Case Study: Decision Tree and Ensemble-Based Tree Classifiers on Human Resource Data

Taihua Li
DePaul University

1. Introduction

Within an organization, human capital is viewed as scarce asset that are hard to acquire and expensive to lose. Recent study shows that in January 2016, the median number of years for wage and salary workers staying with one organization was 4.2 years, down from 4.6 years in January 2014. In addition, the median number of years of a worker staying with one organization was 10.1 years among those ages from 55 to 64, compared to 2.8 years among those ages from 25 to 34 years (Bureau of Labor Satistics, 2016). Furthermore, the average costs to replace an employee are around 16% of annual salary for high-turnover, low-paying jobs, 20% of annual salary for mid-range positions, and up to 213% of annual salary for highly educated executive positions (Glynn & Jane, 2012). This costly phenomenon has raised many concerns across organizations, specifically in the human resource domain, and many efforts have been made to understand *what are the factors contributing to employees leaving the organization*.

In addition to understand contributing factors, organizations would like to predict if an employee is likely to leave to avert avoidable cost. Statistics and machine learning algorithms are usually applied to build classifiers to achieve the prediction, such as Naïve Bayes, Logistic Regression and Decision Tree. In this work, I will explore the effectiveness of several machine learning algorithms on a public human resource dataset.

For the rest of this paper, I will first discuss the dataset, including the attribute information and the data cleaning and transformation process, followed by a preliminary analysis of the dataset. Then, a comparative analysis of machine learning algorithms will be presented. Finally, the paper will be concluded with a discussion, followed by a conclusion.

2. Data

The human resource dataset is provided by Ludovic Benistant, and it is available to the public on *kaggle.com*. In total, there are 14,999 observations and 9 predictors and 1 predicted variable:

a) Predictors
   - *satisfaction_level*: indicates the level of satisfaction of an employee, ranging from 0 to 1
   - *last_evaluation*: the last performance evaluation on the employee, ranging from 0 to 1
   - *number_project*: number of project the employee completed while at work
   - *average_montly_hours*: average monthly hours at workplace
   - *time_spend_company*: number of years spent in the company
   - *Work_accident*: whether an employee had a work accident (0: No 1: Yes)
   - *promotion_last_5years*: whether an employee was promoted in the last five years (0: No 1: Yes)
   - *sales*: department in which the employee worked for, including accounting, HR, IT, management, marketing, product_mng, RandD, sales, support and technical
   - *salary*: relative level of salary (low, med and high)

b) Predicted variable
   - *left*: whether the employee left the company (0: No 1: Yes)

As indicated above, we have a mix of numerical, categorical, binary and ordinal variables; *satisfaction_level*, *last_evaluation*, *number_project*, *average_monthly_hours* and *time_spend_company* are numerical variables, *work_accident* and *promotion_last_5years* are binary variables, *sales* is a categorical variable and *salary* is an ordinal variable. However, predictors, *number_project*, *average_monthly_hours* and *time_spend_company* can be treated as categorical variable as well because each of them only have a certain number of unique values.

2.1 Data Cleaning

In this dataset, a minimum amount of data pre-processing was performed due to its completeness and simplicity. First, in R, the code, *sum(is.na(data))*, was executed to inspect if there was any missing value, and since the returned value was 0, indicating there was no missing value, we proceeded to examine if there was any abnormal value in each attribute. Table 1 shows the summary statistics of all numerical predictors in our dataset, and it indicates that no numerical predictor is suspected to contain any abnormal value, per the distribution of each predictor.

Predictors	*satisfaction_level*	*last_evaluation*	*number_project*	*average_monthly_hours*	*time_spend_company*
Minimum	0.0900	0.3600	2.000	96.0	2.000
1st Quartile	0.4400	0.5600	3.000	156.0	3.000
Median	0.6400	0.7200	4.000	200.0	3.000
Mean	0.6128	0.7161	3.803	201.1	3.498
3rd Quartile	0.8200	0.8700	5.000	245.0	4.000
Maximum	1.0000	1.0000	7.000	310.0	10.000

Table 1: Summary statistics of numerical predictors

The correlation plot in Figure 1 visualizes the correlation among numerical variables in the dataset, where the color indicates positive (blue) or negative (red) correlation, the shades of color indicates the level of correlation (the darker the shade is, the stronger the correlation is), and the sharpness of the ellipse also indicates the level of correlation (the sharper/thinner the ellipses are, the stronger the correlation is). As shown on the right, there is no pair of variables that have strong correlation. It indicates that there is no redundant variable nor multicollinearity in the dataset.

Next, we will inspect the binary and ordinal predictors and predicted variable. Figure 2 shows the bar graphs of all categorical, binary and ordinal variables in the dataset, and no bar graph indicate any abnormal value in any of these variables. Therefore, no data cleaning is necessary for this dataset. To fit the machine learning algorithms on this dataset, we will need to transform all categorical variables, except the predicted variable, *left*, into dummy or indicator variables. For example, *Work_accident* will be transformed into two variables since it has two categories, and the variables are *Work_accident0* and *Work_accident1*.

2.2 Preliminary Analysis

In this section, more detailed analyses of predictors are performed.

2.2.1 *Salary*

Intuitively, *salary* sometimes plays an important part in whether an employee would leave a company. First, a distribution of low, medium and high salary employees across department is shown in Figure 2. We can see that most of the employees in the dataset are in the sales, technical and support departments, while employees with low and medium salary levels make up most the dataset. IT department has the most employees in the low and medium salary levels, excluding sales, technical and support departments, and in these two salary categories, management department has the lowest number of employees. However, in the high salary category, management department has the highest number of employees, excluding the top three most populated departments.

2.2.2 *promotion_last_5years*

From Figure 2, it shows that most employees did not get a promotion in the last five years. After calculating the percentage of employees getting either promoted or not promoted, as shown on the right, we can see that almost 6% of employees who have high salary were promoted in the last five years. It also shows a pattern that the higher the salary level is, the higher the percentage of promoted employees is.

