This project analyzes a dataset for Human Resource Analytics to determine what causes employees to leave prematurely and possibly predict which valuable employees may leave.

**Human Resource Analytics: Why Are Employees Leaving Prematurely and Who’s Next?**

OMIS 665

Carlo Aseron | Z1667059

Swati Botuwar

Swathi Prakasha

Ashwini Ravindra

**Abstract**

**Project Goal**

The goal of this project is to analyze a dataset using data modeling techniques in order to find significant insight and develop actionable recommendations. The dataset used in this model is a simulated Human Resource Analytics dataset that focuses on various employee factors and indicates whether the employees left the company prematurely and which valuable employee might leave company next. The focus of this project is to identify those key factoring impacting the decision to leave the company through statistical modeling and analysis. SAS Enterprise Miner and R studio will be used to make statistical models.

**Dataset Description**

The dataset for this project, "Human Resource Analytics", was obtained on kaggle.com and is a simulated dataset that tracks various employee attributes/experiences (listed in detail below) to provide insight as to why some employees leave the company prematurely. The dataset contains 14999 records and 10 columns/attributes. Each row corresponds to an employee.

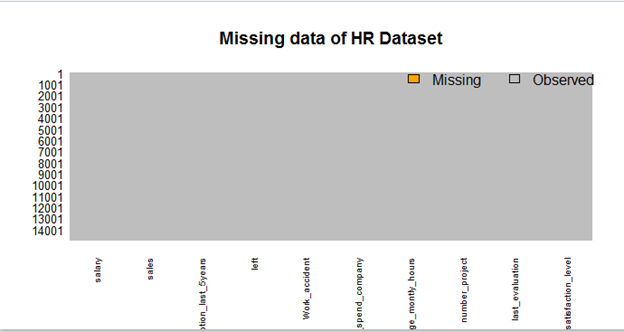
The dataset measures the following employee attributes:

* **satisfaction\_level**
  + *Data Type*: numeric
  + *Description*: the level of satisfaction measured as a value between 0 and 1.
* **last\_evaluation**
  + *Data Type*: numeric
  + *Description*: the last evaluation score of the employee
* **number\_projects**
  + *Data Type:* numeric
  + *Description:* the number of projects completed while at work
* **average\_montly\_hours**
  + *Data Type:* numeric
  + *Description:* the average monthly hours at the workplace
* **time\_spend\_company**
  + *Data Type:* numeric
  + *Description:* the number of years spent in the company
* **Work\_accident**
  + *Data Type:* numeric
  + *Description:* whether an employee had a workplace accident
* **Promotion\_last\_5years**
  + *Data Type:* numeric
  + *Description:* whether an employee was promoted in the last 5 years
* **sales**
  + *Data Type:* string
  + *Description:* the department in which the employee works (sales, accounting, hr, technical, support, management, IT, product-mng)
* **salary**
  + *Data Type:* string
  + *Description:* relative level of salary (low, medium, or high)
* **left**
  + *Data Type:* numeric
  + *Description:* whether the employee has left (1=yes, 0=no)

