# Predictive Model for Employee Attrition - Phase 2

## SAXA 3

Mike Johnson | Kesh Kamani | Ryan Mathis | Khushi Patel | Andrew Singh

OPAN 6607 - Machine Learning I
Final Project
SAXA 3
Mike Johnson | Kesh Kamani | Ryan Mathis | Khushi Patel | Andrew Singh

## Business Problem & Context

As external consultants, we support our client, Canterra, with data-driven solutions to address challenges related to managing their current talent pool due to an attrition rate of 15% per year (affecting approximately 600 employees) leaving the company each year. For Canterra, one of their goals is to replace lost talent with new talent each year. Leadership believes their level of attrition is attributed to the following three organizational factors: 1) prior project timelines set by past employees are not maintained, which damages Canterra's existing relationship with their consumers and partners; 2) adequately sized departments have to be maintained with new talent, and 3) incoming talent has to be trained to learn about their new roles and accompanying company policies.

Our aim in this report is to put forth a set of recommendations that shed light on the leadership's initial hypothesis of the relationship between higher job satisfaction and a higher number of total working years reducing employee attrition while at the same time supporting an in-depth analysis of the marketing department's ask to understand how demographic factors affect employee attrition.

## Executive Summary & Recommendations

Phase 2 of this project evaluated multiple predictive models to understand the factors that Canterra management should focus on in order to curb attrition. In addition to a logistic regression model used in phase 1, decision trees, bagged decision trees, and random forest models were evaluated.

From the models evaluated, our team recommends that the logistic regression model to move forward for use by the Canterra management team. Our recommendation is based on the interpretability of variables on attrition in the model, which we believe can be translated into actionable insights for stakeholders.

The main trade-off with going with the logistic regression model is predictive performance. The logistic regression model has a 74% chance of correctly distinguishing between attrition and non-attrition employees. In comparison, the other models performed better, with the random forests model achieving 97%. While the other models had better predictive performance, these models are more problematic for stakeholders to interpret. Given that this project is more explanatory than predictive, we believe the trading predictive performance for interpretability is reasonable.

<u>Recommendations</u>

Based on our analysis, the following is recommended to curb attrition at Canterra:

- **Provide additional support for employees who need to travel.** This includes improved travel allowances and investing in wellness programs to help alleviate the stress associated with travel. In addition, technology solutions such as video conferencing should be considered to reduce the need for travel.

- **Develop targeted employee engagement strategies for single employees to help them feel more connected to the organization**. This could include social events, employee mentorship programs, and community-building initiatives.
- **Conduct regular employee satisfaction surveys to identify pain points and areas for improvement**. Invest in programs that improve job satisfaction, such as employee recognition, career development opportunities, and initiatives to improve employee work-life balance.
- **Develop programs to incentivize long-term commitment to the organization.** This includes offering clear career progression or promotional paths, tailored professional development plans, and retention bonuses.

## Logistic Regression

A logistic regression model was developed in phase 1 of the project. To summarise, logistic regression is a method for binary classification problems used to predict a certain class's probability for a given input. In the context of this project, a logistic regression model predicts the probability of attrition for a given employee.

The logistic regression model selected (figure 1) had an Area Under Curve (AUC) value of 0.75 on testing data, meaning there is a 75% chance that the model will correctly distinguish between attrition and non-attrition employees.

When measuring the average marginal effect (AME) of the independent variables on attrition, the largest increase in the likelihood of attrition includes Business Travel (Frequent) and Marital Status (Single). Conversely, the largest decreases in the likelihood of attrition include Environment Satisfaction (High) and Job Satisfaction (Very High). For all AMEs, reference Figure 3 in the appendix.