Salary Level	Not Promoted	Promoted
Low	0.9909	0.0091
Medium	0.9719	0.0281
High	0.9418	0.0582

2.2.3 *work_accident*

From the bar graph visualizing counts of *work_accident* levels in Figure 2, it shows that most employees did not have a work accident but more than 2,000 employees did have a work accident. Intuitively, we correlate work accident with whether an employee leaves the company, and it can be examined through a cross tabulation, as shown on the right, where each entry represents the percentage of employees whether leaves the company based on

whether they had a work accident. It shows that among those who had work accident, only 7.79% left the company and the rest stayed with the company. This indicate that *work_accident* might not be an informal predictor to classify if an employee would likely to leave the company.

	Work Accident	
Left Company	Yes	No
Yes	0.0779	0.2652
No	0.9220	0.7348

2.2.4 *time_spend_company*

The number of years an employee stays with a company usually is a strong indicator of whether the employee would leave the company. Generally, especially in the labor markets like that in Silicon Valley, the longer an employee stays with a company, it means that the employee has more professional experience in the field and he or she is more likely to go to another company due to higher salary offer, etc. First, let's look at the distribution of low, medium and high salary levels among different years spent with the company,

	Number of Years Spent with Company							
	2	3	4	5	6	7	8	10
Low	0.2087	0.4381	0.1777	0.1092	0.0455	0.0049	0.0082	0.0077
Medium	0.2194	0.4217	0.1682	0.0943	0.0512	0.0177	0.0130	0.0146
High	0.2449	0.4204	0.1399	0.0534	0.0445	0.0307	0.0145	0.0518

As shown above, most the employees stayed with the company for less than 5 years; more than 40% of employees in each of the low, medium and high salary levels spent three years with the company, around 20% from each salary level spent two years, and around 15% from each salary level spent four years. As indicated in this table, there is not much variation in the distribution of number of years spent with the company among different salary levels.

	Number of Years Spent with Company							
Left	2	3	4	5	6	7	8	10
Yes	0.0148	0.4441	0.2492	0.2332	0.0585	0	0	0
No	0.2792	0.4250	0.1459	0.0560	0.0445	0.0165	0.0142	0.0187

The table above shows the relationship between number of years an employee spent with the company and whether the employee left. It shows that among the employees who left the company, most of them spent from three to five years with the company. This potentially indicates that majority of employees who left the company might age between 25 and 34 years, but the connection cannot be established until demographic data about the employees are provided.

2.2.5 *average_montly_hours*

average_montly_hours indicates the average number of hours an employee spent at the company. Intuitively, the higher value this variable is, the more likely an employee would leave the company because of the exhausting work. As shown in Figure 4, there is a bimodal distribution for the average number of hours an employee spent with the company, where one maximum point is roughly at 150 hours and another one is at 255 hours. It indicates that most of employees work between 37.5 hours and 63.75 hours per week, on average. Since most of employees in this dataset are from the sales department, this number would make sense because of sales representatives do generally work more than other departments' employees. A more detailed relationship among average monthly hours, whether left the company and department is indicated in Figure 5. From the graph, it shows that employees among all departments show almost identify densities in terms of monthly average number of hours spent at company, across groups that left or stayed with the company; all densities showed bimodal distributions that are like the population density distribution. Since there is a lack of variation, this predictor might not be effective at predicting if an employee is likely to leave.

2.2.6 *number_project*

This number indicates the total number of project completed by the employee while at work. The following table is constructed to see the distribution of number of project completed across groups of employees left or stayed with the company,

	Number of Project Completed					
Left	2	3	4	5	6	7
Yes	0.4388	0.0202	0.1145	0.1714	0.1834	0.0717
No	0.0718	0.3485	0.3461	0.1880	0.0454	0

As shown above, among those who left the company, around 43.88% of them completed two projects while at work. For those stayed with the company, in general, they completed more projects than those who left.

2.2.7 *last_evaluation*

This predictor can be used to validate the hypothesis that employees left the company for higher paying positions. As stated in 2.2.4, most of those who left the company spent three to five years with the company and as shown in Figure 6, among those who left the company and spent between four to six years with the company, the majority had a last evaluation score of more than 0.8. It indicates that majority of employees who left the company were high performing labors and their leave costs the company a lot more than what it takes to re-hire to fill the positions, but also their experiences and skills with the company. This is also the motivation of predicting whether an employee will leave the company to prevent potential human resource cost.

2.2.8 *satisfaction_level*

Despite of how well the benefit company gives to its employees, if the employee does not like the company or the position, it is highly likely the employee will leave. This variable captures the overall satisfaction level of the employee to the company. From Figure 7, it shows that among the employees who stayed with the company, in general, they were satisfied with the company. However, among the employees who left the company, most of the satisfaction rated were below 0.5 and the patterns are almost identical across all departments. It indicates that employees who left the company were generally not satisfied with the company.

2.3 Data Transformation

To fit the machine learning models onto this dataset, data transformation was performed. Categorical variables, *Work_accident*, *promotion_last_5years*, *salary* and *sales* were converted into dummy variable indicating the positive and negative cases of each level of the corresponding variable. After dummy variables were generated, there were a total of 21 variables in the dataset, including the predicted variable.

3. Classification Algorithms

In this section, a comparative analysis is conducted to study the effectiveness of ensemble-based approaches in improving predictive performance of machine learning algorithms. In total, four models were built and evaluated, including Decision Tree, Logistic Regression, Random Forest and Boosted Logistic Regression.

All models were built on the training dataset and evaluated the performances on the testing dataset, where training dataset was bootstrapped with replacement and took up 80% of the original dataset. As shown in Figure 2, there is a class imbalance phenomena in the predicted variable, *left*. To build an effective model that discriminates well against the negative class, synthetic up-sampling on the positive class was performed using the SMOTE package in R. After up-sampling the positive class and down-sampling the negative class proportionally, there were 8,545 observations in the training set, where the positive and the negative classes were equally distributed, and 3,014 observations in the testing set. To train the model with best parameters possible, 5-fold cross validation was performed during each model's training and accuracy was used as the metric to select the final model.

In this experiment, Random Forest is the ensemble-based (bagging) model to be compared with Decision Tree and Boosted Logistic Regression is the ensemble-based (boosting) model to be compared with Logistic Regression.

3.1 Decision Tree

The configuration parameters to find the best decision tree included: minimum split equaled to 20, minimum bucket number equaled to 7 and the maximum tree depth equaled to 30 levels.

