Machine Learning Final Project

Munerah Al-Fayez

Prof. Murali Shanker

Abstract

The project describes the real-world data for employees in a company that faces an issue of employee attrition. The human resource department needs to keep its employees from leaving. The approach used for that issue is to identify machine learning techniques that help the human resource department to understand which factors contribute to employee attrition.

Contents

Introduction

A large company employs about 4000 employees. However, yearly, around 15% of its employees leave the company and need to be replaced with new talents. The management believes that this level of attrition (employees leaving, either on their own or because they got fired) affects the company negatively, for the following reasons

1. Projects of former employees are delayed, making it harder to fulfill deadlines, resulting in a loss of reputation among customers and partners.
2. For the aim of hiring new personnel, a sizable department must be maintained.
3. New employees are frequently required to be trained for their jobs.

Business Problem

The human resources department personnel want to know what changes they should make in their workplace, to get most of their employees to stay.

Machine Learning Techniques

Two machine learning techniques are used.

Rapt (Recursive Partitioning and Regression Trees)

Recursive Partitioning and Regression Trees or Rpart is a popular machine learning model that can be used for both classification and regression applications. For this case, the identifier is

a factor with two levels whether yes or no, which means the Rpart decision tree classifier is the choice.

Random Forest

Random Forest is a commonly used machine learning algorithm that combines more than one tree output for more accurate prediction.

Rpart and Random Forest can help to understand what factors the human resources department personnel should focus on to minimize attrition. And, which of these factors are the most important and need to be considered. There are also additional reasons regarding choosing these techniques which are summarized in the good visualization, which is easy to understand without a need for analytical background. Moreover, it can handle both numerical and categorical variables.

Analysis

About data

Variable	Meaning	Levels
Age	Age of the employee	
Attrition	Whether the employee left in the previous year or not	
Business travel	How frequently the employees traveled for business purposes in the last year	
Department	Department in company	
DistanceFromHome	Distance from home in km	
Education	Education Level	1 'Below College'
		2 'College'
		3 'Bachelor'

		4 'Master'
		5 'Doctor'
EducationField	Field of education	
EmployeeCount	Employee count	
EmployeeNumber	Employee number/id	
EnvironmentSatisfaction	Work Environment Satisfaction Level	1 'Low'
		2 'Medium'
		3 'High'
		4 'Very High'
Gender	Gender of employee	
JobInvolvement	Job Involvement Level	1 'Low'
		2 'Medium'
		3 'High'
		4 'Very High'
JobLevel	Job level at the company on a scale of 1 to 5	
JobRole	Name of job role in the company	
JobSatisfaction	Job Satisfaction Level	1 'Low'
		2 'Medium'
		3 'High'
		4 'Very High'
MaritalStatus	Marital status of the employee	
MonthlyIncome	Monthly income in rupees per month	
NumCompaniesWorked	Total number of companies the employee has worked for	
Over18	Whether the employee is above 18 years of age or not	
PercentSalaryHike	Percent salary hike for last year	
PerformanceRating	Performance rating for last year	1 'Low'
		2 'Good'
		3 'Excellent'
		4 'Outstanding'
RelationshipSatisfaction	Relationship satisfaction level	1 'Low'
		2 'Medium'
		3 'High'
		4 'Very High'
StandardHours	Standard hours of work for the employee	
StockOptionLevel	Stock option level of the employee	

TotalWorkingYears	Total number of years the employee has worked so far	
TrainingTimesLastYear	Number of times training was conducted for this employee last year	
WorkLifeBalance	Work life balance level	1 'Bad'
		2 'Good'
		3 'Better'
		4 'Best'
YearsAtCompany	Total number of years spent at the company by the employee	
YearsSinceLastPromotion	Number of years since last promotion	
YearsWithCurrManager	Number of years under current manager	

This is the real-world data collected from Kaggle. Kaggle is a platform for data scientists to compute and contribute where real-world datasets can be found.

There are 4382 rows and twenty-four columns (variables) in this dataset. The variables contain the employee age, whether the employee left the previous year or not, education field, distance from the employee home, years at the company. The dependent variable for this analysis is Attrition, and it is a factor with two levels yes and no. The rests are the predictors or the independent variables.