Based on the logistic regression model, the following was recommended in phase 1 to curb attrition:

- **Provide additional support for employees who need to travel.** This includes improved travel allowances and investing in wellness programs to help alleviate the stress associated with travel. In addition, technology solutions such as video conferencing should be considered to reduce the need for travel.
- **Develop targeted employee engagement strategies for single employees to help them feel more connected to the organization**. This could include social events, employee mentorship programs, and community-building initiatives.
- **Conduct regular employee satisfaction surveys to identify pain points and areas for improvement**. Invest in programs that improve job satisfaction, such as employee recognition, career development opportunities, and initiatives to improve employee work-life balance.

- **Develop programs to incentivize long-term commitment to the organization.** This includes offering clear career progression or promotional paths, tailored professional development plans, and retention bonuses.

## Decision Trees

Another approach evaluated for this project was a decision tree model. Decision trees use a tree-like structure to deliver consequences based on input decisions. The main advantages of decision trees are that they are easy to understand and can capture non-linear relationships in the data. On the other hand, they can easily overfit, as was the case in the model selected.

The decision tree model developed for this project (figure 4) has an area under curve (AUC) value of 0.79 on testing data. While an improvement over the logistic regression model, the tree is overly complex and challenging to read. This is due primarily to the complexity parameter that was selected: 0.001. Smaller values allow for more splits, while larger values result in a more straightforward tree. In this model's case, more extensive complexity parameters resulted in a lower AUC (figure 5). As a result, the more complex model was selected.

While complex, the following variables of importance (>80) were identified (figure 6): Total Working Years, Years with Current Manager, Age, and Income.

Based on the decision trees, the following is recommended:

- **Develop programs to incentivize long-term commitment to the organization.** This includes offering clear career progression or promotional paths, tailored professional development plans, and retention bonuses.
- **Focus on improving the quality of management.** Provide training for managers to ensure they are effective leaders who can support and retain their team members.
- **Develop targeted employee engagement strategies for younger employees to help them feel more connected to the organization**. This could include social events, employee mentorship programs, and community-building initiatives.
- **Conduct regular salary reviews to ensure that compensation is competitive**. Consider performance-based incentives and bonuses to reward top performers.

## Bagged Decision Trees

Bagged decision trees, also called bootstrap aggregating, is an ensemble learning technique to improve the performance of and stability of decision trees. The main idea with this approach is to reduce the model variance by training multiple versions and then averaging the predictions.

The bagged decision tree model developed had an area under curve (AUC) value of 0.96 on testing data - a remarkable improvement over previous models. Variables of importance (>80) include (Figure 9): Age, Income, and Total Working Years. In this model, as we were building the initial model, we focused on re-factoring several variables to become factor data types and removed the Distance From Home variable in the model, as these modifications would improve the performance of the model. To improve the decision tree model, we incorporated the use of a

tuned random forest where we changed the number of decision tree leaves, improving the validity of the model. We found that 100 decision tree leaves were the optimal amount to receive a performance model.

Based on the bagged decision trees model, the following is recommended:

- **Develop targeted employee engagement strategies for younger employees to help them feel more connected to the organization**. This could include social events, employee mentorship programs, and community-building initiatives.
- **Conduct regular salary reviews to ensure that compensation is competitive**. Consider performance-based incentives and bonuses to reward top performers.
- **Develop programs to incentivize long-term commitment to the organization.** This includes offering clear career progression or promotional paths, tailored professional development plans, and retention bonuses.

## Random Forest

The final approach tested was a random forest model. It is an ensemble method that combines multiple decision trees to improve classification accuracy. This type of model had the advantage of reduced overfitting and high accuracy. On the other hand, these models can be more complex and computationally intensive.

The random forest model developed had an area under curve (AUC) value of 0.97 on testing data - the best of all models. Variables of importance (>80) include (Figure 9): Income, Total Working Years, and Age.

Based on the random forest model, the following is recommended:

- **Conduct regular salary reviews to ensure that compensation is competitive**. Consider performance-based incentives and bonuses to reward top performers.
- **Develop programs to incentivize long-term commitment to the organization.** This includes offering clear career progression or promotional paths, tailored professional development plans, and retention bonuses.
- **Develop targeted employee engagement strategies for younger employees to help them feel more connected to the organization**. This could include social events, employee mentorship programs, and community-building initiatives.