In the final decision tree model, the five most important variables include *statisfaction_level*, *time_spend_company*, *number_project*, *last_evaluation*, *average_monthly_hours*. This potentially validated our assumptions, stated previously, of the relationship between whether an employee is likely to leave the company and whether the company is satisfied with the company itself. As shown in Figure 8, the most important variable is *satisfaction_level*.

3.2 Logistic Regression

The final logistic regression model used to evaluate on the test dataset was,

$$\begin{aligned} left = {} & 2.165 + 4.55satisfaction\_level - 1.19last\_evaluation + 0.4number\_project \\ & - 0.43time\_spend\_company - 1.55Work\_accident\_0 + 1.75promotion\_last\_5years\_1 \\ & - 1.89salary\_Low - 1.45salary\_Med - 0.03hr + 0.44IT + 0.76management \\ & + 0.08marketing + 0.3product\_management + 0.69RandD + 0.26Sales + 0.03support \\ & + 0.13technical \end{aligned}$$

As shown above, satisfaction level, salary level, whether the employee was promoted in the last five years and whether the employee had a work accident were the top contributing factors to whether the employee is likely to leave the company.

3.3 Random Forest

After performing 5-fold cross validation during the training process and keep tuning and evaluation on different number of trees parameter for Random Forest, a final model of 500 trees were selected to build the final model because it outputted the highest accuracy and lowest overall error. As shown in Figure 9, black line represented the overall model classification error on the training data, and it showed that when the number of trees approached 500, the error rate was at the minimum point, which was at 1.49%. In the Random Forest model, the five most important variables included *statisfaction_level*, *time_spend_company*, *number_project*, *last_evaluation*, and *average_monthly_hourr*, which are the same output as those from the Decision Tree model.

3.4 Boosted Logistic Regression

Boosted Logistic Regression is a Logistic Regression that is trained based on the AdaBoost model. In this model, during each iteration of the model building, the weights assigned to the observations that are falsely classified increases and for those that correctly classified, the weights would decrease. The final model was built on 8,545 observations after 31 iterations of training.

4. Discussion

The following table summarizes the performance statistics of all four models described above,

	Models			
Metrics	Decision Tree	Logistic Regression	Random Forest	Boosted Logistic
Accuracy	0.8086	0.7565	0.9887	0.9237
95% Conf. Interval (lower)	0.7941	0.7407	0.9843	0.9136
95% Conf. Interval (uppper)	0.8225	0.7717	0.9922	0.9329
Sensitivity	0.9695	0.7701	0.9806	0.9474
Specificity	0.7579	0.7522	0.9913	0.9162
Precision	0.5578	0.4947	0.9806	0.7808
Recall	0.9695	0.7700	0.9725	0.9474

As shown above, Decision Tree and Logistic Regression did not perform well in predicting if an employee is likely to leave the company overall, where their accuracies are 0.8086 and 0.7565, respectively. However, Decision Tree is better at predicting positive class than predicting negative class since its sensitivity is higher than its specificity, while Logistic Regression performs almost equally well at predicting both positive and negative classes.

Ensemble-based models have proven to improve the predictive power of machine learning algorithms, and it holds true in this experiment study. As shown above, Random Forest has an overall accuracy of 0.9887 while its sensitivity is 0.9806 and specificity is 0.9913, which has a 0.18 improvement on overall accuracy and 0.24 improvement on the predictive power on the negative class. Boosted Logistic Regression has an overall accuracy of 0.9237, a sensitivity of 0.9474 and a specificity of 0.9162, which is a 0.17 improvement in the overall accuracy and 0.16 improvement on both sensitivity and specificity. In this comparison, both ensemble-based classifiers performed much better than the base classifiers.

Predicting if an employee is likely to leave the company is an ideal solution for companies to save avoidable cost, and in this case, precision and recall are important measurements. In this context, precision measures how accurate the predictive model is and recall measures the accuracy of predicting an employee to be likely to

leave among all who are likely to leave. Therefore, an ideal model shall have both high recall and precision. As shown in the table above, Decision Tree has a high recall of 0.9695 and a low precision of 0.5578, which means it captures those who are likely to leave the company well but there is too much noise associated with overall predicted output. However, for Random Forest, the precision increases to 0.9806 and the recall increases to 0.9725. This shows that Random Forest can minimize the output noise while keeping the predictive performance of the Decision Tree model. For Logistic Regression model, the precision is 0.4947 and the recall is 0.77. This shows that the model is not suitable to predict the positive class while there is still room for improvement in predicting the negative class. Boosted Logistic Regression improved the performance of Logistic Regression, and its precision is 0.7808 and the recall is 0.9474. It indicates that Logistic Regression and Boosted Logistic Regression both have worse performance in predicting the positive class and the negative class in the human resource dataset used in this study.

Among the four models evaluated, a ROC curve plot is presented in Figure 10. It shows that overall, Random Forest is the best model among all because it has the highest true positive rate when the false positive rate is low, and it has the highest true positive rate across all levels of false positive rate. The accuracy, sensitivity, specificity, precision and recall also indicate that Random Forest is the best model among four to predict if an employee is likely to leave the company.

5. Conclusion

In this paper, we study the performance of four machine learning algorithms, including Decision Tree, Logistic Regression, Random Forest and Boosted Logistic Regression. Due to the imbalanced predicted variable class, a synthetic up-sampling of positive class and down-sampling of negative class was performed on the training dataset. Each model was trained using 5-fold cross validation and the best model parameters were used to build the model, which was then evaluated on the testing dataset. Per evaluation metrics, including accuracy, sensitivity, specificity, precision and recall and the visualization of true positive and false positive rates on ROC curves, Random Forest had the best performance, followed by Boosted Logistic Regression, Decision Tree and lastly, Logistic Regression.

In this dataset, there were several limitations restricting from a better predictive model being built. First, many variables have almost identical distributions as discussed in Section 2, which indicates that there is not much valuable information contributing to the associative relationships with the predicted variable. Second, the class imbalanced problem in the predicted variable led to synthetic up- and down-sampling of classes, and it might introduce noise into the dataset, which could possibly skew the model performance. In addition, the size of the dataset that contributes to predicting if an employee is likely to leave is small; among 14,999 observations, only 3,571 observations are in the positive class. To improve the performance and applicability of the predictive models, more data about the positive class, which is when the employee left the company, should be collected, and more information on the employee information could be collected as well, such as demographic data.