Building a decision tree model

Figure number one shows a decision tree model using the(Rpart) library. By choosing the dependent variable (Attrition) and the other twenty-three variables, including age, business travel, department, distance from home, etc. as independent variables. Only six variables out of

twenty-three, as well as seven nodes, are represented in the tree. Those variables are significant predictors for the dependent variable attrition. However, to prevent tree overfitting, further processing is needed.

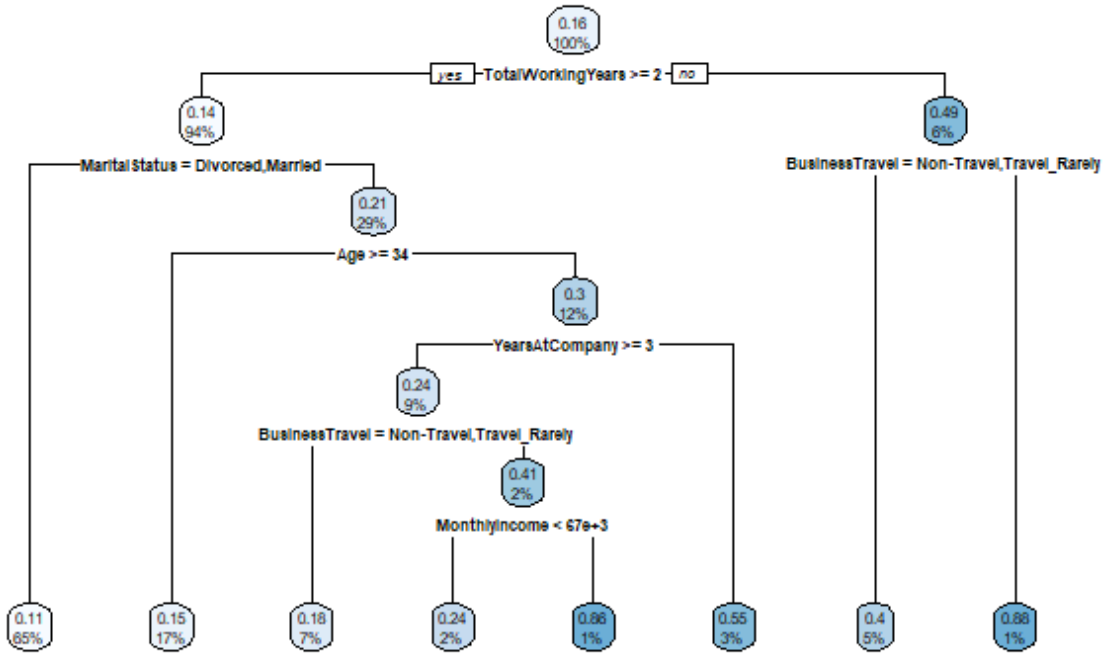

Figure 1 *R-Part Tree*

Overfitting

Once the number of splits is increased, the error rate in the training data decreases. This is because the tree finds different splits and can fit all the data points in the training data, so the number of errors on the data set is reduced. This results in capturing unnecessary noise data. This case is called overfitting.

Pruning the tree

Choosing a tree with an optimal number of splits is essential to avoid overfitting issues. Post-pruning the tree is a technique to reduce the size of the tree until the optimal size is reached.