**Data Exploration**

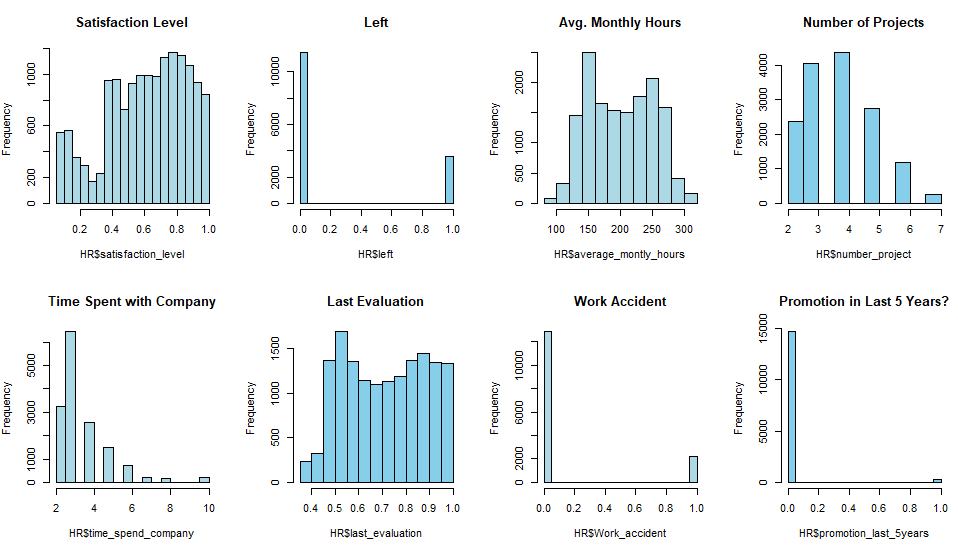
**Check for Missing Values:**

Before performing analysis on data, we did a check for missing values. Based on the plot below, we can see that our dataset has no missing values.

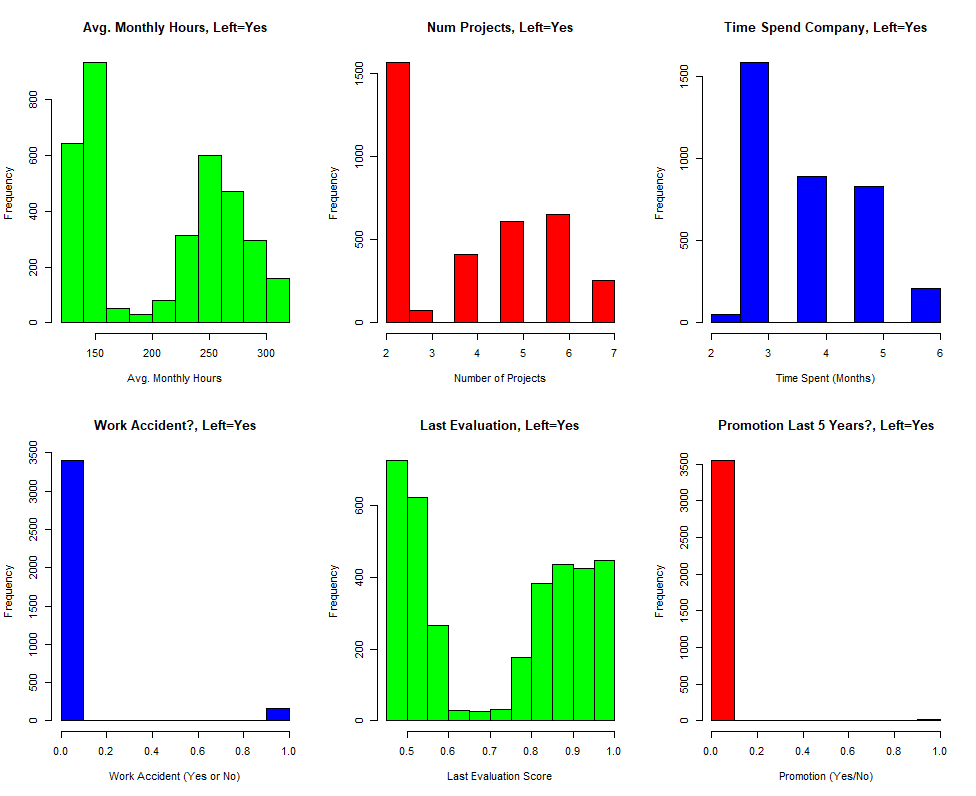


**Histograms:**

To further explore our data, we analyzed histograms to study the distribution of values for each of the numeric variables (see chart below). An important thing to note is the number of work accidents and promotions in the last 5 years. There are a relatively low number of promotions compared to the total number of employees, so the promotion variable may have little to no impact on our analysis. Work accidents are also scarce, though there is still a fair amount that may account for a good portion of employees who left prematurely.