Overall, in this study, ensemble-based classifiers, including both bagging and boosting approaches, performed better at predicting whether an employee is likely to leave the company than classical machine learning algorithms.

Appendix: Visualization Figures

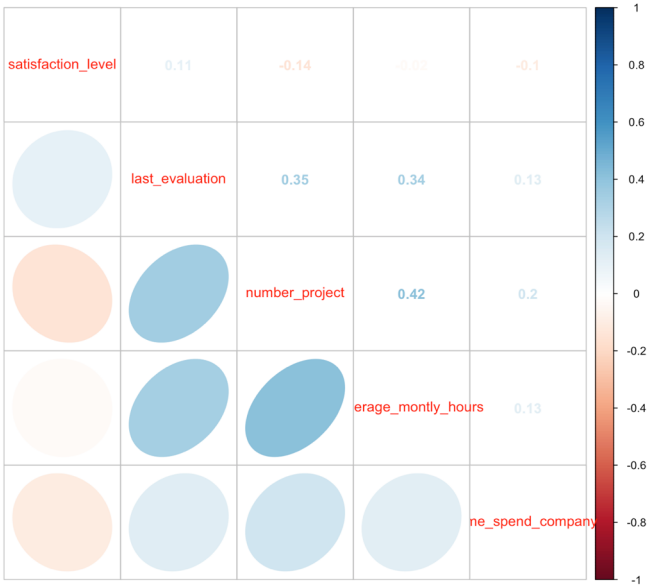

Figure 1: Correlation Plot of Numerical Predictor

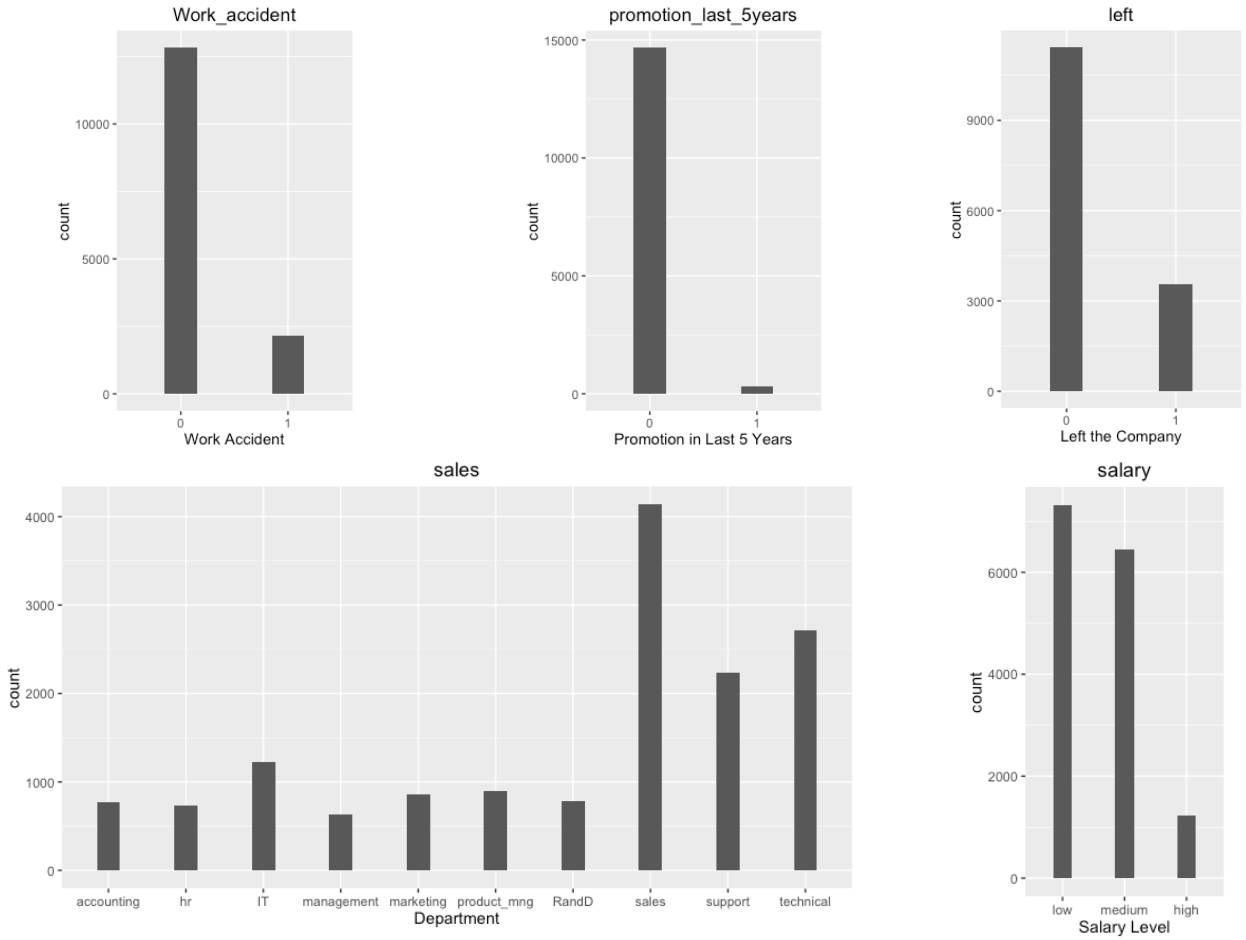

Figure 2: Bar Graphs of Categorical, Binary and Ordinal Predictors and Predicted Variable