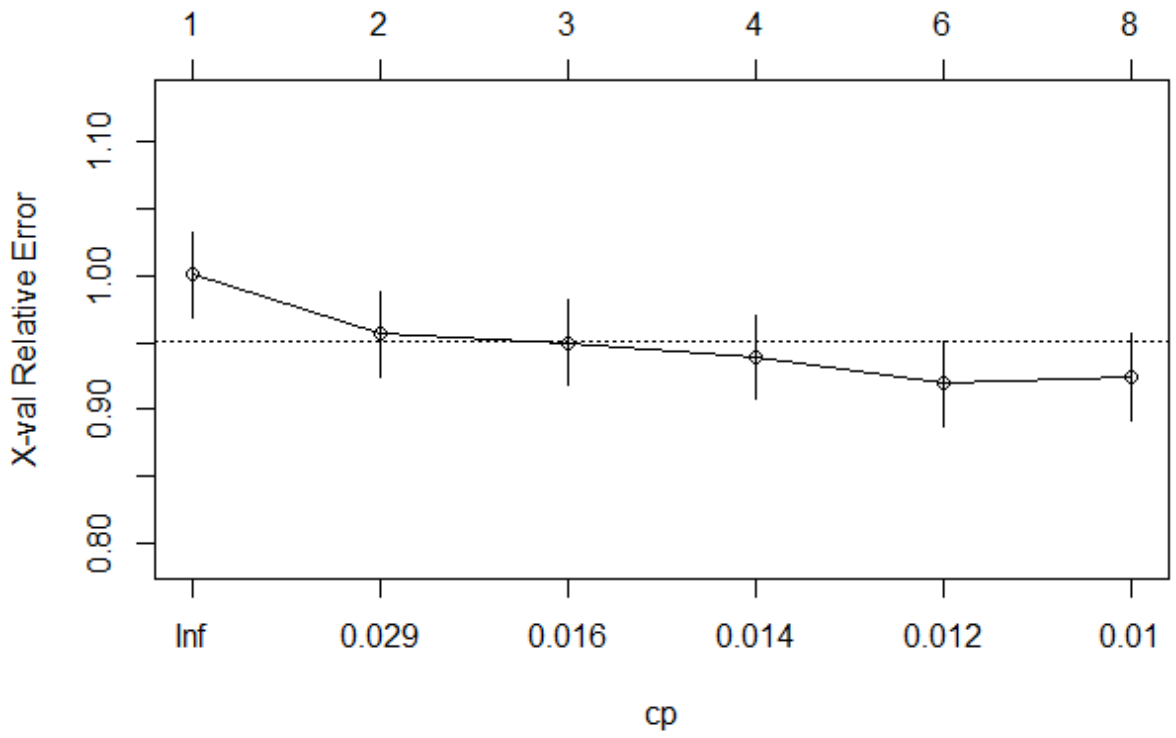

Figure 2 Cp vs Relative error

Figure number two shows a plot for the cost complexity(cp) versus relative error values, and the size of the tree as well on the top of the graph. Using this graph is to identify a point where the tree has a low error, but at the same time low cp as well.