## Model Selection

Based on the business problem at hand and the models tested, **it is recommended that the logistic regression model move forward for use by Canterra**. While the other models have high predictive performance compared to the logistic regression model (figure 12), the logistic regression model has the benefit of interpretability that can be translated into actionable insights for Canterra Management.

Canterra initiated this project with a consultant because management wanted *to understand what factors should focus on to curb attrition*. The logistic regression model's coefficients provide an interpretation of the impact of each variable on the likelihood of attrition. The average marginal effects (figure 3) will help identify the factors that increase or decrease the likelihood and the amount of change to the probability. This interpretability benefit makes it easier to implement targeted strategies to curb attrition.

The other models have higher predictive performance but are more complex and less interpretable. Variables of importance can be identified in these models, but the average marginal effect on the probability of attrition cannot be quantified.

# Appendix

**Figure 1: Logistic Regression Model**

| | Estimate | Standard Error | Z Value | P-Value | |
|---|---|---|---|---|---|
| Intercept | 0.35709 | 0.41532 | 0.86 | 0.389898 | |
| Business Travel Frequently | 1.71007 | 0.30928 | 5.529 | 3.22e-0.8 | *** |
| Business Travel Rarely | 0.92314 | 0.2773 | 3.329 | 0.000872 | *** |
| Marital Status Married | 0.29744 | 0.20484 | 1.452 | 0.146487 | |
| Marital Status Single | 1.05026 | 0.20753 | 5.062 | 4.18E-07 | *** |
| Number of Companies Worked | 0.1264 | 0.03099 | 4.079 | 4.52E-05 | *** |
| Total Working Years | -0.1069 | 0.01716 | -6.229 | 4.69E-10 | *** |
| Training Times Last Year | -0.105 | 0.05713 | -1.838 | 6.61E-02 | |
| Years At Company | 0.11018 | 0.02509 | 4.392 | 1.12E-05 | *** |
| Years With Current Manager | -0.16316 | 0.03431 | -4.755 | 1.98E-06 | *** |
| Environment Satisfacton Medium | -0.67387 | 0.22409 | -3,007 | 2.64E-03 | *** |
| Environment Satisfaction High | -0.69045 | 0.19977 | -3.456 | 5.48E-04 | *** |
| Enviornment Satisfaction Very High | -0.53393 | 0.20234 | -2.639 | 8.32E-03 | *** |
| Job Satisfaction Medium | -0.71543 | 0.22925 | -3.121 | 8.18E-03 | *** |
| Job Satisfaction High | -0.58908 | 0.20339 | -2.896 | 0.003776 | ** |
| Job Satisfaction Very High | -0.96023 | 0.2123 | -4.523 | 6.10E-06 | *** |