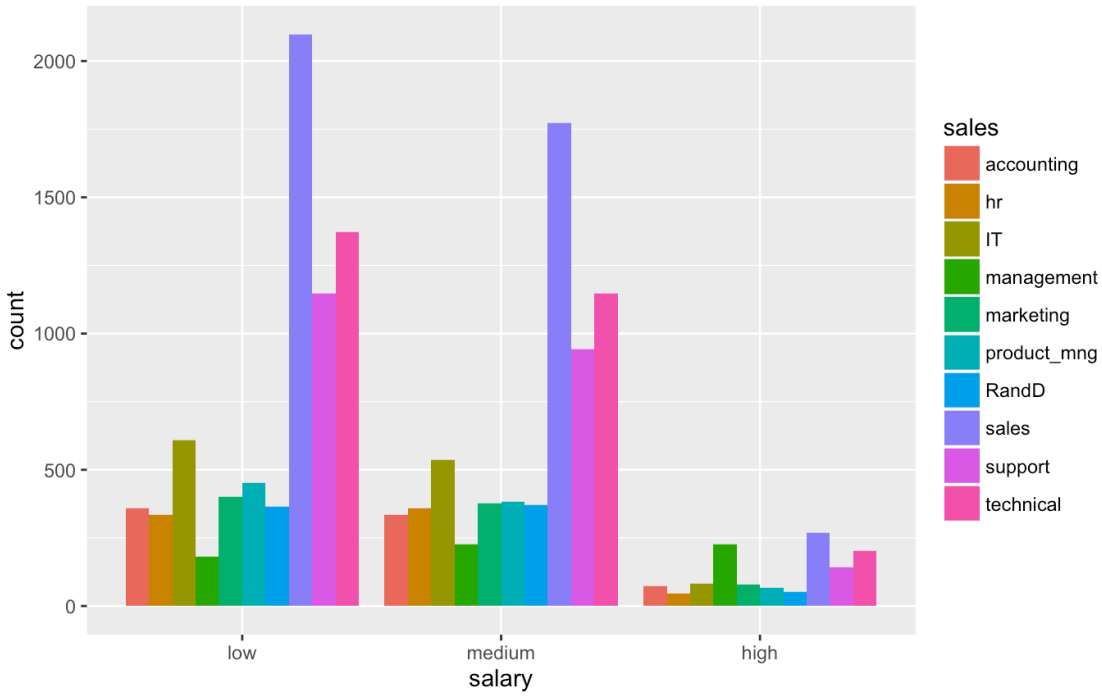

Figure 3: Distribution of Salary Level Across Department

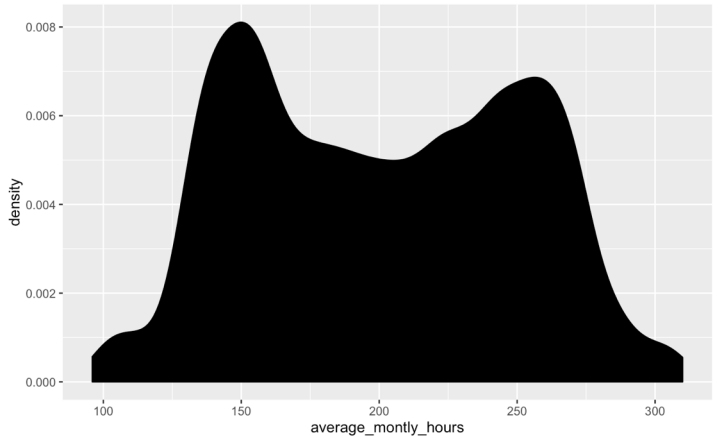

Figure 4: Density Plot of Average Monthly Hours Spent with the Company

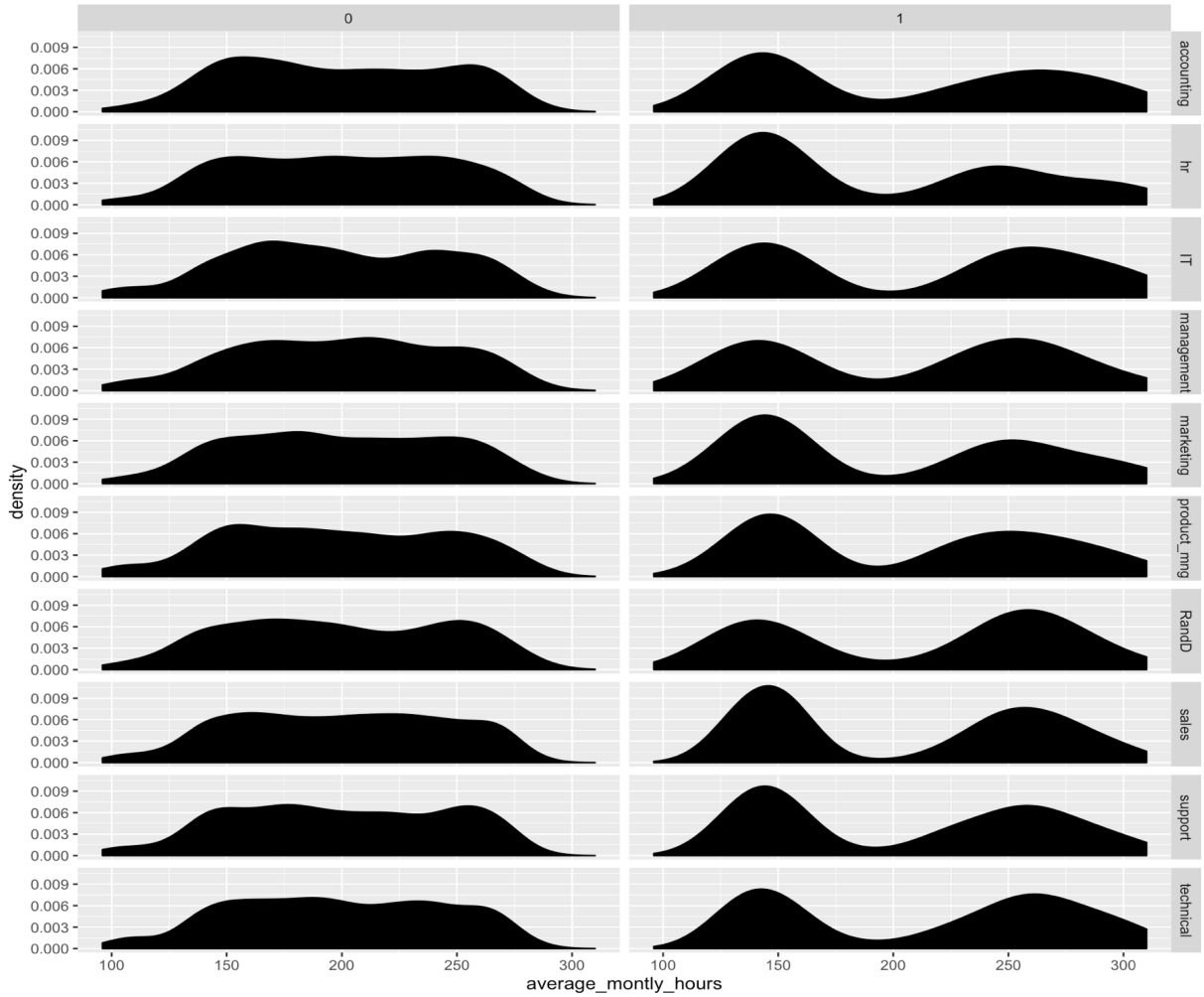

Figure 5: Density Plot of Average Monthly Hours Spent with the Company, mapped by department and whether the employee left the company (0: did not leave 1: left the company)

