our outcome was whether students received a high or low exam scores relative to the median. Using these features, our decision tree model demonstrated an 80% accuracy rating and 82% precision. We also generated a tree depth of 21, indicating a high degree of branching based on these features, so the model could be prone to overfitting. These findings indicate that these more academically-related features are able to predict exam score outcomes with a high degree of accuracy.

Our second decision tree model focused on more tertiary factors, such as teacher quality, parental involvement, and family income. Using the same outcome, our decision tree revealed an accuracy 56%, a precision rate of 60% and a depth of 6, indicating moderate performance. Hyperparameterization did not improve model performance and, in fact, reduced the model's depth. These findings indicate that these tertiary factors are adequate predictors of academic performance; however, more academically-related factors serve as better predictors.

Our analysis shows that a single decision tree using classification struggled to capture the data's complexity, with modest accuracy, precision, and recall scores that didn't improve significantly even after tuning. However, the random forest model, which is an ensemble of multiple decision trees, and the decision tree regression models performed far better. The random forest model accounted for issues with the decision trees because it used many decision trees to form a prediction. It was able to find general patterns in many different ways when creating the individual trees, and then the combined results of those trees formed the overall model. That was beneficial because it led to more stability and fewer errors. Accuracy could not be specifically determined for the random forest regression model, but we analyzed the model with other means like error and stability. The results showed that the model had low error on average and the model was stable. This highlights the advantage of using ensemble methods like random forest over a single decision tree, as they effectively capture complex relationships in data and deliver more accurate predictions. Furthermore, feature selection for single decision trees is important when developing them because different input variables lead to dramatically different results.

Thus, we have aimed to avoid the potential pitfalls of using decision trees as outlined by the blog post. We have generated several models to determine the importance of predictors, and all of our predictors are relatively straightforward in terms of conceptual depth and labeling. Furthermore, while our random forest model is rather feature dense, our decision trees are simpler in construction. Perhaps additional analyses can include a combination of academic and tertiary features to determine how the combination of these factors can influence performance in a simple model.

Sources:

1. https://decizone.com/blog/the-good-the-bad-the-ugly-of-using-decision-trees
2. https://www.analytixlabs.co.in/blog/random-forest-regression/