

Machine Learning Assignment

COURSE

Machine Learning

Submitted To

Institute of System Science

***Submitted By***

* Sreekumar Premalakshmy Gopakumar-A0148572X
* Junaid Ramis -A0163287X
* Thushara Joseph -A0163270M
* Monalisa Debnath – A0163196Y

Project Proposal

**Description**

We intend to consider the HR survey results and understand why employees are leaving prematurely?

Fields in the dataset include:

* Satisfaction Level
* Last evaluation
* Number of projects
* Average monthly hours
* Time spent at the company
* Whether they have had a work accident
* Whether they have had a promotion in the last 5 years
* Departments (column sales)
* Salary
* Whether the employee has left

Using Machine learning we will try to predict which valuable employees will leave next. And also create data explorations based on the dataset in hand.

At large the application will be developed using R shiny and R for data analysis and presentation.

At the data exploration phase, we are also considering using Python along with R so as to understand the differences in the platforms.

Dataset Links

The data is from Kaggle

<https://www.kaggle.com/ludobenistant/hr-analytics-1/data>

**Technologies and Tools**

Technologies: - R, R Shiny

Tools: - R Studio

**Problem Statement**

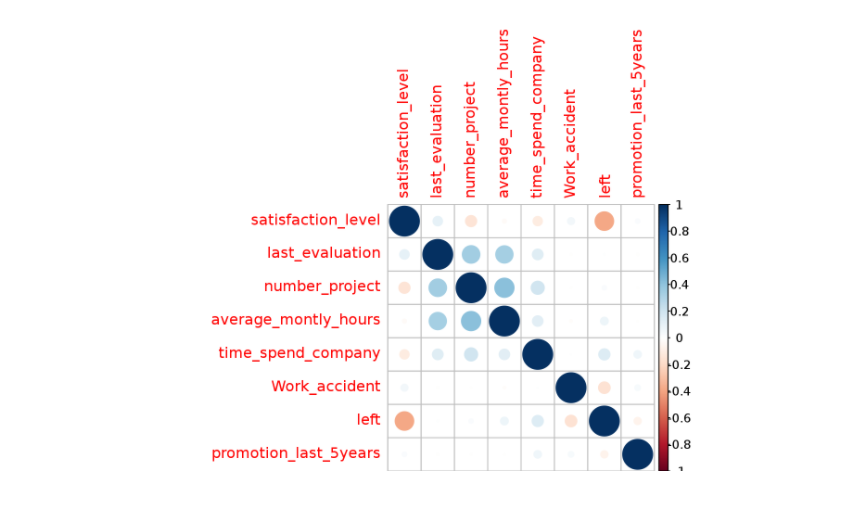
* Big company that wants to understand why some of their best and most experienced employees are leaving prematurely.
* The company also wishes to predict which valuable employees will leave next.

**Analysis**

* We have two goals: first, we want to understand why valuable employees leave, and second, we want to predict who will leave next.
* Therefore, we propose to work with the HR department to gather relevant data about the employees and to communicate the significant effect that could explain and predict employees' departure.
* We have the following details
* For our 15 000 employees we know: satisfaction level, latest evaluation (yearly), number of project worked on, average monthly hours, time spend in the company (in years), work accident (within the past 2 years), promotion within the past 5 years, department and salary.

**Data exploration**

This graph presents the correlations between each variable. The size of the bubbles reveals the significance of the correlation, while the colour present the direction (either positive or negative).

****

**Analysis based on Exploration**

* On average people who leave have a low satisfaction level, they work more and didn't get promoted within the past five years.

**Who is leaving?**

* Let's create a data frame with only the people that have left the company, so we can visualize what is the distribution of each features:
* Code

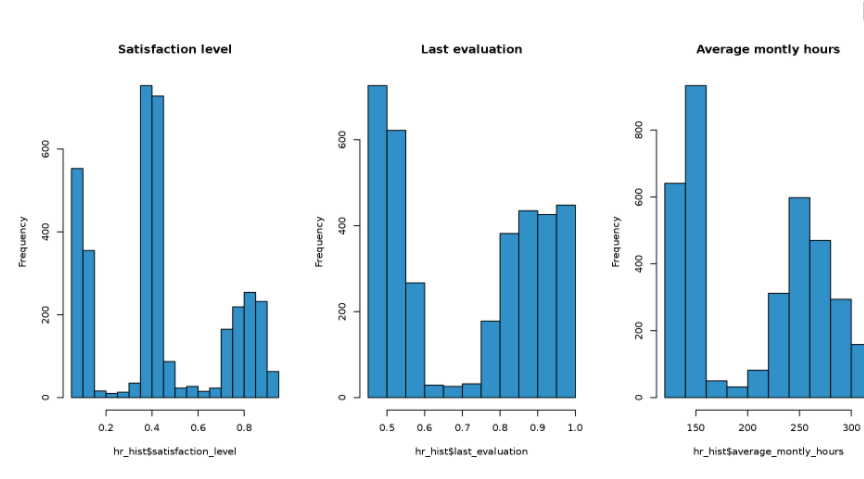
hr\_hist <- hr %>% filter(left==1)

par(mfrow=c(1,3))

hist(hr\_hist$satisfaction\_level,col="#3090C7", main = "Satisfaction level")

hist(hr\_hist$last\_evaluation,col="#3090C7", main = "Last evaluation")

hist(hr\_hist$average\_montly\_hours,col="#3090C7", main = "Average montly hours")