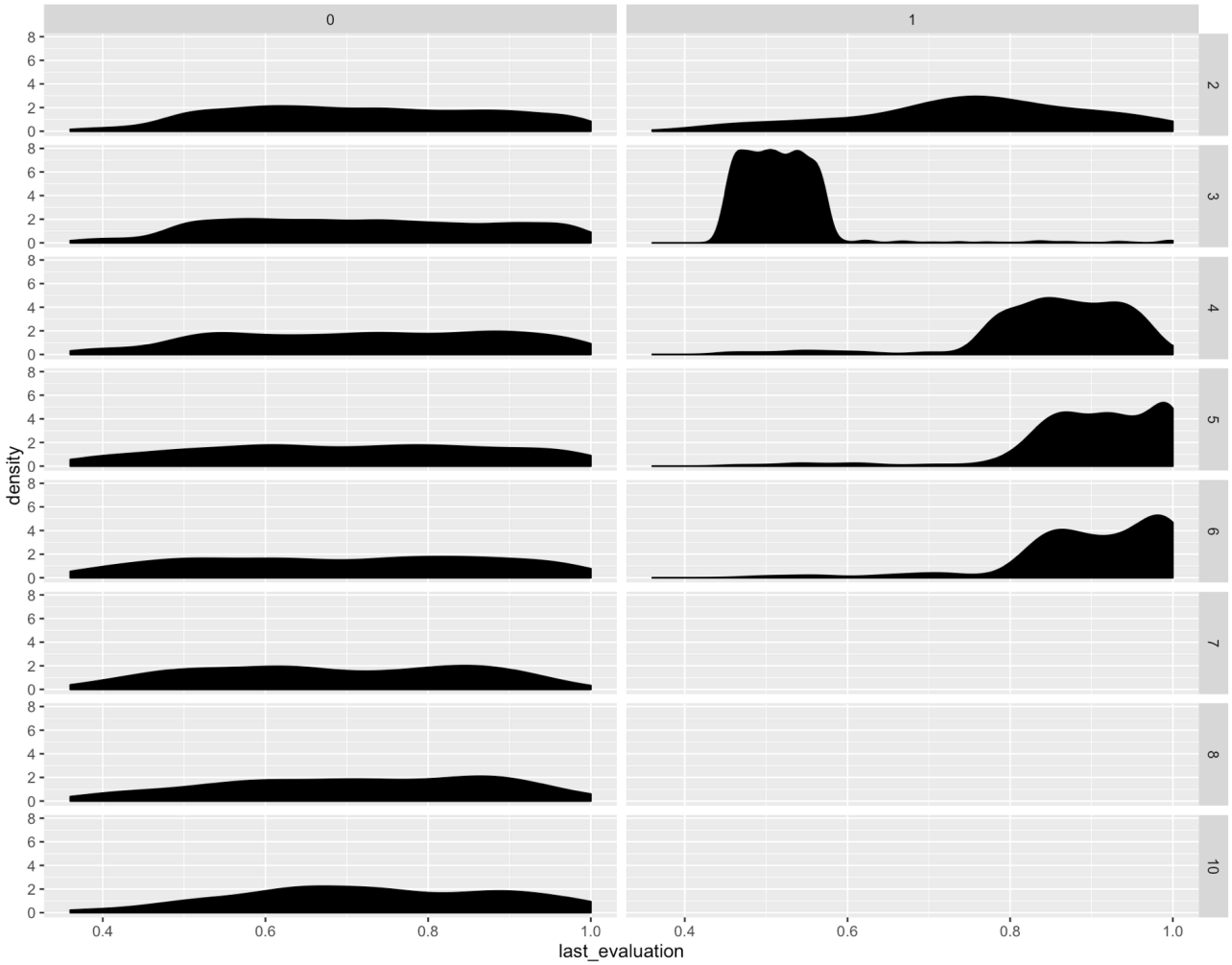

Figure 6: Density Plot of last employee evaluation, mapped by number of years spent with the company and whether the employee left the company (0: did not leave 1: left the company)