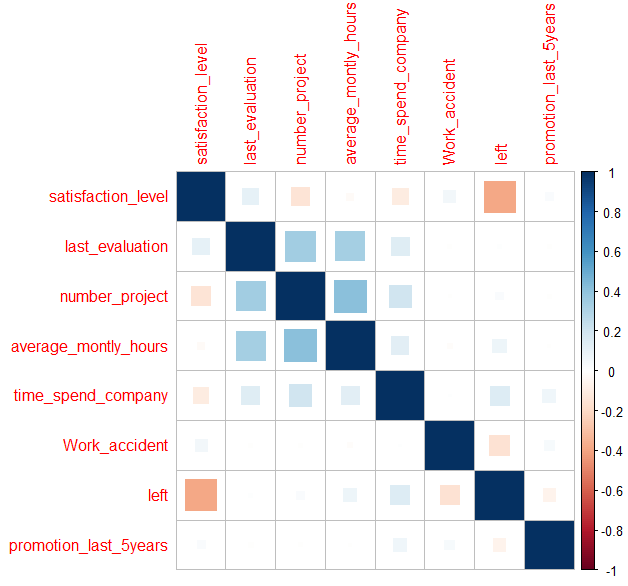


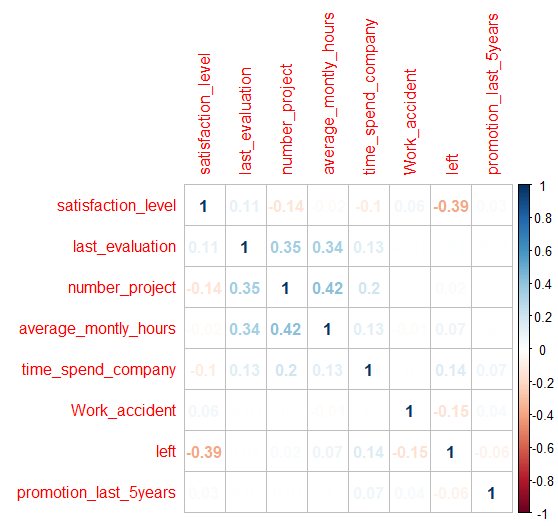
We did a similar analysis for the rest of the continuous variables for Left=1/Yes (see chart below). Here, we can see the distributions of only those employees who left. We can see from a glance that promotion has minimal impact since most people who left did not have an accident. Promotions may have an impact as most people who left did not have a promotion, possibly meaning the lack of promotions was a deciding factor. The other factors have somewhat of a bimodal curve, except for time spent with the company. We can see that the attributes of employees who left have high distributions where time with the company was short (low avg. monthly hours, number of projects, and time spent with company), indeed indicating that there are a lot of employees who left prematurely.



**Variable Correlations**

The following chart shows the correlations of between each of the numeric variables. One important correlation is with "left" and "satisfaction\_level", suggesting a relatively strong negative correlation; this means that employee satisfaction may be an important factor in whether an employee decides to leave, which is reasonable and expected. The other notable correlations are between positive correlations between "last\_evaluation" and "number\_project", "last\_evaluation" and "average\_montly\_hours", and "number\_project" and "average\_monthly\_hours". It is important to note, however, that no of the correlations are larger than 0.42, we speak about "strong correlations" in relative terms.

