

Master of Science in Business Analytics

Machine Learning Final Project—Logistic Regression, Decision Trees, and Random Forest

Team 7: Breanna Barton, Aaron Beverly, Peter Shoemaker, Ryan Taylor, and Pedro Voyer

**Executive Summary**

Canterra struggles with employee retention and is looking to understand better the factors that influence an employee's decision to stay or leave the company. This large company employs around 4000 employees; with a turnover rate of 15%, the company is replacing approximately 600 employees a year. This high attrition rate increases training costs and reduces productivity while hurting timelines and customer trust.

Our team performed model analysis and evaluation to determine the primary causes of attrition and provide management recommendations to decrease the attrition rate. The data provided by Canterra includes both demographic information on employees and information related to their working experience, such as employee satisfaction, years with the company, time with their current manager, and whether travel is required.

We developed logistic regression, decision tree, bagging, and random forest models to identify the most significant variables. The random forest method proved the most accurate and specified **age**, **income**, and **total working years** as the most significant variables.

We recommend Canterra target experienced workers with more years of experience who may exhibit lower attrition rates. Canterra can also explore incentives for younger workers. For example, a flexible work environment or education assistance may motive younger workers to stay with the company and increase the average tenure. Canterra should also compare its pay scale with competitors to ensure their compensation is on par with the industry. These measures would ensure employees are not leaving for financial reasons and offer more motivation to stay at the company.

**Data Description**

To better understand employee attrition, Canterra has provided data on employee demographics and work experience. This data includes age, gender, education, number of years worked, number of years at the company, level of travel, number of years an employee has with their current manager, and job satisfaction. Our team used these variables to understand better the factors that influence employee attrition and make recommendations about what Canterra can do to decrease the employee turnover rate.

Before performing any data analysis, our team used random sampling to split the data into training data to build the model (70%) and testing data to evaluate the model (30%). We isolated the testing data to assure the accuracy of the report. We also used downsampling to account for the unbalanced data, which resulted in a 50-50 split between employees who stayed or left in the data used to train the models[[1]](#endnote-2).

**Methods**

The team first developed a logistic regression model with the data—a combination of demographic and numeric variables describing working characteristics. Dummy variables were added for s**ingle** and **travel frequency**, as these were the subsets of marital status and business travel that had the most significant impact. A stepwise procedure was run to develop the best combination of variables. The final model identified **travel frequency,** **single**, and **job satisfaction** as the top variables[[2]](#endnote-3) for determining attrition. Job satisfaction is an important indicator, and employees with low satisfaction are more likely to find new employment. We quantified the accuracy of the model by applying it to the test data that we previously partitioned. With a probability cut-off of 0.5, we are 95% confident that the true overall accuracy of this model is between 68.6% and 73.6%[[3]](#endnote-4).

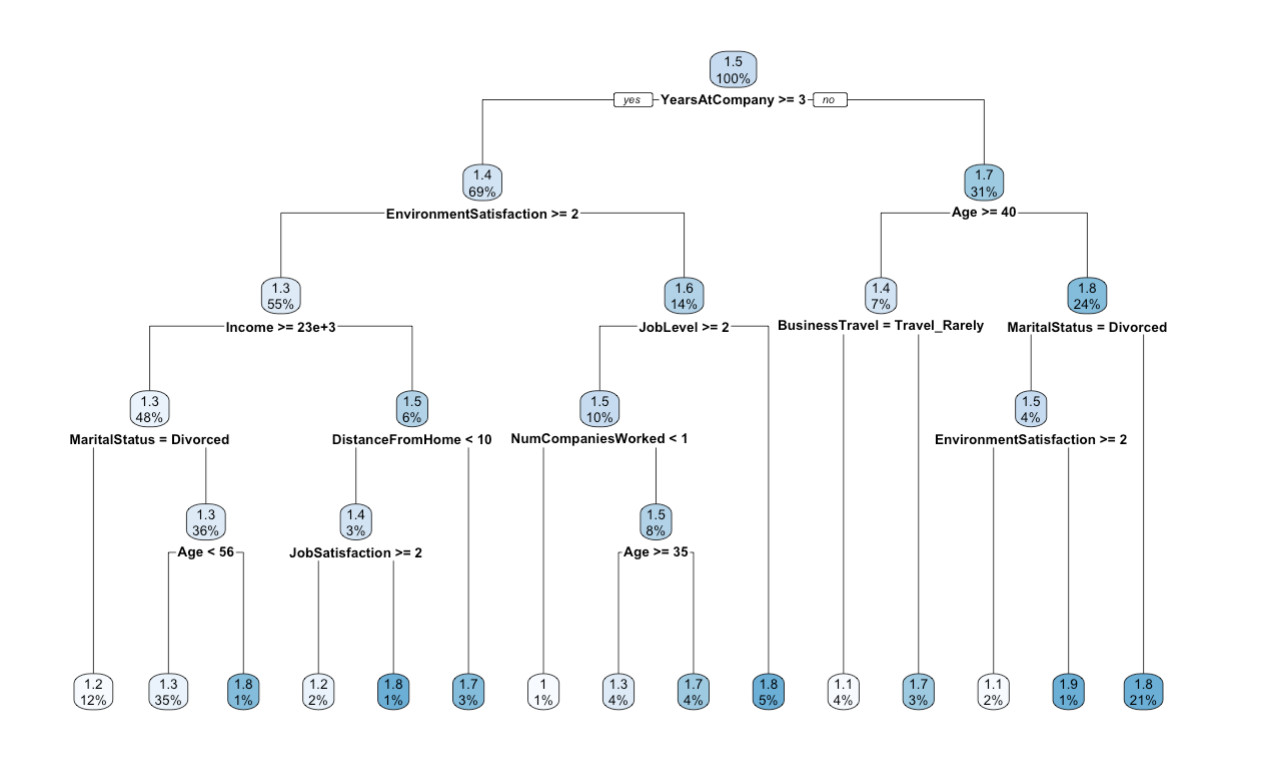
The team exponentiated the coefficients on the logistic regression model to interpret the impact of each variable on employee attrition (Table 1). For every one-unit increase in age with eighteen as the baseline, the odds of employee attrition decrease by about 4%. Job and environment satisfaction play a critical role in retaining employees. For every one-unit increase in job satisfaction or environment satisfaction, the odds of employee attrition decrease by about 18% and 22%, respectively. Increases in training times last year and years with current manager also positively impacted employee attrition, with one-unit increases in both categories showing the odds of employee attrition decreasing by a little over 11%.