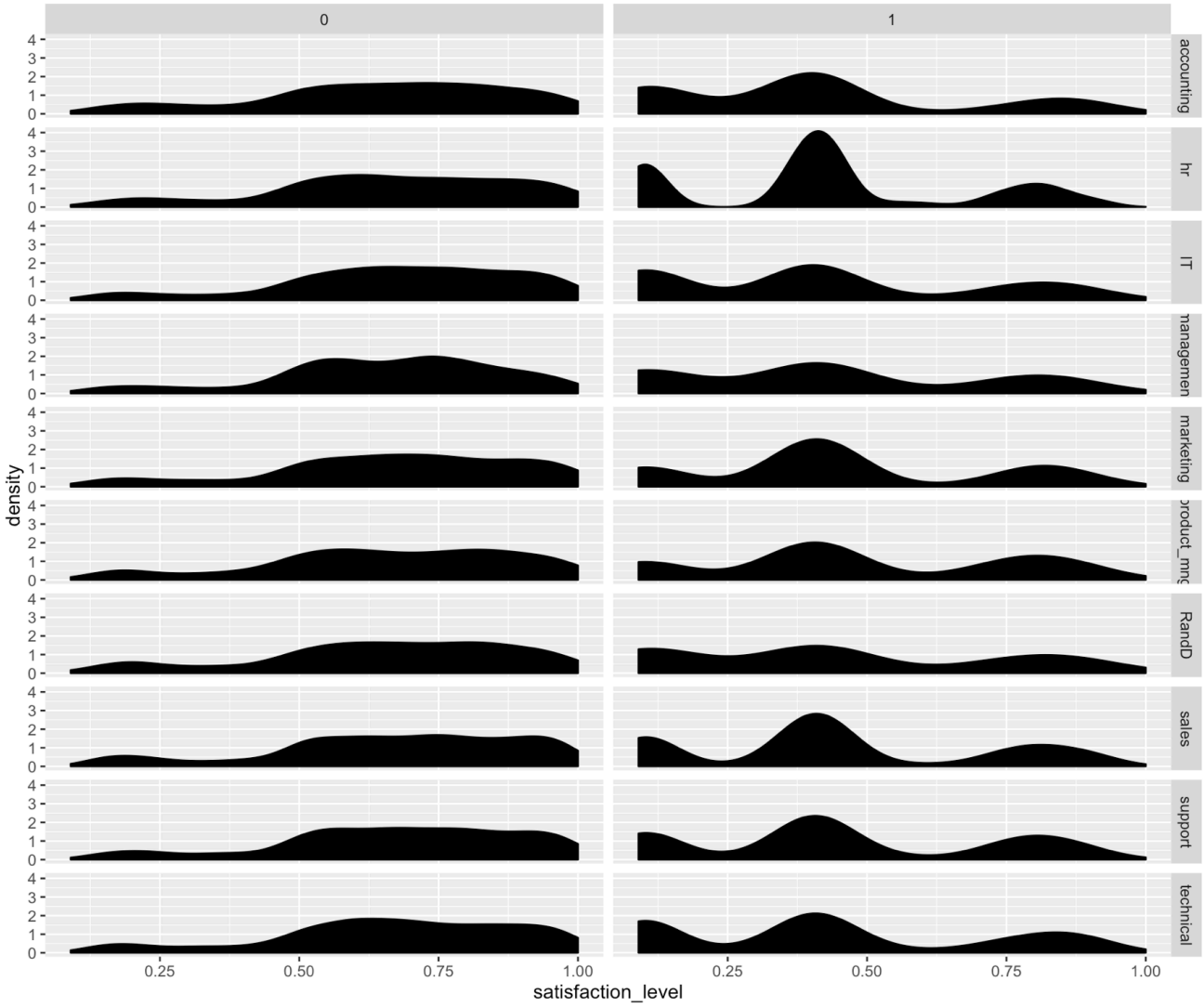

Figure 7: Density Plot of employee satisfaction level, mapped by department and whether the employee left the company (0: did not leave 1: left the company)