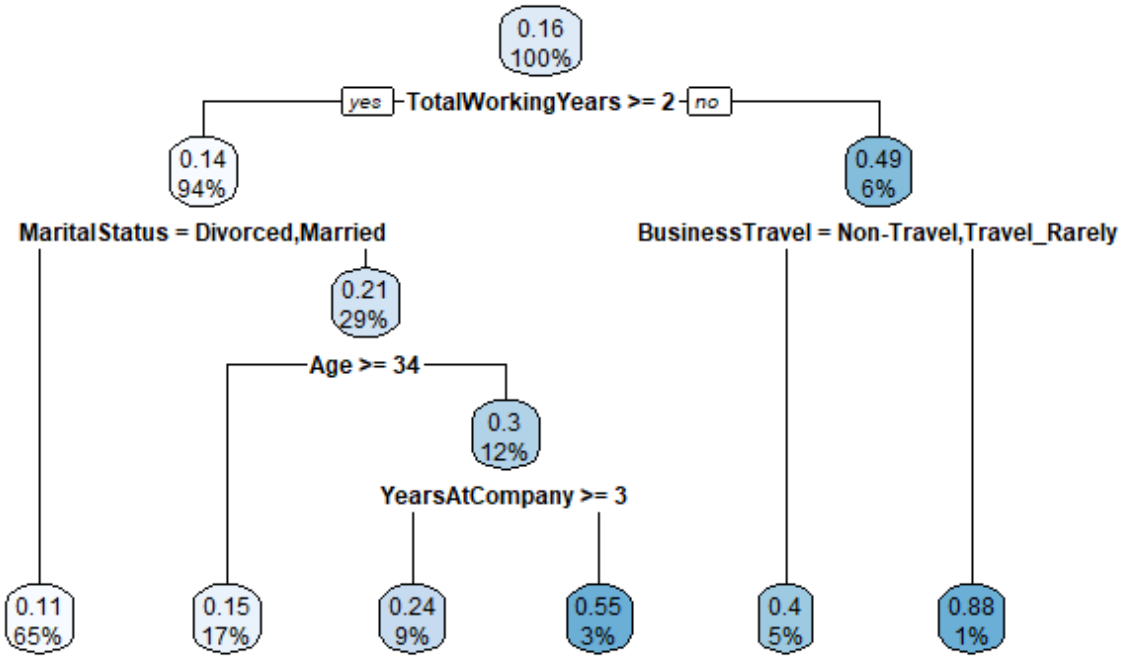

Figure 3 R-Part Pruned Tree

This tree (Figure number three) results in using CP value = 0.012 which has the lowest number of errors. As we notice there are five nodes and six terminals. By default, the left branch is selected if the result is positive.

Some statistics can be inferred from this decision tree.

Employees who are divorced and have two or more total working years have a probability of 65 % to leave the company.

Employees who have three or more years at the company and their age less than 34 and married have a probability of 9% to leave the company.

Employees who are married and have two or more total working years have a probability of 29 % to leave the company.

However, the model must have a minimum error factor with better cp as well.

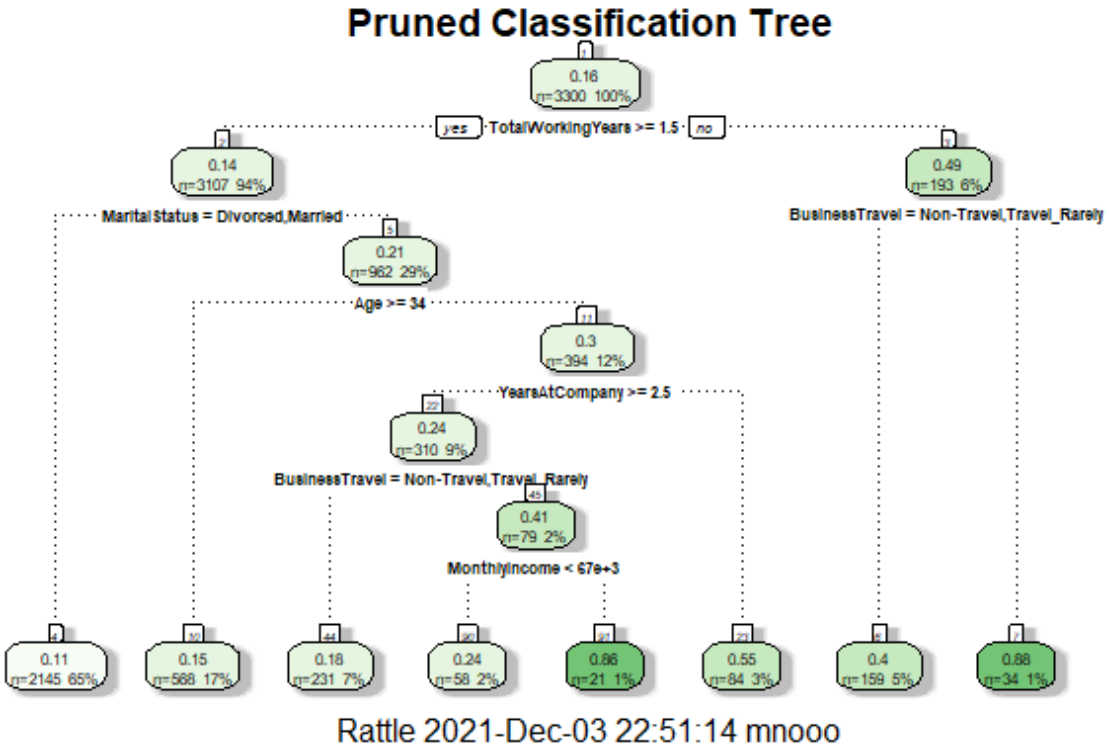

Figure 4 Pruned Tree

Using different parameters, this tree (Figure number four) has a balanced error and cp value as well. This tree has the optimal number of splits with seven nodes and eight terminals. By default, the left branch is selected if the result is positive. The tree shows only seven variables (nodes) out of twenty-three, which are significant predictors for the dependent variable Attrition. Those variables are total working years, marital status, age, years at the company, business