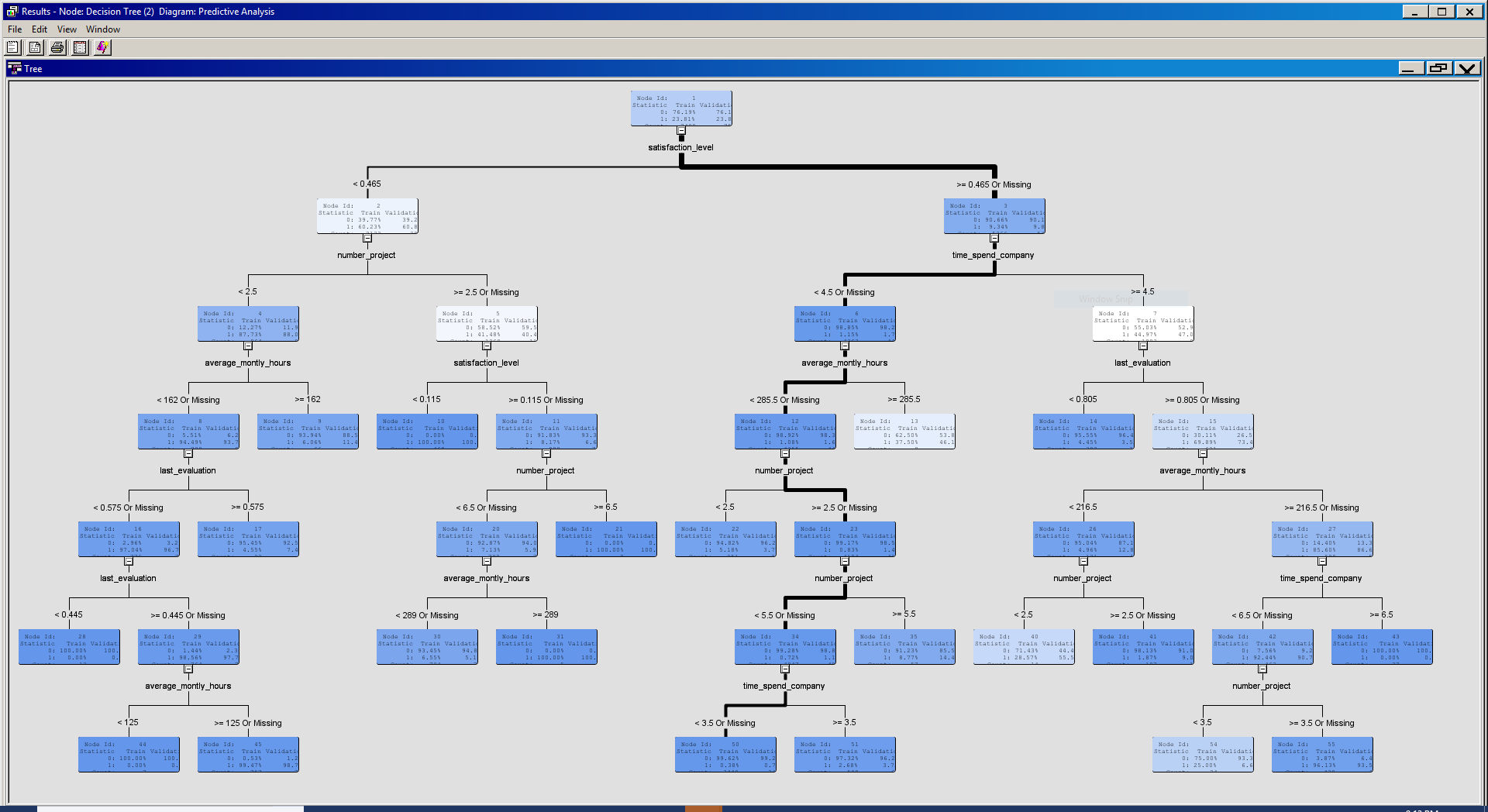


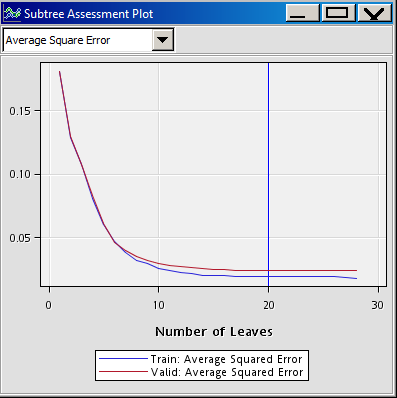
s

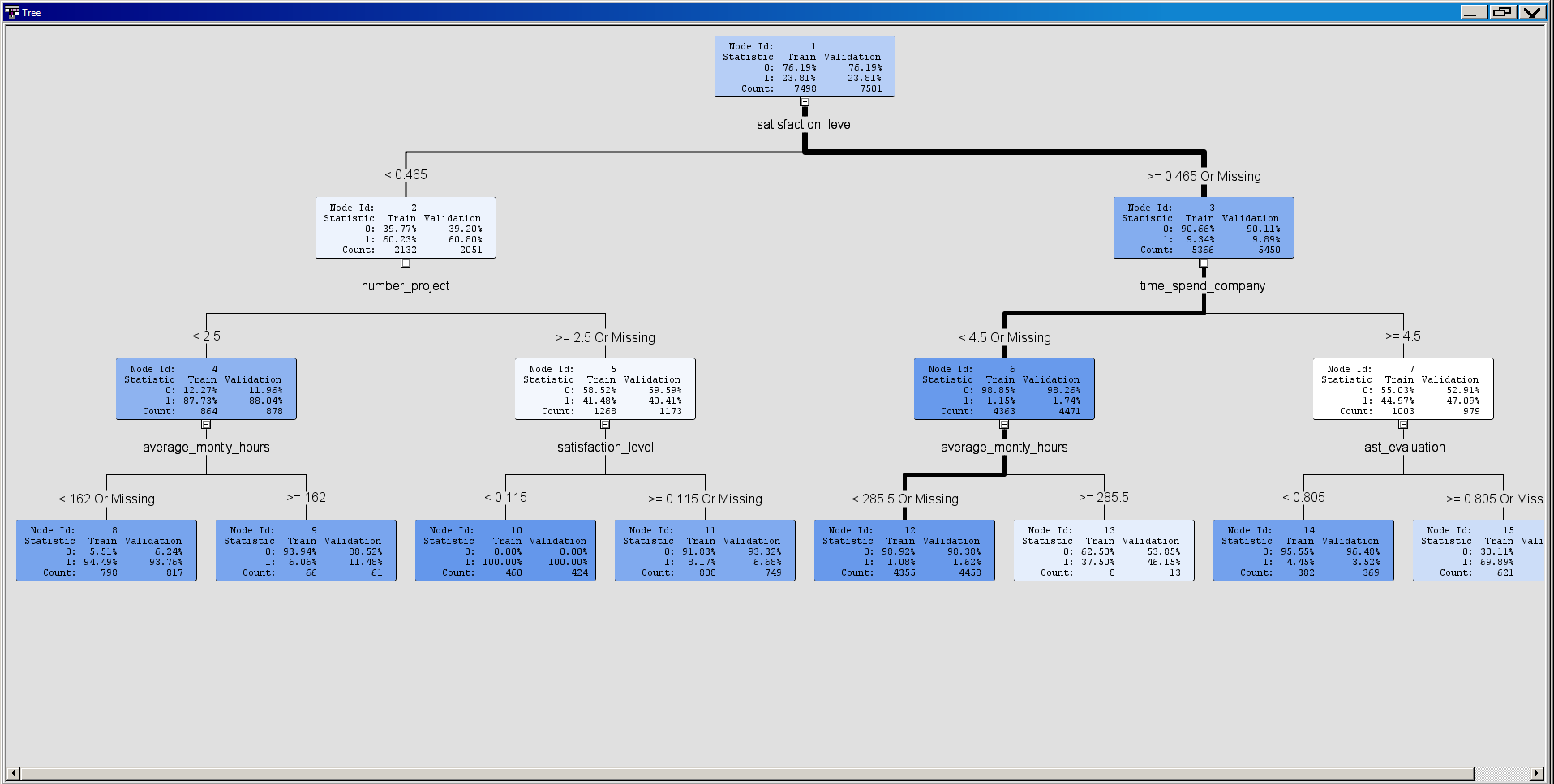
**Dataset Analysis: Models for analyzing why Employee tends to leave Company**

**Model 1: Decision Tree Analysis: Determine which employees will leave**

The first model we built was a decision tree. Below is the initial decision tree using all variables. For this model, we used a 50/50 split