  
**Table 1: Exponentiated Coefficients on Logistic Regression Model**

However, employees that frequently travel (FreqTravel) or are single tend to have higher odds of attrition than their counterparts who travel less or are currently married or divorced. The odds of attrition are 103% higher for employees who travel frequently versus employees who rarely travel or do not travel at all. Likewise, single employees' odds of attrition are 126% higher than those married or divorced.

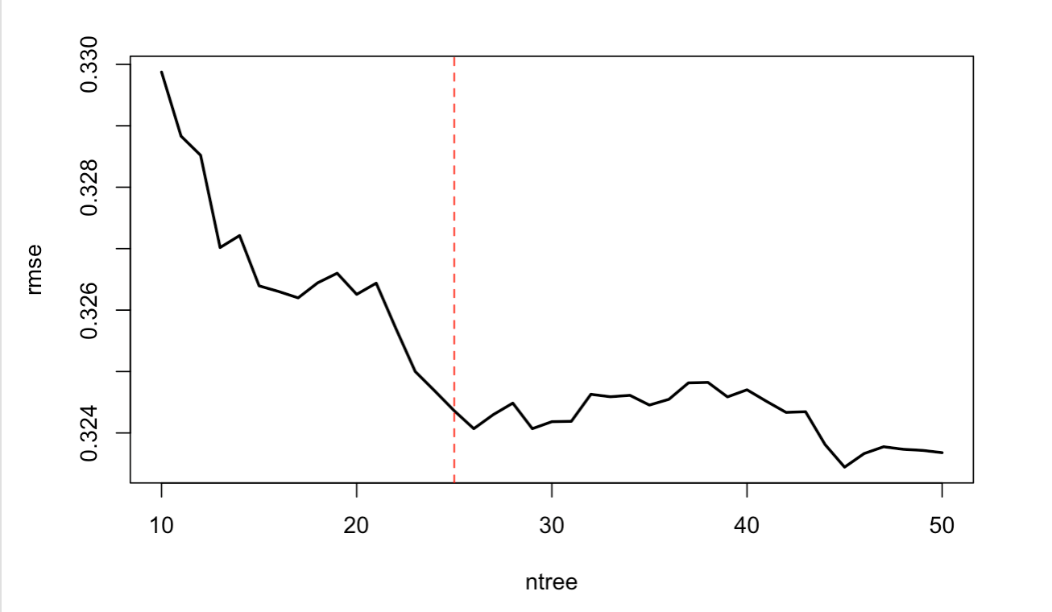
The area under the curve for logistic regression (or AUC) is 0.725. The ideal model maximizes the AUC, so the team was pleased with this result and can confidently use this model to evaluate employee retention. We can infer that the model can correctly differentiate between the employees leaving the company and staying at the company because the AUC is greater than 0.5.

Next, we created a model using decision trees. Decision trees are used for classification by dividing data into smaller and smaller subsets. The initial decision tree model[[4]](#endnote-5) resulted in 12 decision nodes and 13 terminal nodes with **years at company**, **environment satisfaction**, and **age** being the first three variables used to split. The cost complexity parameter[[5]](#endnote-6) plot showed that similar results could be obtained with a small margin of error using 12 terminal nodes. We chose to tune the initial model using a grid search[[6]](#endnote-7) to minimize the error. The smallest error of 0.8052 with a cp of 0.01 was obtained using a minimum split value of 13 and a maximum depth value of 8. After tuning, the final model[[7]](#endnote-8) resulted in 14 decision nodes and 15 terminal nodes. The optimal decision tree is shown below in figure 1.

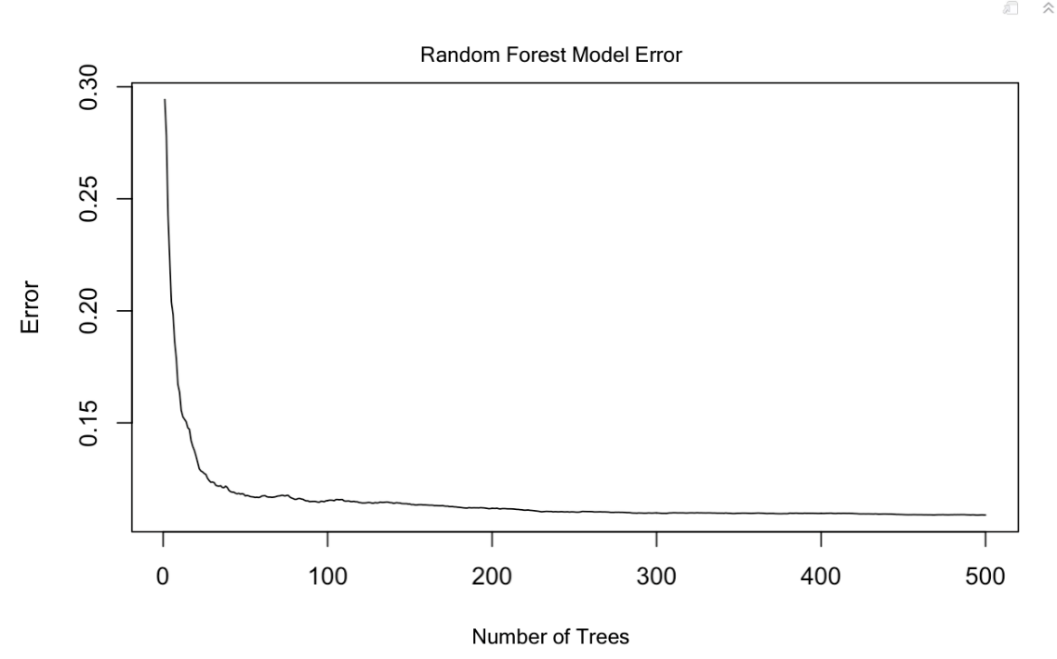


**Figure 1: Optimal Decision Tree**

Next, the team considered the bagging technique, an ensemble method that uses bootstrapping to generate subsamples that can then be analyzed. The decision trees for each subsample are combined to form a more efficient predictor. As the number of subsamples increase, the RMSE decreases. Figure 2 shows the RMSE stabilizing at approximately 25 trees. This model averaging decision tree results, which reduces variance and can minimize overfitting. By predicting with the model generated through bagging and test data, we find that the most important variables[[8]](#endnote-9) are **age**, **income**, and **total working years**.