**Figure 2: Logistic Regression ROC Curve**

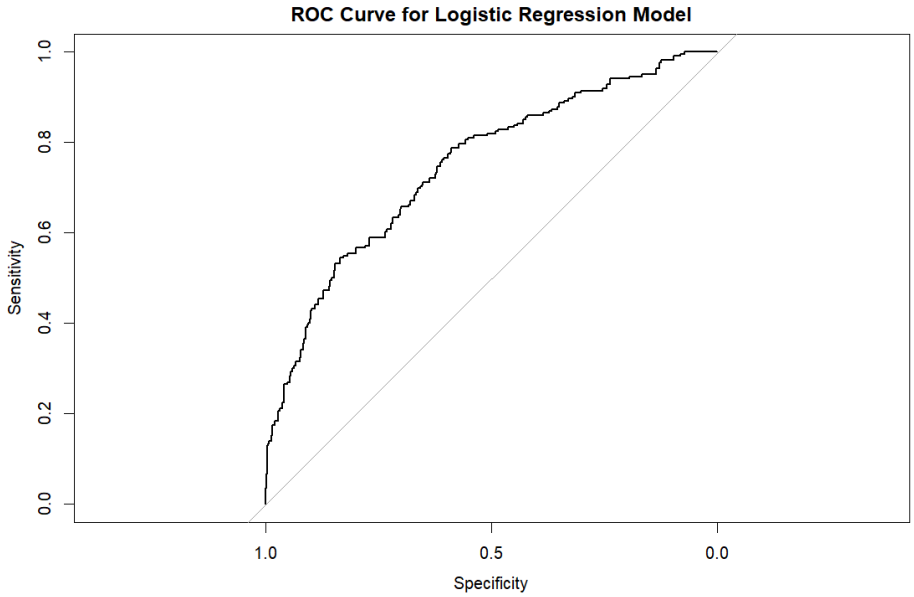



**Figure 3: Logistic Regression Average Marginal Effects**

| Variable | AME |
|---|---|
| Business Travel: Frequently | +0.28 |
| Business Travel: Rarely | +0.15 |
| Environment Satisfaction: Very High | -0.17 |
| Environment Satisfaction: High | -0.17 |
| Environment Satisfaction: Medium | -0.17 |
| Job Satisfaction: Very High | -0.16 |
| Job Satisfaction: High | -0.09 |
| Jab Satisfaction: Medium | -0.11 |
| Marital Status: Married | +0.07 |
| Marital Status: Single | +0.24 |
| Number of Companies of Worked | +0.03 |
| Total Working Years | -0.02 |

| | |
|---|---|
| Years at Company | +0.02 |
| Years With Current Manager | -0.03 |