for training and validation data sets. As it is rather large, we wanted to see what the optimal number of leaves is.

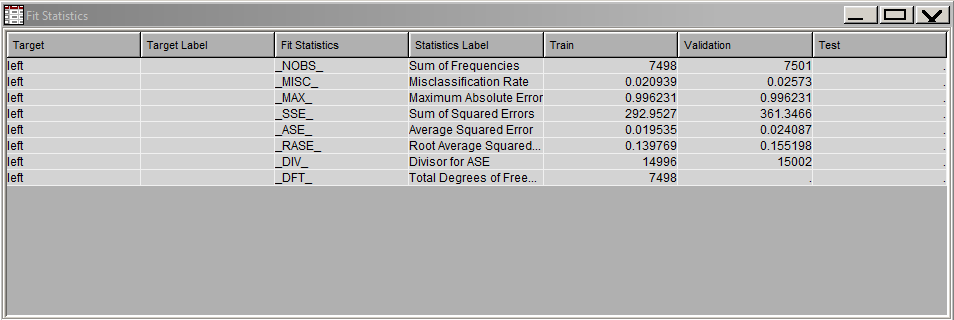
The Subtree Assessment Plot below shows the optimal number of leaves to be around 5-6, as that is where the training and validation sets begin to diverge and have larger difference in errors. Our original model already had a leaf size of 6, so we kept it as is.