  
**Figure 2: RMSE for Bagging**

Random forests are another ensemble method that uses bootstrapped subsamples to obtain a model recommendation. This method is comparable to bagging, except that only a subset of features is considered for each split. In Figure 3, we see that the error stabilizes around a sample of approximately 100 trees. We randomly viewed a few of the individual models, and the initial split was on total working years for each of the models. Secondary splits ranged from training times to income. Using the random forest method, we can see that age, income, and total working years are the most significant variables from the aggregated samples.



**Figure 3: Error for Random Forest**

**Results and Insights**

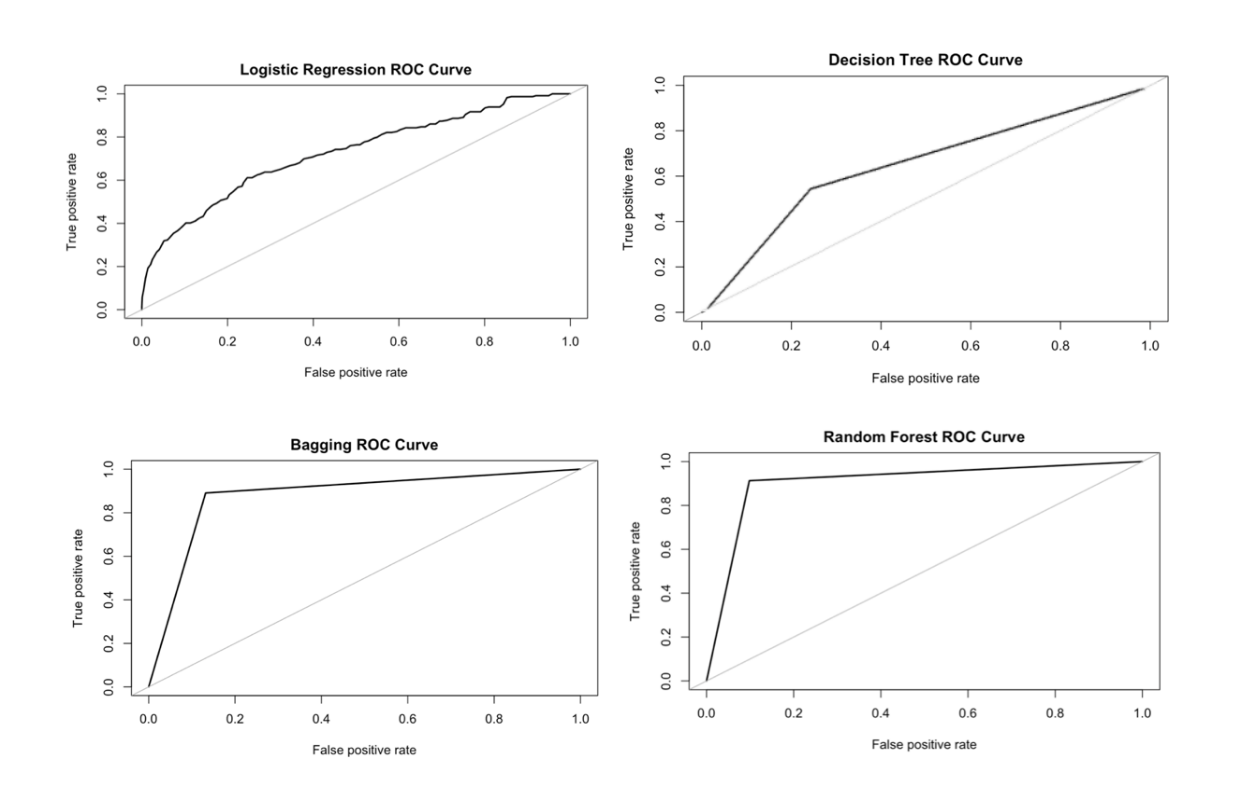
The sensitivity and specificity of the logistic regression, decision trees, bagging, and random forest methods are important considerations. For this scenario, we define the model's sensitivity as the proportion of employees who leave Canterra correctly predicted by the model. For Canterra, maximizing the model sensitivity should be the top priority. The company must accurately identify who is susceptible to attrition so that management can take corrective action. By picking up on trends across employees who attrit, management can tailor Canterra's corporate policies and work environment to appeal to these at-risk employees. While of secondary importance, Canterra also needs a model with high specificity. The model's specificity is the proportion of employees who stay at Canterra accurately predicted by the model to stay. Specificity provides the company with an idea of which group of employees are stable in their jobs and not at risk of leaving. When implementing new changes to target employees at risk of attrition, Canterra must be sure to avoid alienating happy employees. It would be unproductive to sacrifice the stability of the employees who stay to appease the employees who leave.

As seen in Table 2 below, the bagging model had the highest sensitivity, closely followed by random forest. Random forest had the highest specificity and area under the ROC curve. We also compared RMSE values for each model and confirmed that random forest has the lowest value.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Logistic Regression | Decision Tree | Bagging | Random Forest |
| Sensitivity | 0.624 | 0.541 | 0.917 | 0.912 |
| Specificity | 0.730 | 0.763 | 0.846 | 0.905 |
| AUC | 0.725 | 0.652 | 0.881 | 0.907 |
| RMSE | 0.438 | 0.362 | 0.340 | 0.198 |

**Table 2: Model Performance**

Figure 4 contains the ROC curves for the four models described above. By plotting sensitivity against specificity and comparing the area under the curve, we can see that the bagging and random forest models are the most accurate.