**Figure 4: Decision Trees Model**

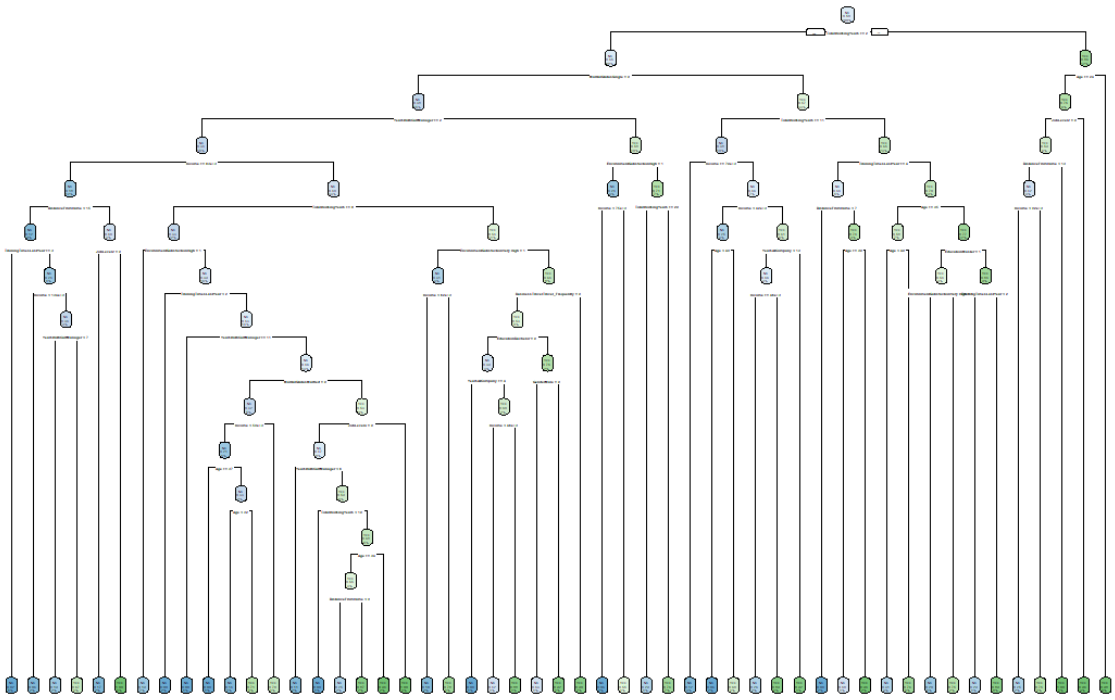


**Figure 5: Decision Trees ROC Curve**

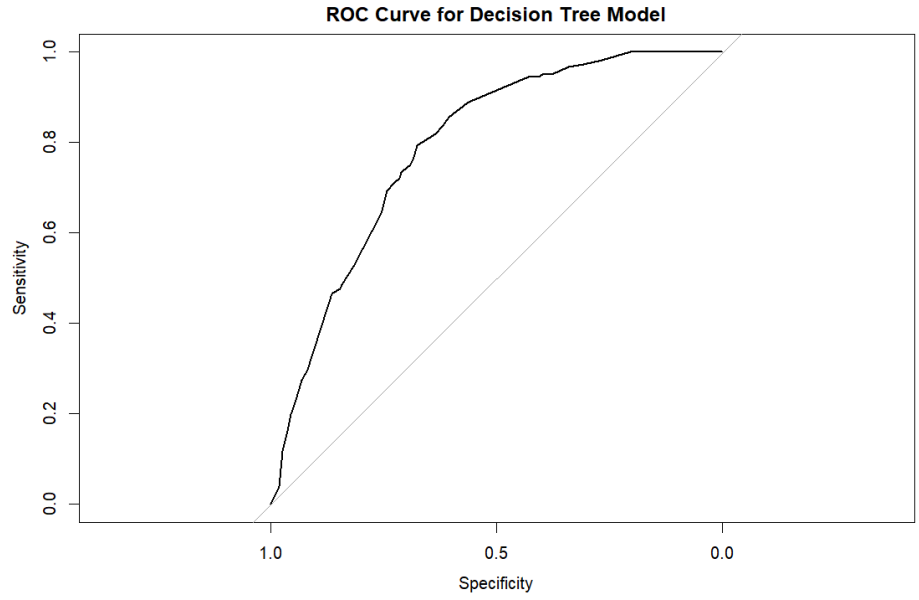
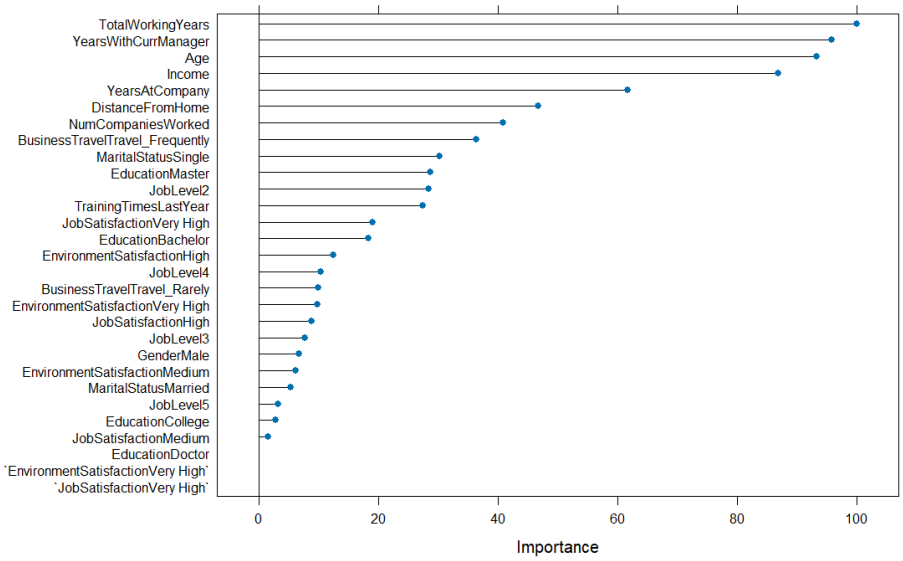