The original tree was set at a depth of 6, but we lowered it to 3 as beyond 3 we began to see some variables split in the same variable. Below is our modified tree.

**Model Evaluation:**

These are the fit statistics for the first model:

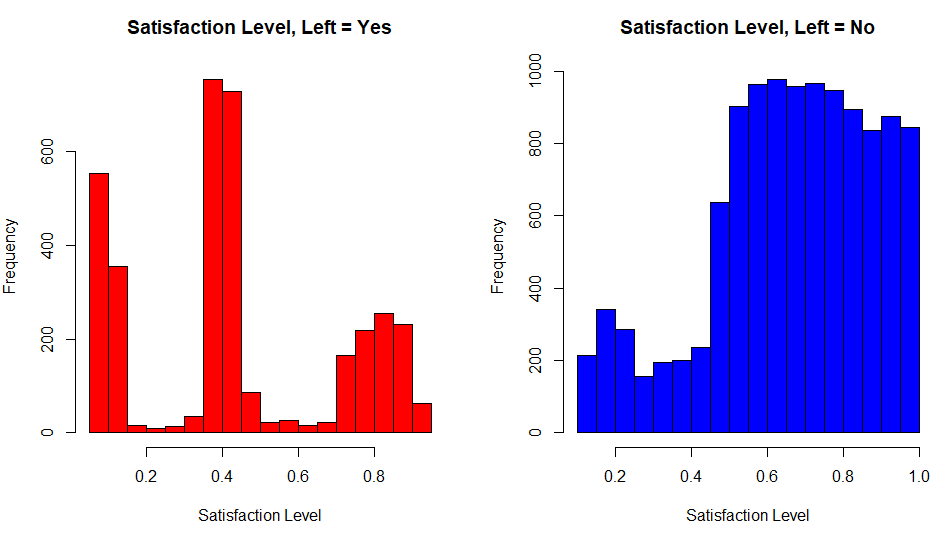


The model has a low average square error which shows a good fit for the model.

We also tried removing variables from the model that weren't as highly correlated to the first split, *satisfaction\_level,* though the results were the same.

**Model Conclusions:**

The first key takeaway from this model is the first branch, we divided by *satisfaction level*. This, in addition to the regression model, indicates that *satisfaction level* is the variable that has the most indication of whether the employee will leave. From here, we wanted to further explore *satisfaction level*:



Based on this histogram, in which we divided the distributions of satisfaction level by employees who left (left graph) and employees who did not (right graph), there are some notable distributions in employees who left and have a satisfaction level in the ranges 0.0 to 1.0, 3.5 to 4.5, and 7.0 and above, with the most being in the 3.5 to 4.5 range. This is interesting as it shows that even those employees with high satisfaction level may still leave, so the other variables may provide an explanation as to why.

After analyzing the tree, we see that the employees most likely to leave are those with *satisfaction\_level* of less than 0.465, *number\_project* of less than 2.5, and *average\_montly\_hours* of less than 162. This suggests that employees who left were unsatisfied and appeared to have worked less projects and less hours.

Conversely, employees more likely to stay had satisfaction\_level of 0.465 or greater, time\_spend\_company of less than 4.5, and average\_montly\_hours of less than 285.5. For employees who's time\_spend\_company was greater than 4.5, there was a more even split between employees who left and employees who didn't.

**Model 2: Analysis for Predicting Which valuable employee**

**might leave company**

The next model we built was a Logistic Regression model to predict the which valuable employees who might leave the company. To build this model, we first took a subset of the data that we found most useful.

**Step 1: Valuable Employees**

For analyzing valuable employee, first we get subset of dataset which contains only those employees who are valuable to company which includes -