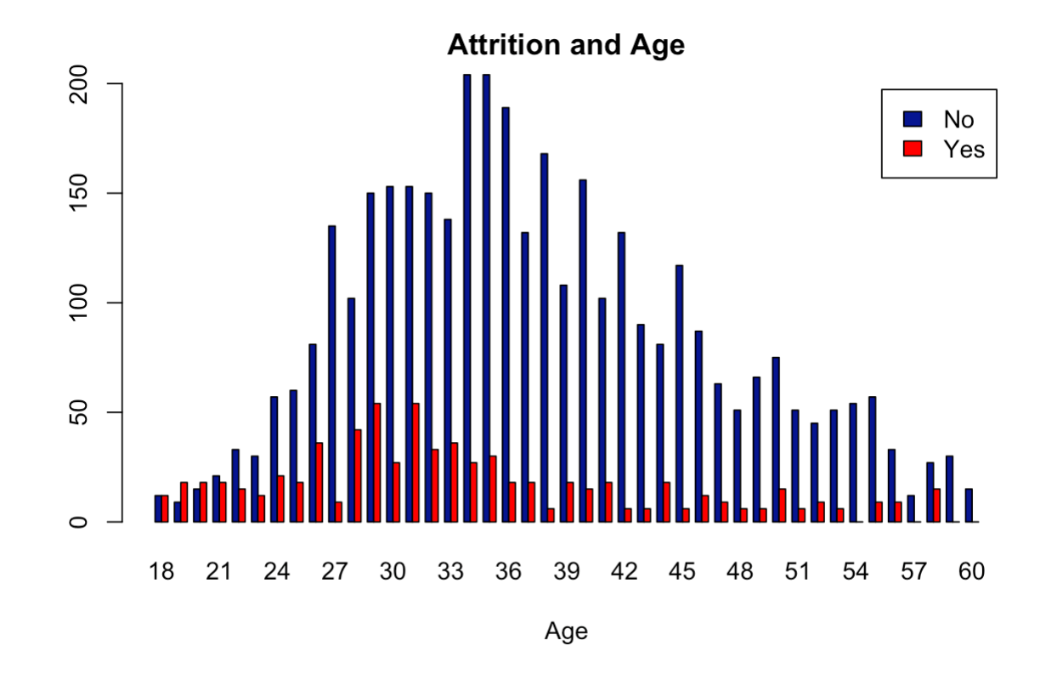


**Figure 4: ROC Curves**

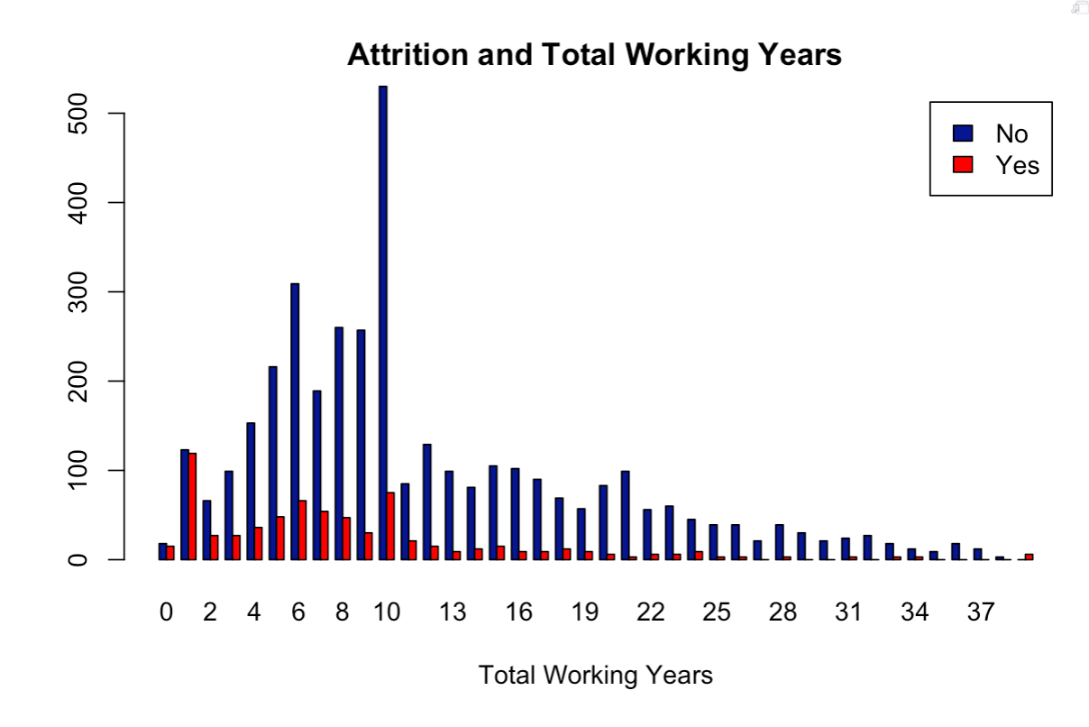
For all the reasons mentioned above, we have decided to utilize the random forest model to present the results to the client. It provided superior results with regards to prediction and error metrics.

**Recommendations**

Total working years are a significant indicator of employee attrition. Employees who have been working less than three years are more likely to leave Canterra. In figure 6 below, we see that nearly all employees with total working years under two are leaving the company, which may also be connected to age. In figure 5, we see that almost 50% of the younger employees are also leaving the company. As age increases, total working years increases; given this relationship, it is vital to notice that younger, less experienced workers are more likely to leave the company. Canterra's HR department could aim to hire experienced workers who may exhibit lower attrition rates. However, this strategy could have unintended consequences such as increased salary costs and lower innovation rates. Older employees may be more willing to stay, but Canterra should consider how a more aging workforce with lower turnover could lead to stagnation.



**Figure 5: Attrition vs. Age**



**Figure 6: Attrition vs. Total Working Years**

Similarly, Canterra could offer incentives to younger employees to motivate them to stay with the company and increase the average tenure. Increased vacation time, flex time, and stock options may encourage employees to remain at the company. Further analysis could explore what younger employees are looking for in an employer and how Canterra can better fit and meet those needs. Exit surveys may also provide valuable insight into why former employees are leaving and the future opportunities they have found.