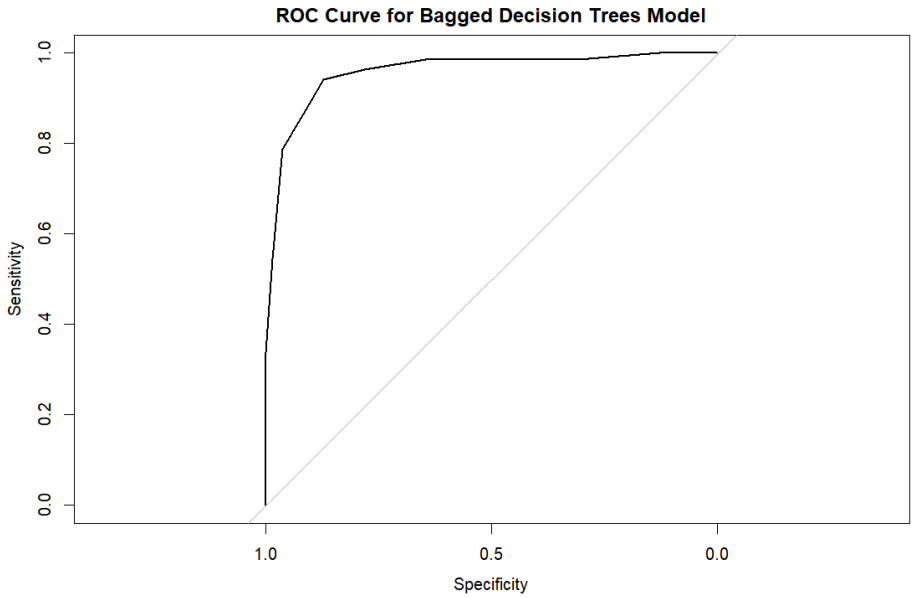

**Figure 6: Decision Trees Complexity Parameters**

| Complexity Parameter | AUC |
|---|---|
| 0.001 | 0.75 |
| 0.011 | 0.67 |
| 0.021 | 0.61 |
| 0.031 | 0.60 |
| 0.041 | 0.59 |
| 0.051 | 0.59 |
| 0.061 | 0.59 |
| 0.071 | 0.58 |
| 0.081 | 0.58 |
| 0.091 | 0.58 |

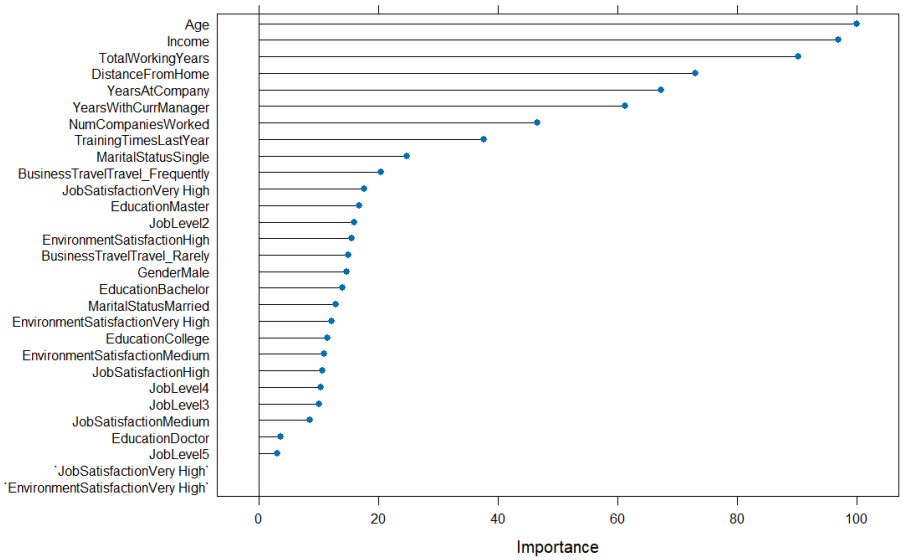
## Figure 7: Decision Trees - Variables of Importance