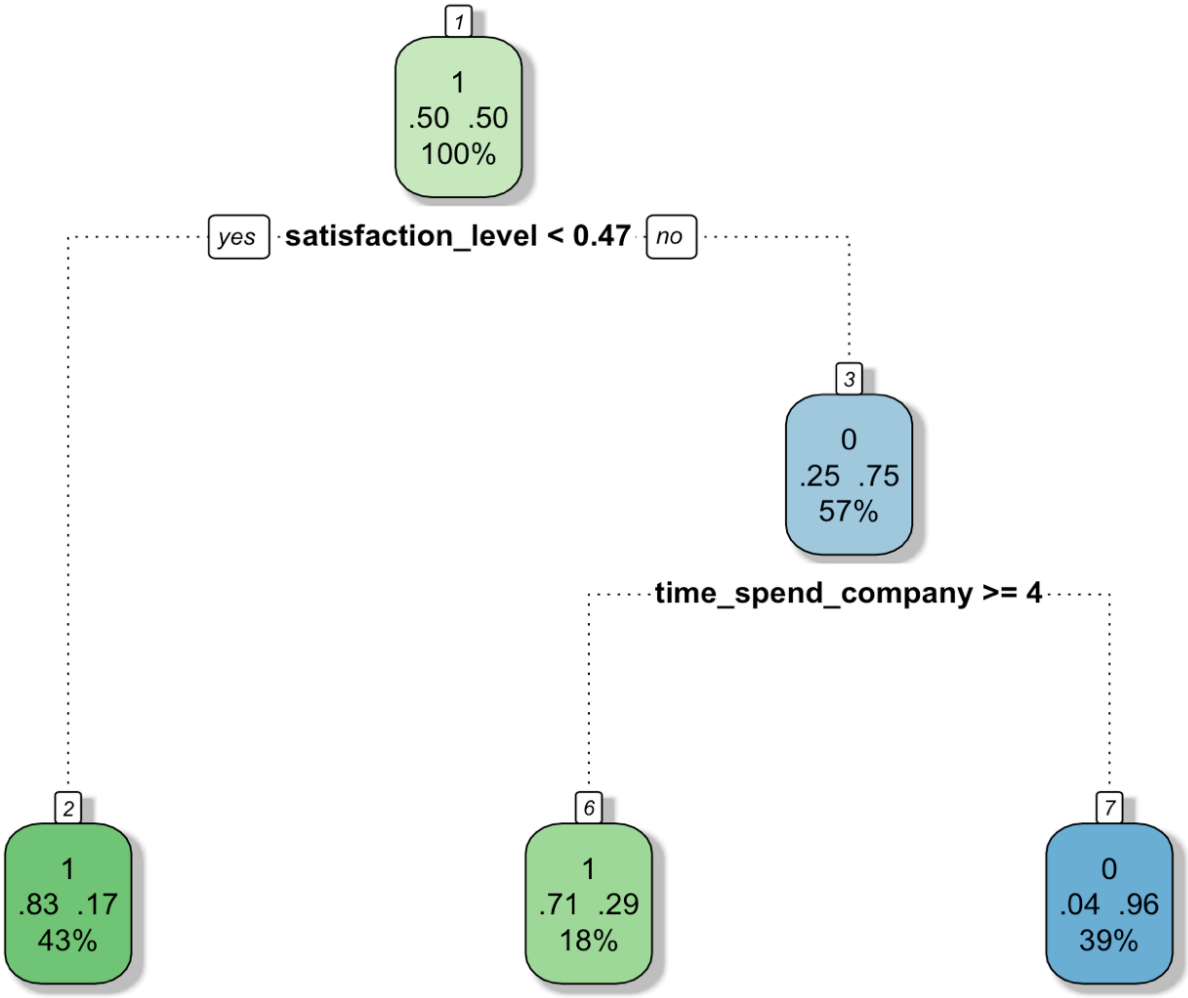

Figure 8: Final Decision Tree Model Plot

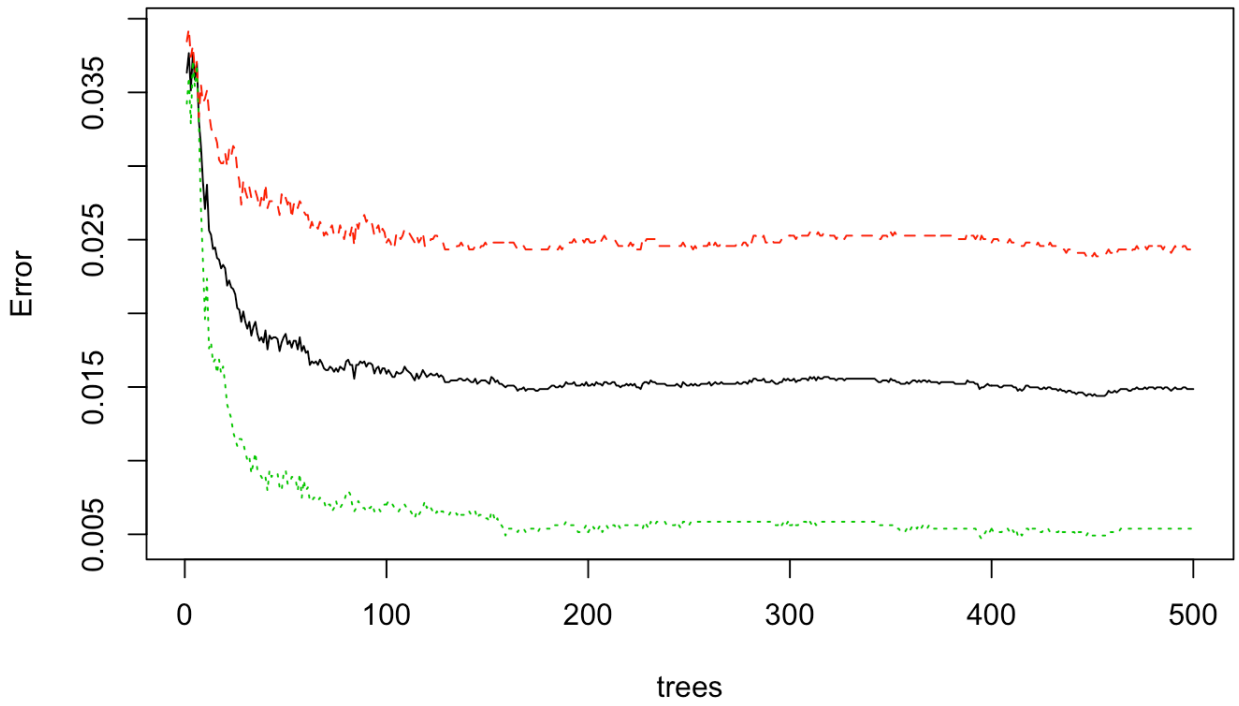

Figure 9: Error rate of different Random Forest models with stochastic number of trees (Black: overall classification error, Red: positive class error, Green: negative class error)

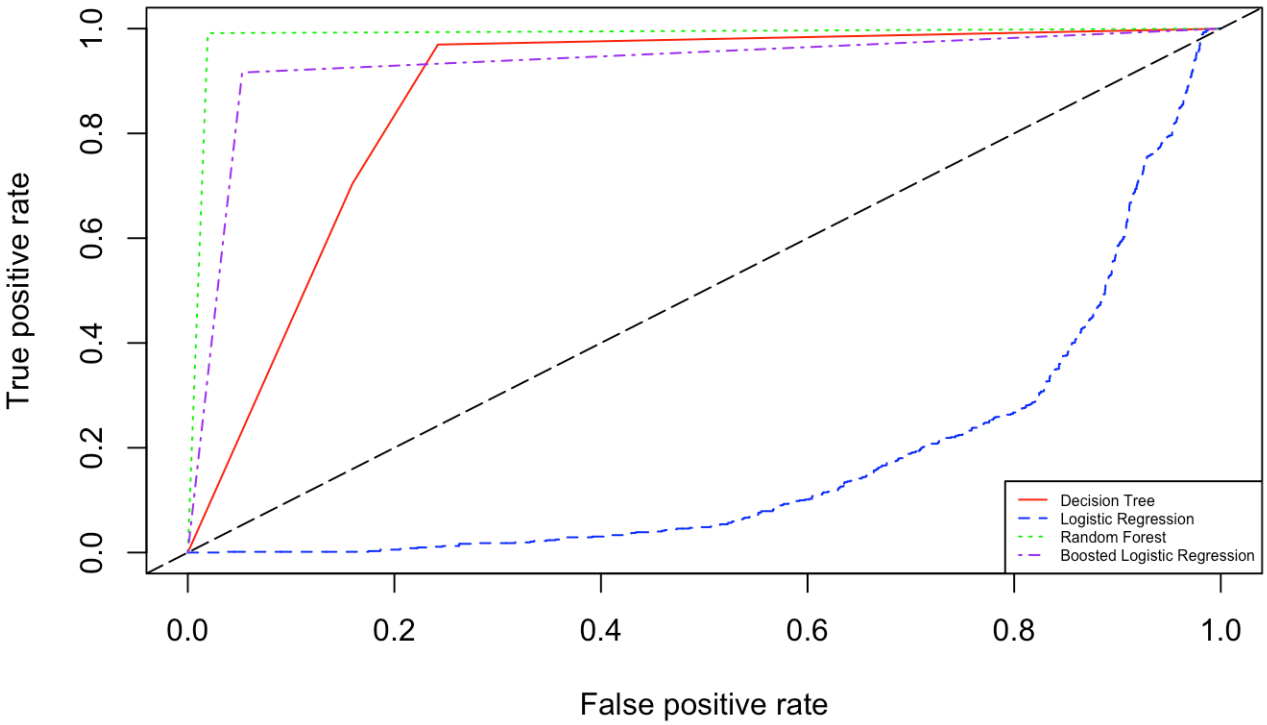

Figure 10: ROC Curves of All Models Trained

Bibliography

Bureau of Labor Satistics. (2016, September 22). *Employee Tenure Summary.* Retrieved from Bureau of Labor Satistics: https://www.bls.gov/news.release/tenure.nr0.htm

Glynn, H. B., & Jane, S. (2012, November 16). *Center for American Progress.* Retrieved from There Are Significant Business Costs to Replacing Employees: https://www.americanprogress.org/wp-content/uploads/2012/11/CostofTurnover.pdf