It would be straightforward for Canterra to compare their pay scales to those of their competitors at each experience level regarding the income variable. Equivalence of pay combined with the above incentive structures could reduce the motivation for employees to leave Canterra for purely financial reasons.

Future analysis should also explore departments within the company that have low attrition rates. With this data, Canterra could observe teams' practices with higher retention and implement their strategies across the company. These strategies vary in complexity and scope. If implemented correctly, they may go a long way in reducing Canterra's attrition rate and ensuring Canterra is a great place to work.

**Appendix**

1. For reproducibility, the data was divided into training and test sets using the following seed.

   *set.seed(456)*

   With a turnover rate of 15% there was more data available for employees who stay. Down-sampling was used to ensure that the data included an even amount of data for both employees leaving and staying.

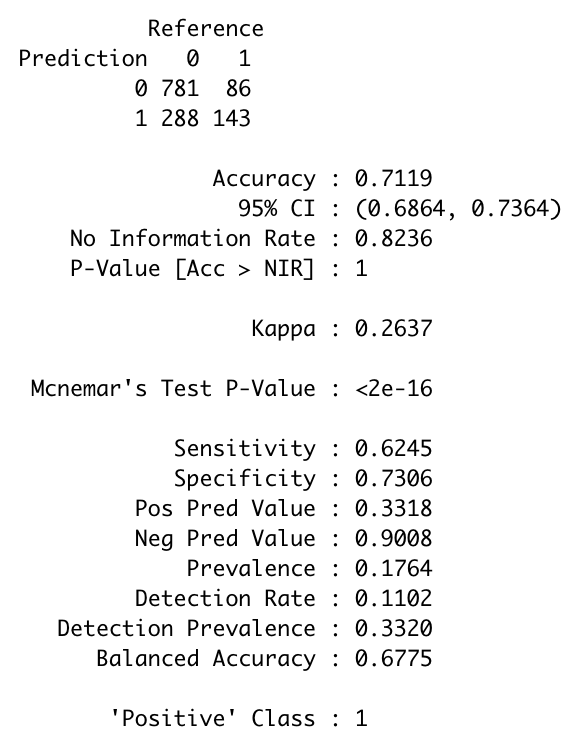
   Using the “under” method, 1000 data points were selected. The training data set resulted in 530 data points for employees who stayed and 470 data points for employees leaving.