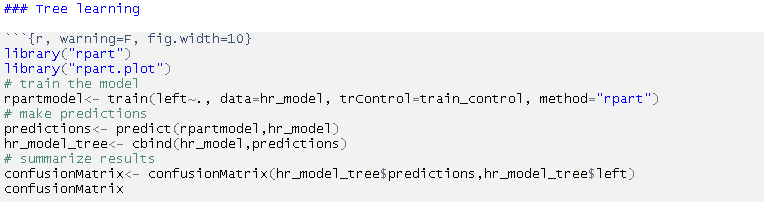
****

**Preparing data**

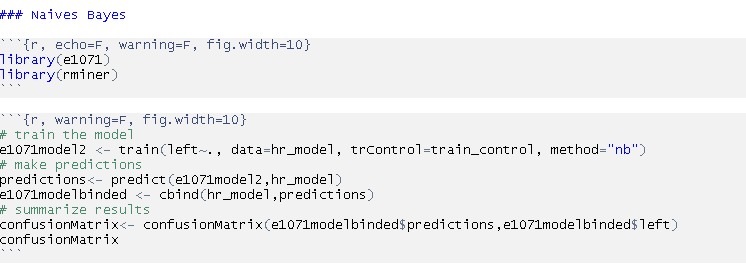
* Now we want to predict which valuable employe will leave next.
* Select database
* Let's use the same database than above where we kept the most valuable employees. Here is the summary of that database.

**Selecting Algorithm**

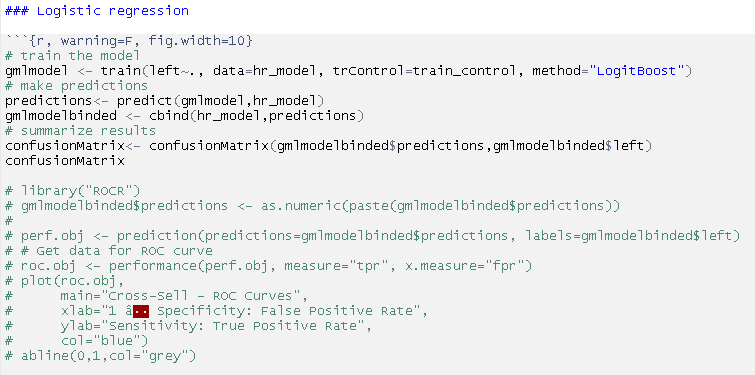
**Tree learning Algorithm**

****

**Naive bayes**

****

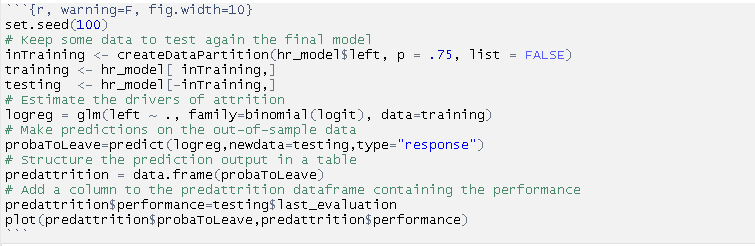
**Logistic Regression**

****

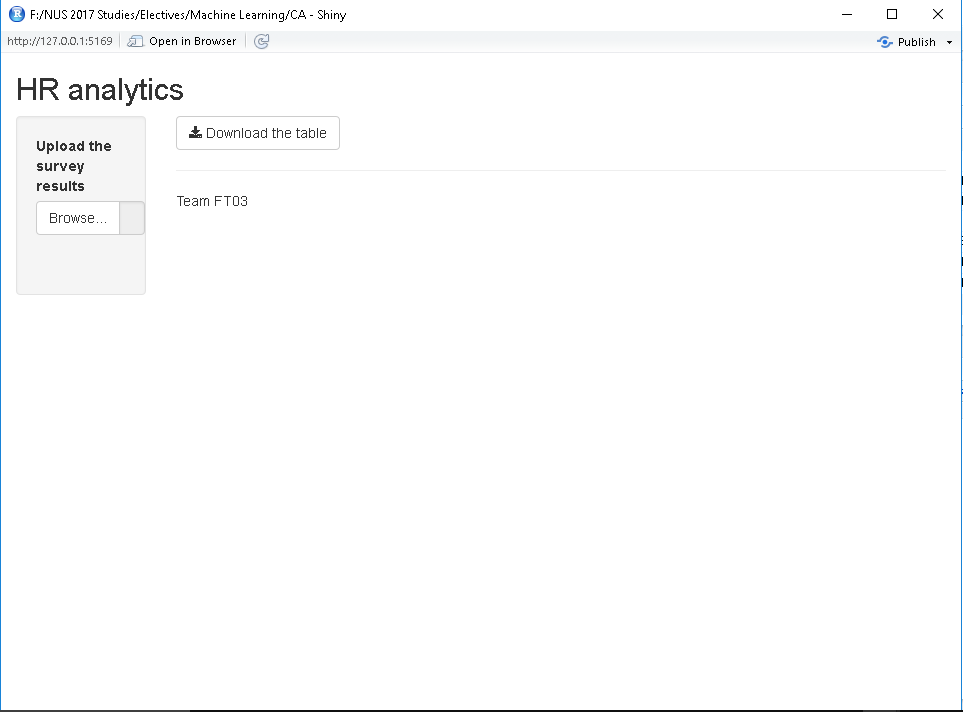
**Conclusion: Logistic Regression is the better option**