travel, and monthly income. We may notice that business travel showed up twice in the tree, but with different conditions for each. The value that split the monthly income is shown as 67+3 which means 67*1000 equals 67,000 rupees. (fancyrpartplot) the function gives an easier and clearer representation, where the number of observations falls in each node (n) as well as the probability.

Some statistics can be inferred from this decision tree.

Employees who are divorced and have one and a half or more total working years have a probability of 65 % to leave the company with 2145 observations.

Employees who have never traveled and worked for the company for more than two and half years and their age less than 34 and married have a probability of 7% to leave the company with 231 observations.

Employees who have never traveled and have less than one and a half working years have a probability of 5% to stay at the company with 159 observations.

```
Variable importance
   TotalWorkingYears                Age       BusinessTravel
                  36                 15                   12
      YearsAtCompany       MaritalStatus        MonthlyIncome
                  10                  9                    9
  NumCompaniesWorked             JobRole YearsWithCurrManager
                   3                  2                    2

Node number 1: 3300 observations,    complexity param=0.04994041
```

Figure 5 Variable importance

Using the summary function, figure number five shows some information about the tree which includes the importance of the variables, which shows the age is the most variable that affects the dependent variable Attrition after the total working years, then the business travel.

Variable Importance using random Forest

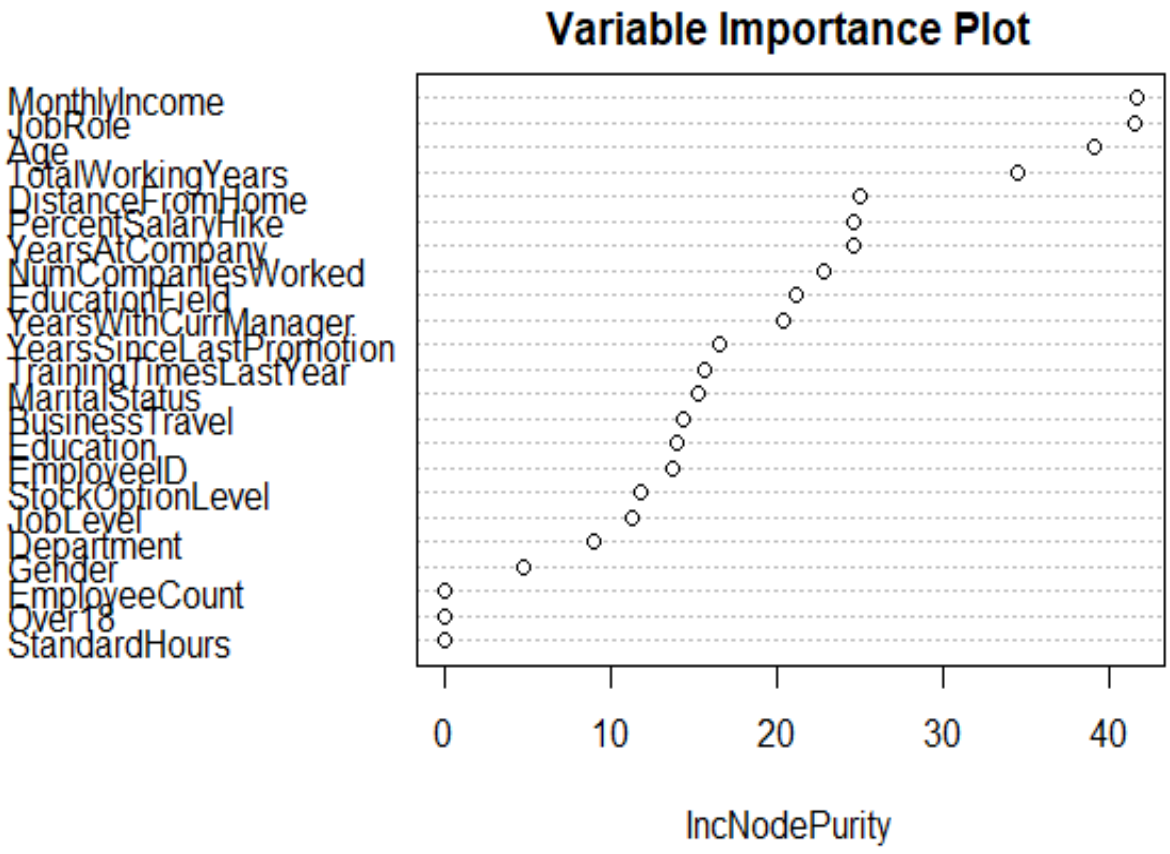

Figure 6 variable Importance using Random Forest

This plot (figure number six) shows the importance of the variables that affect the dependent variable attrition. It shows the monthly income is the most important variable, then by order of

job role, age, total working years, and the distance from home. Gender, department, and job level are the least important variables. Employee count, over 18, standard hours are not important in this case.

Performance evaluation

Performance evaluation of a model is very critical to understand whether the model is a good fit, improve the model accuracy, or compare between more than one model.

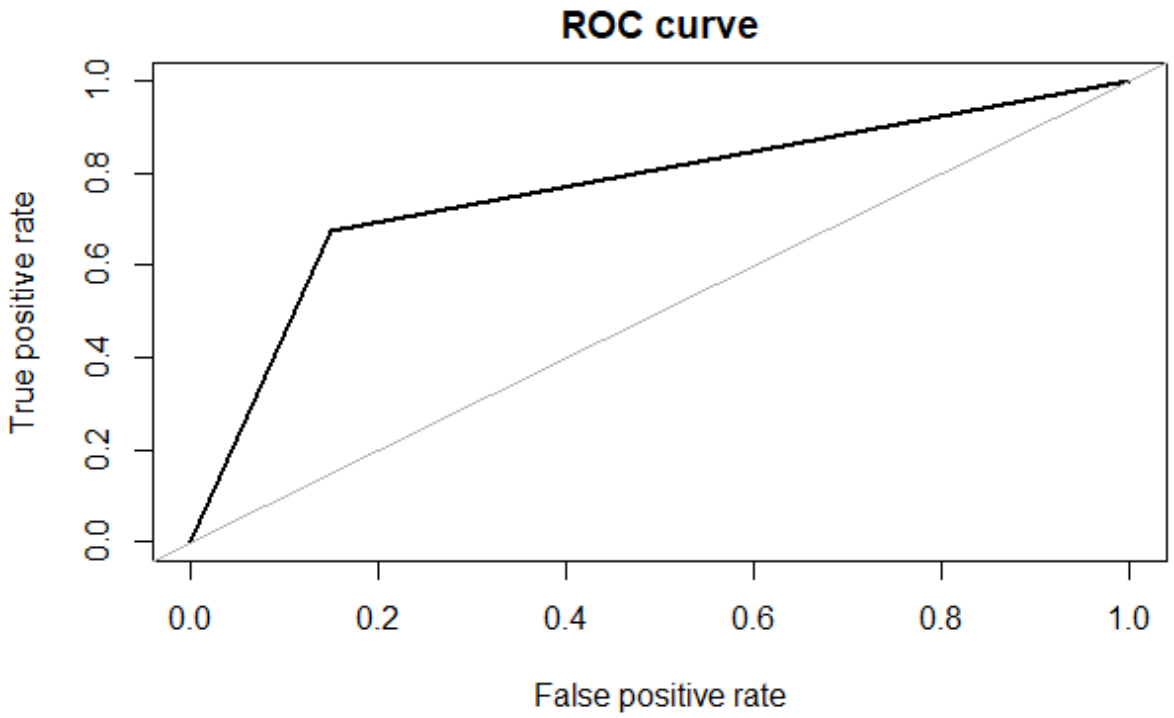

Figure 7 Rpart ROC

Using Roc curve in figure number seven, the accuracy (area under the curve) of the Rpart decision tree model is equal to 76 %