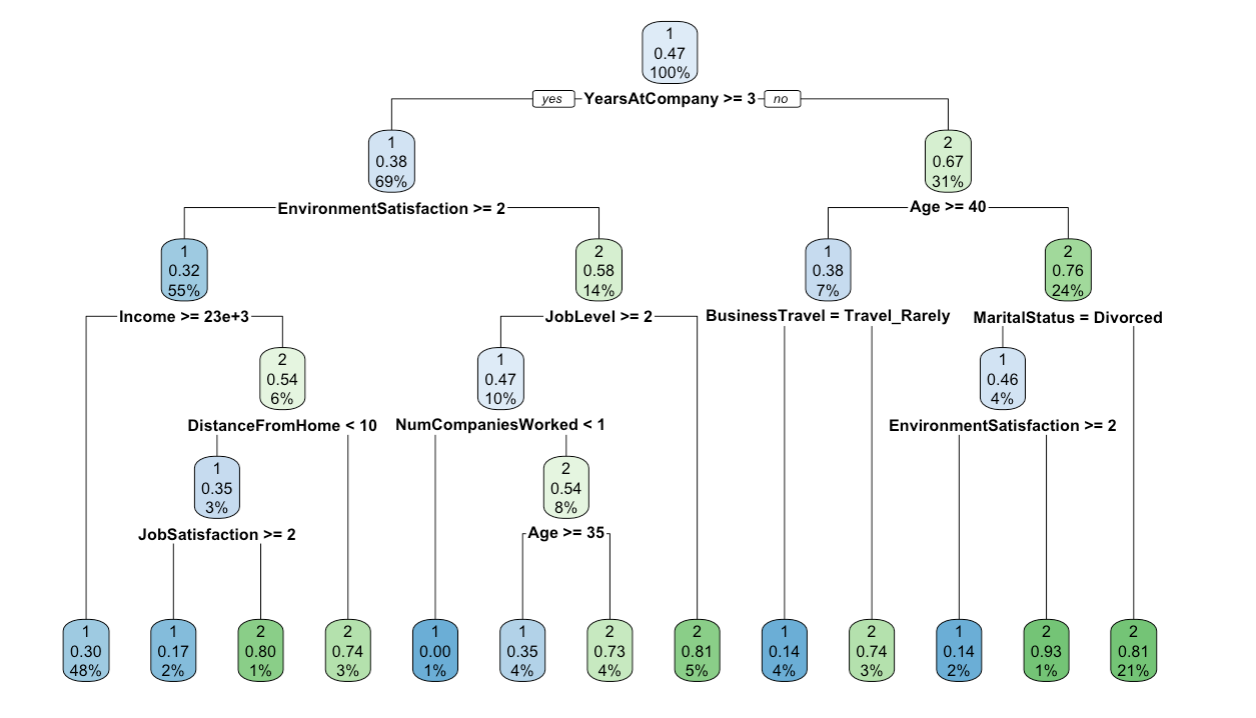
   0 (no) = stays with the company

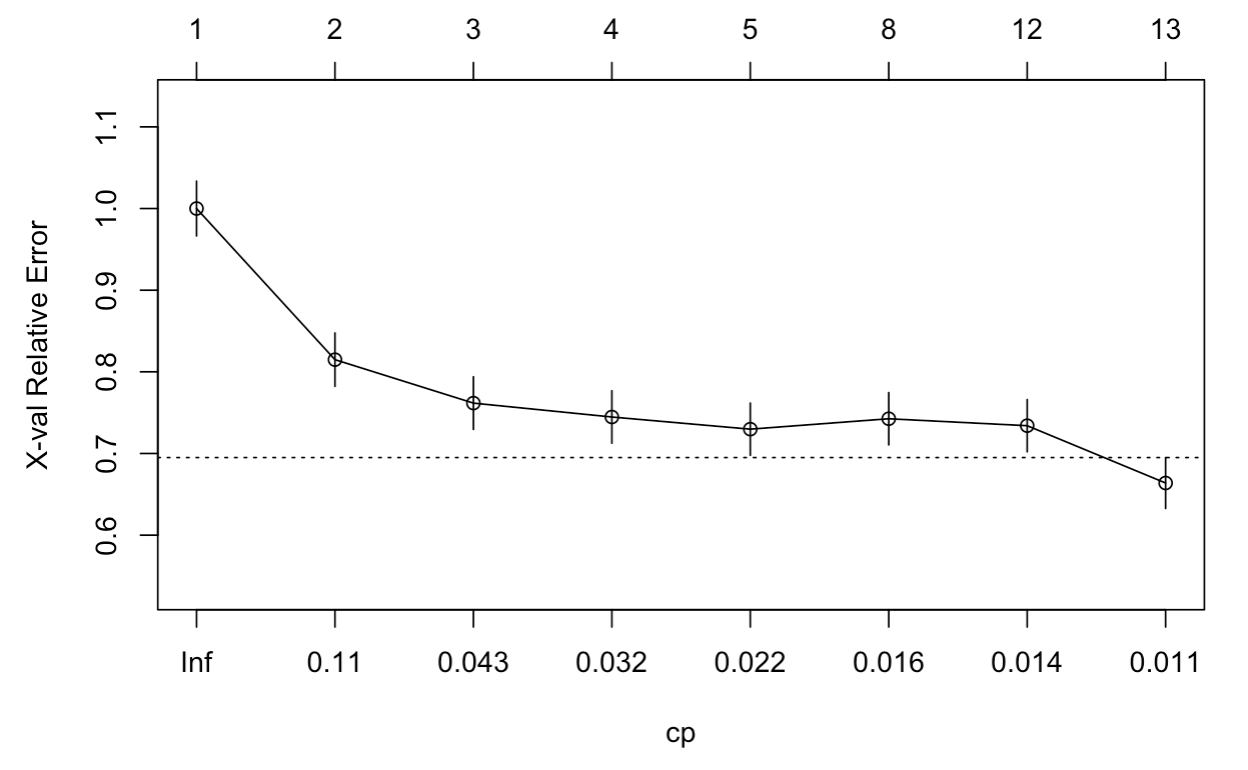
   1 (yes) = leaves the company [↑](#endnote-ref-2)
2. The logistic regression model identified frequent travel, marital status of single, and environment satisfaction as the most significant variables.

   |  |  |  |
   | --- | --- | --- |
   | COEFFICIENT | ESTIMATE | P-VALUE |
   | (INTERCEPT) | 2.63 | 4.76e-10 |
   | AGE | -0.04 | 1.52e-06 |
   | JOBSATISFACTION | -0.22 | 0.000616 |
   | FREQTRAVEL | 0.71 | 1.09e-05 |
   | SINGLE | 0.78 | 4.33e-08 |
   | NUMCOMPANIESWORKED | 0.07 | 0.018772 |
   | TRAININGTIMESLASTYEAR | -0.15 | 0.009466 |
   | YEARSWITHCURRMANAGER | -0.13 | 5.99e09 |
   | ENVIRONMENTSATISFACTION | -0.18 | 0.003689 |

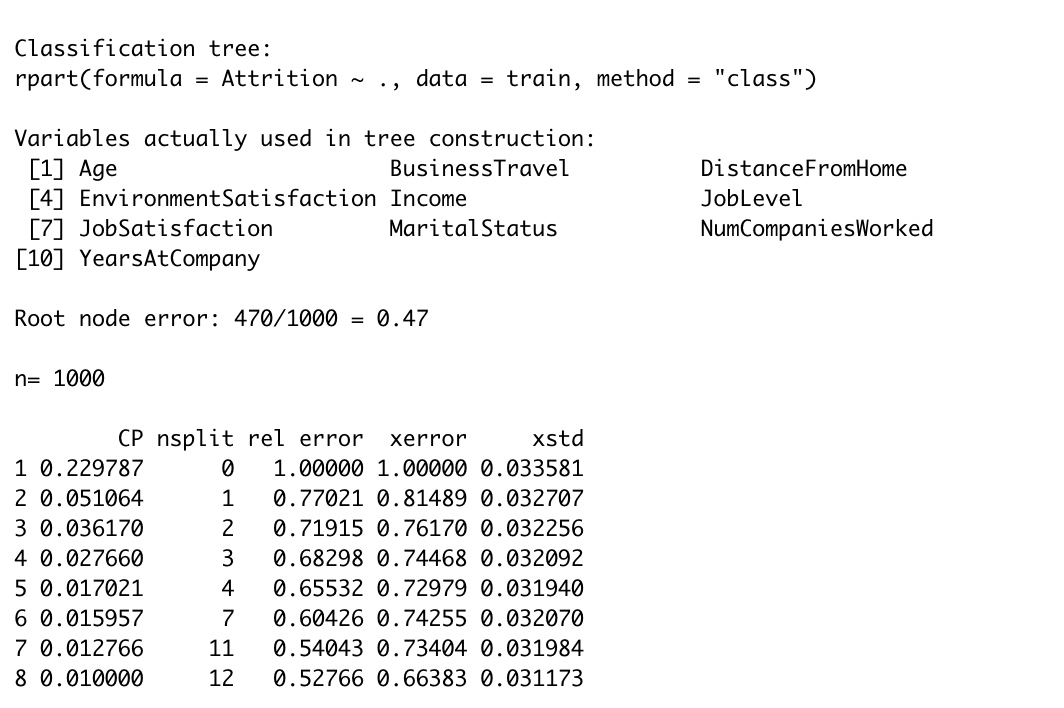
   [↑](#endnote-ref-3)
3. The confusion matrix and accuracy values for the logistic regression model are displayed below.