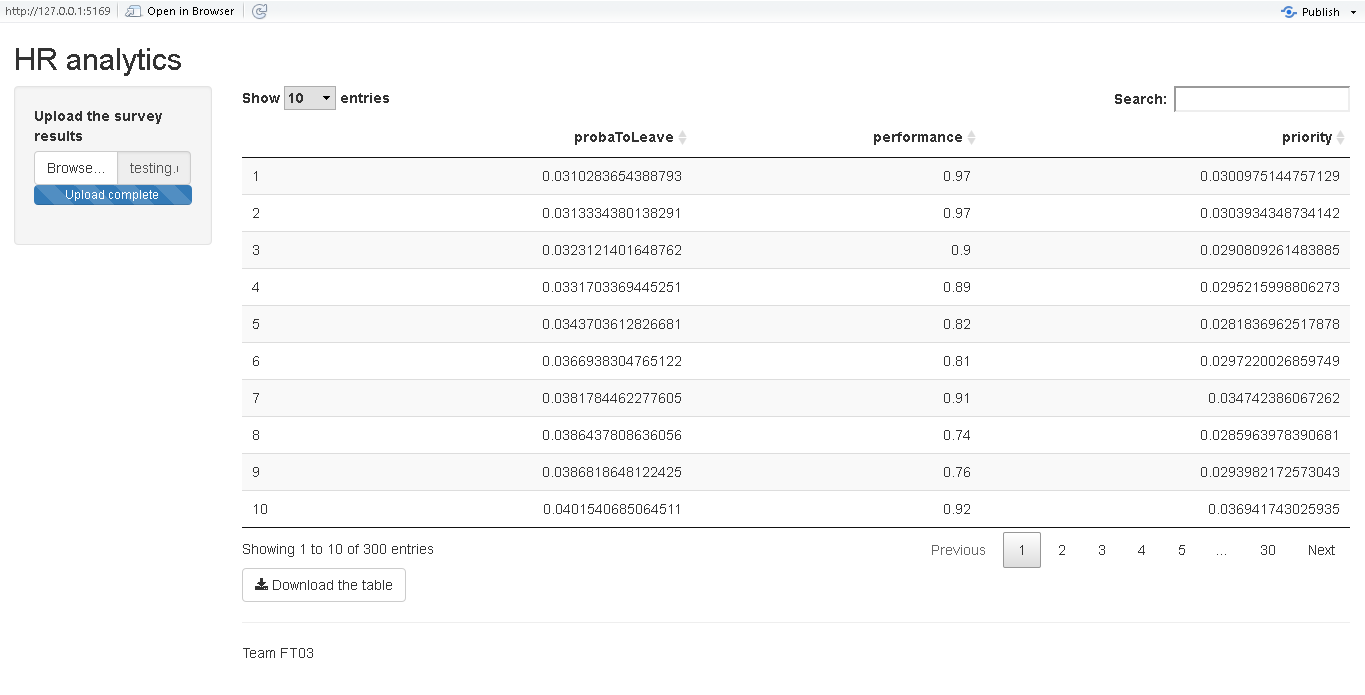
* The confusion matrix and the accuracy figures of the different model show that the predictive power is very similar and seems robust.
* About 95% accuracy and for a Kappa of 84%.
* We decide to keep the logistic regression model to lay out actionable insights. It's a very simple model and give the best results.

**Training the model and Testing with 30%**

****

**Shiny application**

****

****

**Installation steps**

* Run the **“selectingAlgorithm.rmd”** file to understand the data and why we selected logical regression algorithm
* Now install R studio and the necessary libraries listed here
* library(shiny) library(dplyr) library(tidyr) library(ggplot2) library(ggvis) library(corrplot) library(DT) library("caret")
* Now run the **Server.R** file in R studio to get the application deployed .
* Upload the testing.csv survey results and the application will display the predicted results of the first 300 employees most likely to leave.