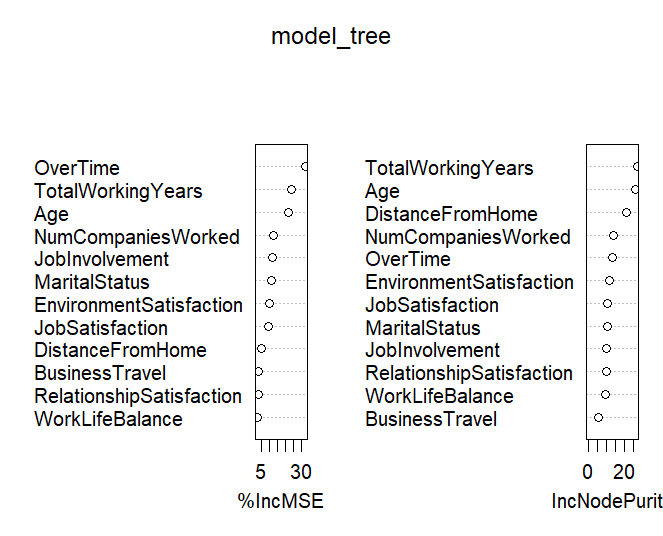
The Exploratory Data Analysis has reduced the number of predictors down to a more manageable twelve and now these predictors can be used with the designated models. Random forest modeling has been chosen for its ability to handle large amounts of data and ability to achieve a high accuracy.

A random forest model consists of a collection of decision trees, where each tree is trained on a given set of data. Each tree makes predictions on the data and the random forest aggregates the predictions from all individual trees to create a final prediction. This aggregation process produces more robust and accurate predictions compared to any single decision tree. While this method can capture complex relationships and potentially mitigate overfitting, it also is a black box process where the underlying process is hard to describe. This black box can be mitigated through validating the results with the other model, logistic regression.

**Random Forest with Selected Predictors**

When running the random forest function with twelve predictors, the model can rate the importance of each variable in two terms: %IncMSE and IncNodePurity. The results of this are produced below.



%IncMSE is the increase in mean squared error and is determined by randomly shuffling the values of a particular variable in the model. If the shuffling produces a less accurate response, then that variable is more likely important for the prediction.

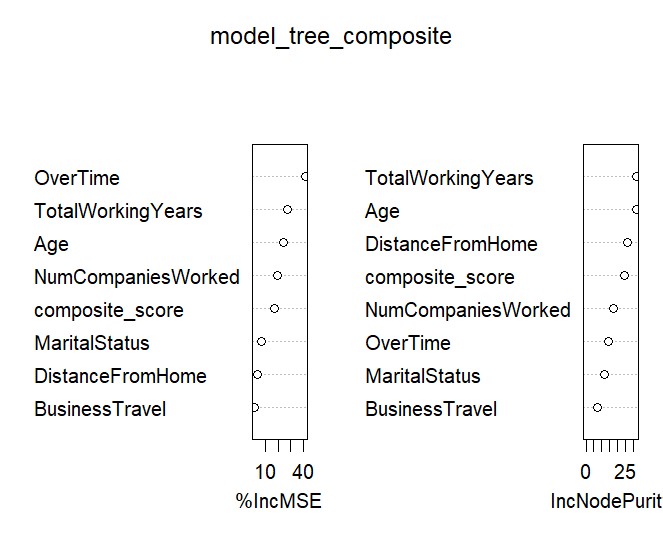
IncNodePurity is the increase in node purity and it measures the quality of the model in splitting the data into groups when making decisions. If splitting the data results in more homogenous groups, then it’s more likely that predictor is more likely important.

This model repeatedly found TotalWorkingYears and Age to be important predictors. It also produced a psuedo R2 value of 0.207 which is relatively low, and there are many predictors that do not have high influence on the model. This indicates that there should be further exploration of random forest models with different predictors.

In conjunction with the logistic modeling, we will explore two more options to reduce and condense the number of predictors: combining the satisfaction ratings of employees and by creating a loyalty score based on an employee’s length of tenure at a company.

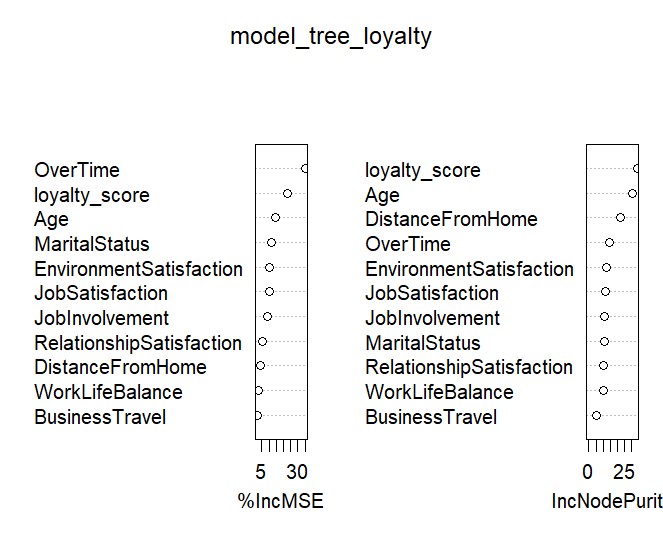
**Random Forest with Composite Rating Score**

A composite score was made by combining WorkLifeBalance, EnvironmentSatisfaction, RelationshipSatisfaction, and JobInvolvement. Rather than keep these predictors separate and further complicating the model, they can be aggregated to create one score that can summarize the employee’s rated quality of life.

Creating a composite score has reduced the number of variables and has generally increased their overall importance as shown in the figure. However, the composite score itself is not very significant which contributes to a lowed pseudo R2 value is 0.184. This method has actually decreased the predictive power of the model.

**Random Forest with a Loyalty Score**

A loyalty score was created by taking the number of companies worked divided by the number of years worked. It is hypothesized that an employee who is more ‘loyal’ is less likely to leave the company.

A model with this loyalty score has marginally increased the pseudo R2 to 0.209. This is a slight increase of predictive power and the loyalty score does show a large degree of importance as shown on the table. Since this model has an increase in R2, it’ll be worth expanding and investigating this model further.

The decision tree is included below as this may help with further improvement. For example, age is particularly important, and the level of grouping is important for employees younger than 27 years old or older than 34 years old.

Going forward, these preliminary models have indicated that we can experiment with more reduction of variables. There’s also the possibility of creating a different model with both the composite score and a loyalty score. And with the preliminary model of the loyalty score, this will give us more insight on how to group age, or environmental satisfaction for logistical modeling.