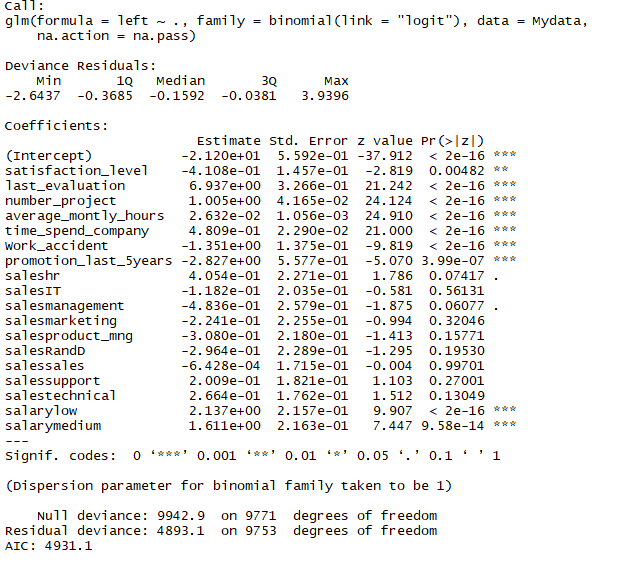
* Employee who are working for more than 4 years or
* Who have worked on more than 5 projects or
* Employee having last evaluation greater than 0.7

Our Initial dataset had 14999 observations whereas after applying these criteria for valuable Employees, our dataset contains 9772 observations.

**Step 2:** **Using Logistics Regression for the prediction:**

mylogit <- glm(left ~ . , data=Mydata, family=binomial(link="logit"), na.action=na.pass)

logistics regression is performed “left” dependent variable Vs. all the other variables of the dataset.



Coefficient value for independent variable are -

**Satisfaction:** -4.108e-01 implies as exp(-0.4108)= 0.66

With one unit decrease in satisfaction level, chance of employee leaving company increase by 0.66.

**average\_montly\_hours** : 2.632e-02 implies as exp(0.02632)= 1.02

Which implies that with one-unit increase in average monthly working hour, odd ratio of employee leaving company increase by 1.02.

**Number\_project** :1.005e+00 implies as exp(1.005) = 2.73

with unit increase in number of projects chance of employee leaving company increase by 2.73

**Time\_spend\_company** : 4.809e-01 implies exp(0.4809) = 1.62

With Unit increase in time spend in company, chance of employee leaving company increase by 1.62

**Promotion\_last\_5years** : -2.827e+00 implies exp(-0.02827) = 0.97

With unit decrease in Promotion\_last\_5years attribute, chance of employee leaving company increase by 0.97.

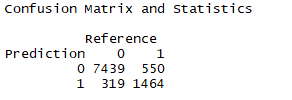
**Step 3: Making predictions based on logit results:**

The predicted data was not a categorical value, instead it was continuous values in between [0-1]. To compare the Prediction and Reference values we make predicted data to categorical with criteria as if the predicted value >0.5, consider them to more likely to leave soon and assign 1, else 0.

Mydata$predicted <- ifelse (pred > 0.5, 1, 0)

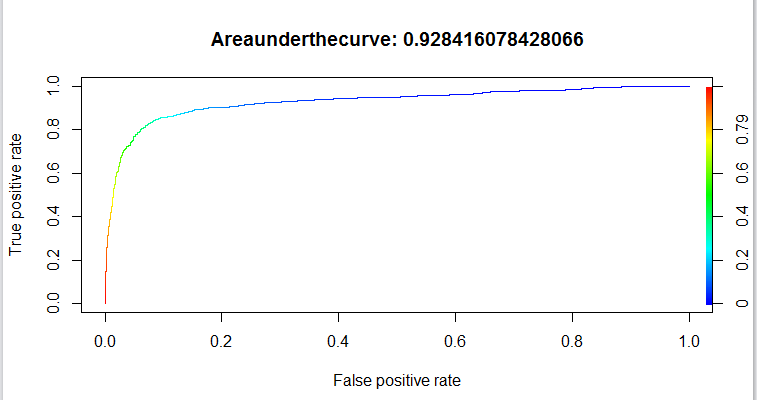
**Step 4: Creating Confusion matrix to compare the prediction & Reference**

To check TPR and FPR and compare the prediction and Reference we created confusion matrix.



Since the False Positive and False Negatives are lower, the prediction can be considered good.

**Step 5: Calculation Accuracy and plotting ROC Curve:**



From the plot, we can see that model is having accuracy of around 92 %,

From the AUC value and ROC curve we can say that the model works well for the data set.

**Model Conclusion:**

From result and accuracy level of model, we can say that by considering predicted values, company HR can get idea about who is going to leave the company next and then accordingly make some arrangement to retain that employee.

**Conclusion:**

In conclusion, we analyzed distributions of variables, created a Decision Tree, and created a predictive regression model to

Based on our analyses, we found

**Recommendations:**

The