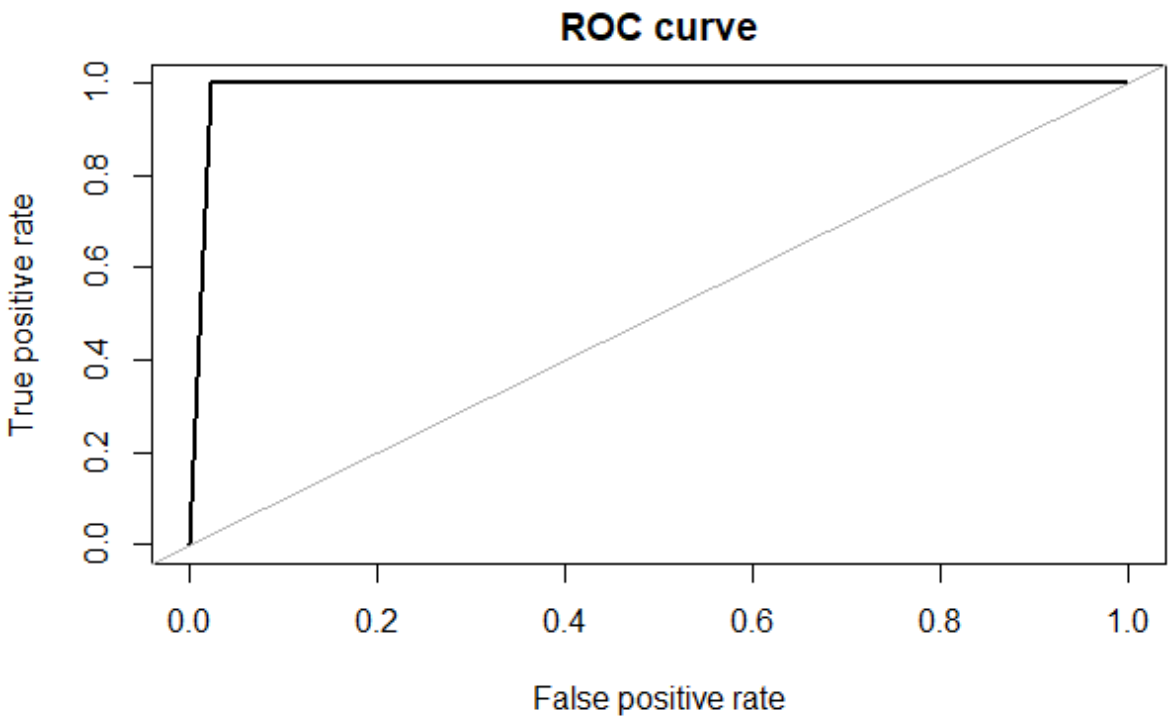

Figure 8 RandomForest ROC

Using the Roc curve in figure number eight, the accuracy (area under the curve) of the Random Forest model is equal to 99% which is very high and significant. This means it is highly suggested to rely on this model output.

Conclusion

By analyzing the output from the random forest model and get the most efficient factors that affect employee attrition, which are Monthly Income, Job Role, Age, Total Working Years, Distance from Home, Percent Salary Hike. Also, by linking those factors to the decision tree, those are a list of solutions ordered by its priority.

1. Raise the salary for employees.

2. Prepare a detailed study about the relationship between job titles and employees leaving.
3. look for underlying causes why experienced employees whose ages above 33 tend to leave the company.
4. Select candidate employees who live nearby the company or provide transportation allowances.

The human resource personnel should consider those points according to the budget and other resources they have, also the company's needs.

GitHub Link

Click here

References

Choudhary, Vijay. (2018). HR Analytics Case Study. *kaggle*. https://www.kaggle.com/vjchoudhary7/hr-analytics-case-study

Meltzer, Rachel. (2021, July 15). What Is Random Forest *careerfoundry*? https://careerfoundry.com/en/blog/data-analytics/what-is-random-forest/#what-is-random-forest