## Figure 8: Bagged Decision Trees ROC Curve



## Figure 9: Bagged Decision Trees - Variables of Importance




## Figure 10: Random Forest ROC Curve

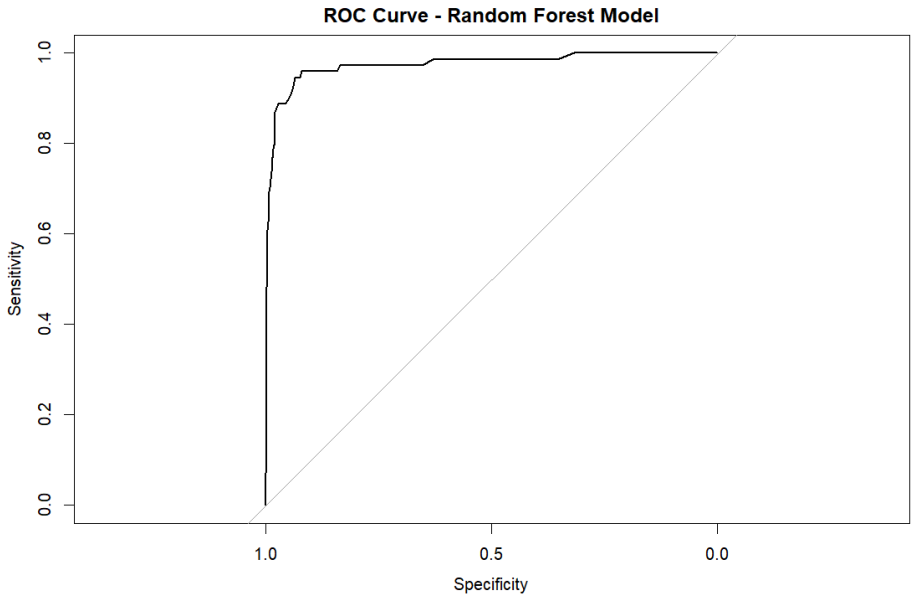


**Figure 11: Random Forest - Variables of Importance**

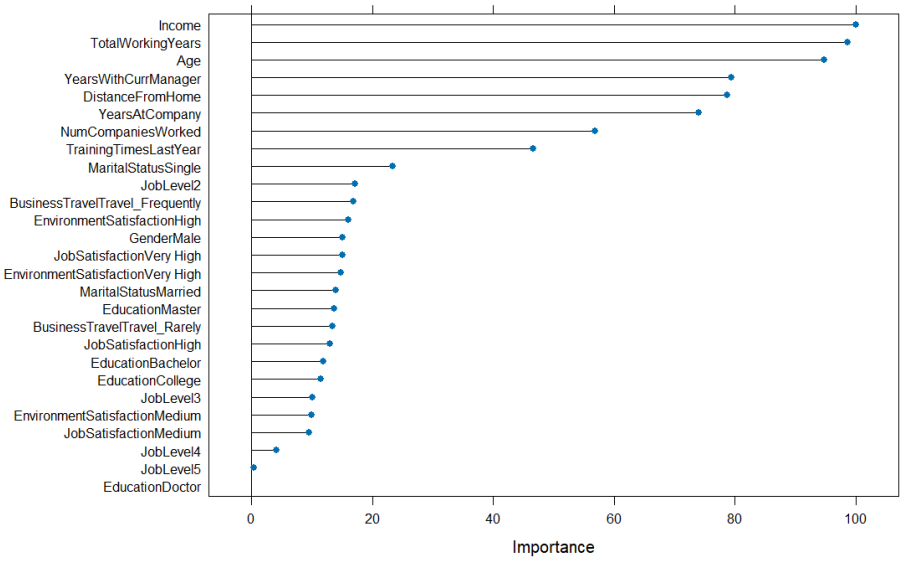



**Figure 12: Model Comparison**

| Model | AUC |
|---|---|
| Logistic Regression | 0.74 |
| Decision Trees | 0.79 |
| Bagged Decision Trees | 0.96 |
| Random Forest | 0.97 |