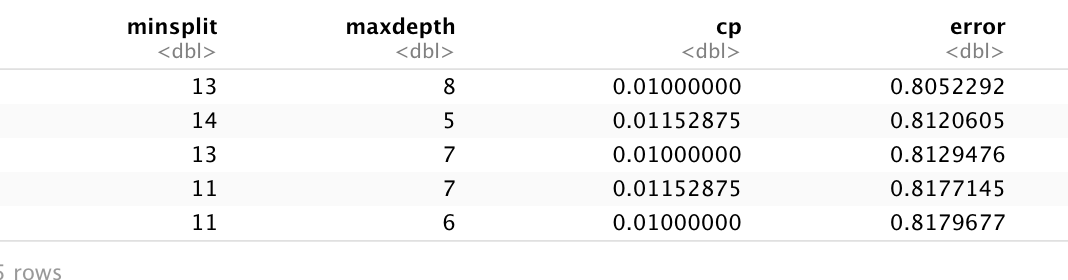
    [↑](#endnote-ref-4)
4. Classification decision tree modeling employee attrition at Canterra, obtained using rpart package. The tree used 1000 data points with years at company being identified as the most significant variable and the initial split, followed by environment satisfaction and age. Income, marital status, job level, and number of companies worked were also of significance.

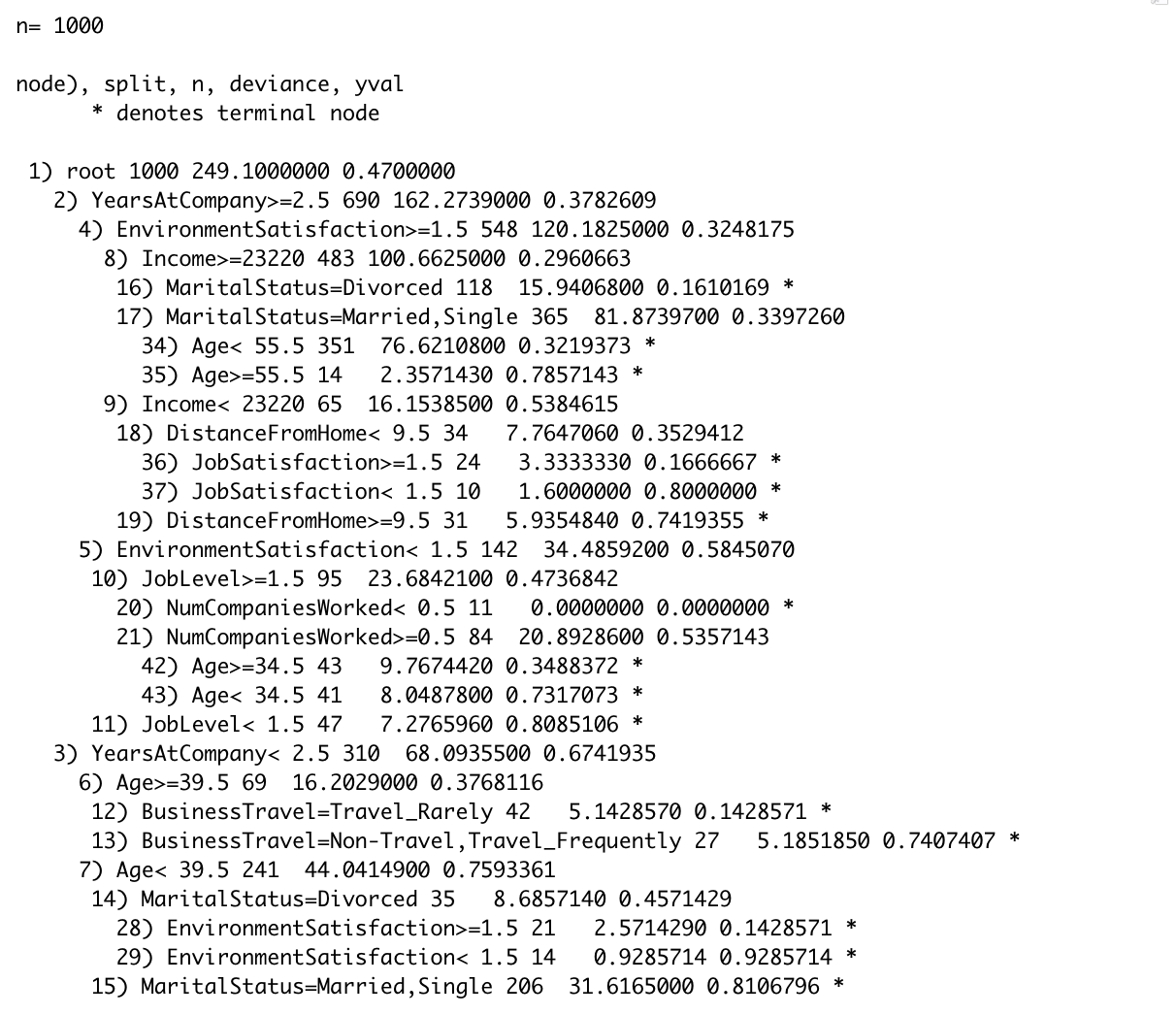
    [↑](#endnote-ref-5)
5. The plot below is the cross-validation error with cost complexity. We can see that the error is minimized with 12-13 terminal nodes.

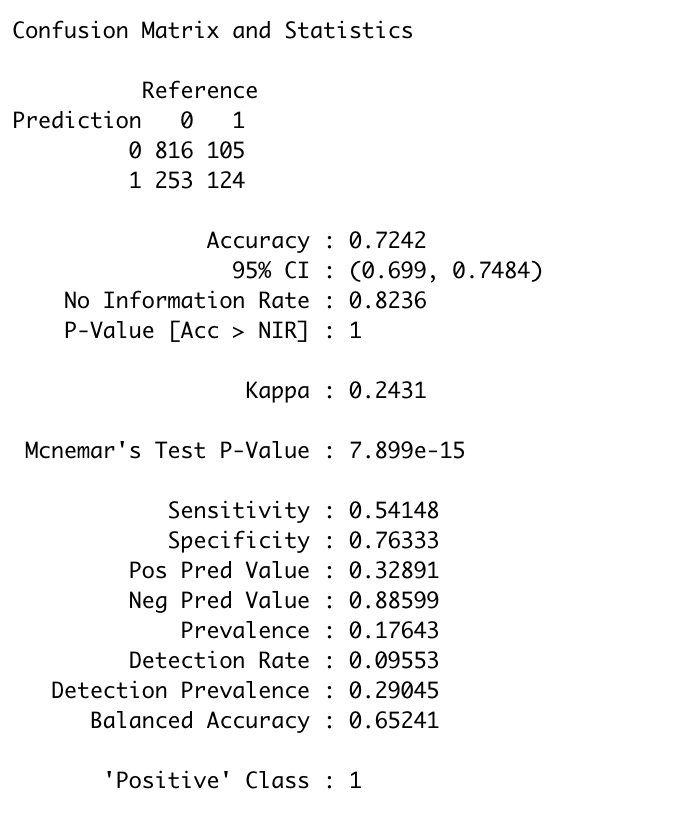
   The xerror value is minimized at nsplit =12 with a cp value of 0.01 and an xerror value of 0.6638.

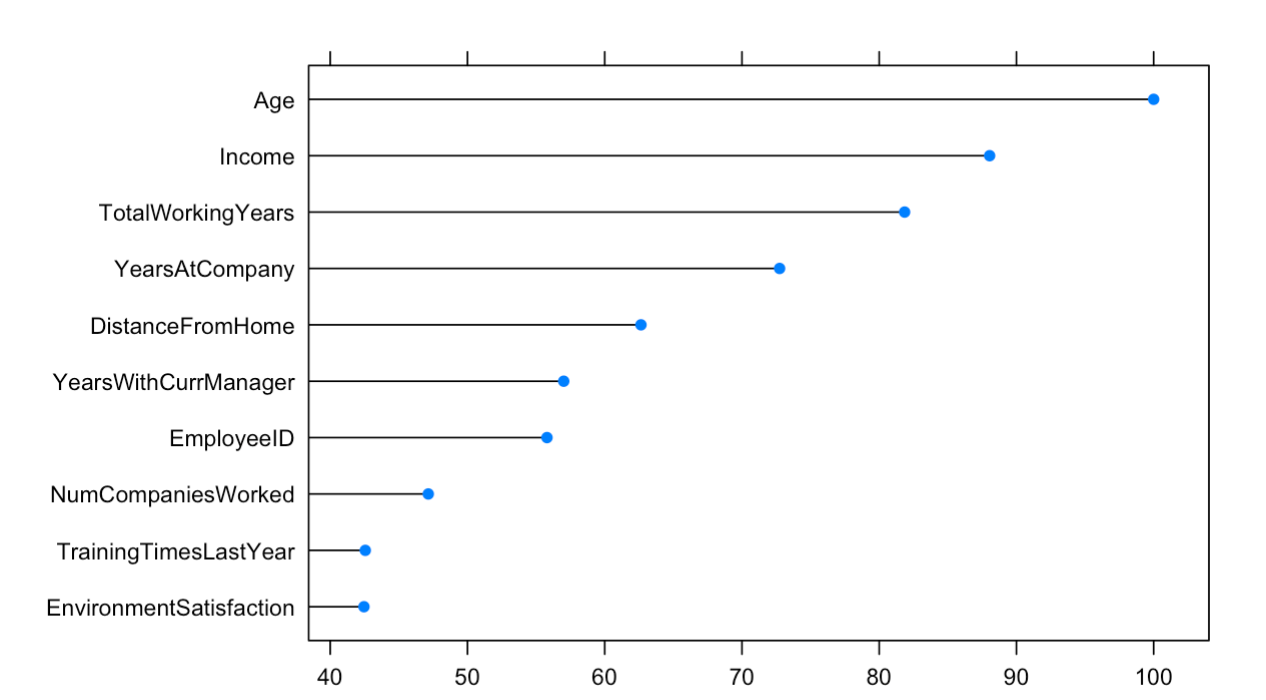
    [↑](#endnote-ref-6)
6. The grid search was run using a range of 8-15 for minsplit and 3-8 for maxdepth. The table below shows the 5 lowest error values that were returned.

    [↑](#endnote-ref-7)
7. The optimal decision tree and the breakdown between nodes are below.

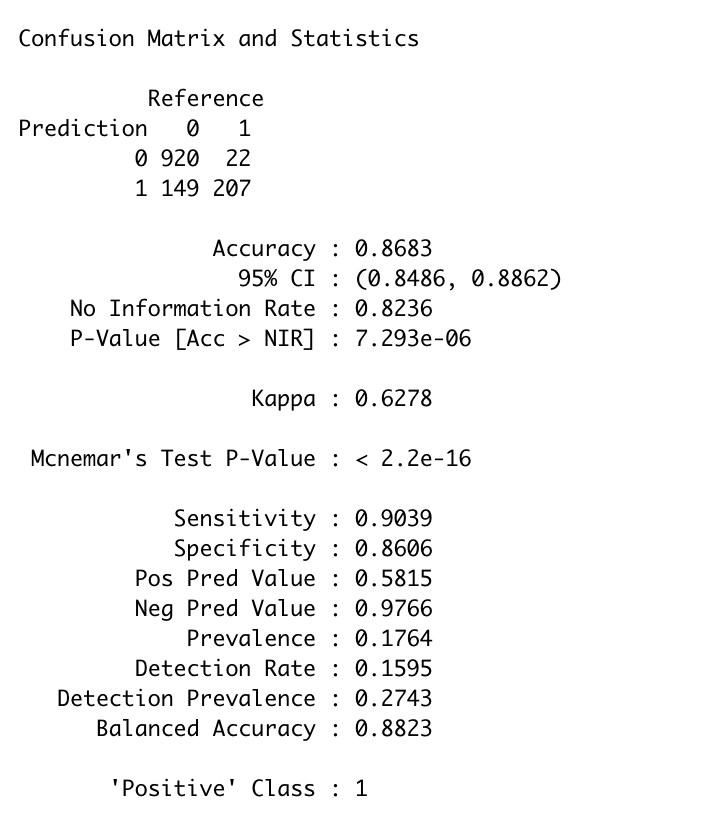
   

   The accuracy of this model is shown below.

    [↑](#endnote-ref-8)
8. The bagging method identified age, income and total working years as the most significant variables.

   The accuracy of the bagging method is shown below.

    [↑](#endnote-ref-9)