Human Resource (HR) Analytics

Hardik Shah, Ishan Dhawan, Varun Srivastava, Wenlong Liao, Yash Kothari

Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907

shah400@purdue.edu, idhawan@purdue.edu, srivas53@purdue.edu, liao100@purdue.edu, ykothari@purdue.edu

1. Abstract

Human Resource Analytics (HR analytics) refers to applying analytic processes to the human resource department of an organization to analyze the employee performance and also make better judgements regarding the employee.

Our team aims to calculate the probability of an employee leaving the organization based on different parameters like employee satisfaction, last evaluation of the employee, average monthly hours worked and various other parameters.

We decided to conduct the project because the HR department has to make calculated judgements regarding an employee's future and also develop retention strategies accordingly so that they do not lose the organization's valuable employees to their nearest competitors.
Initially the team explored the data and generated a data quality report which described the statistic of each parameter in the dataset. From the prepared a Data Quality Report and found three main parameters which calculates the probability of an employee to leave. For the purpose of predictive modelling we used logistic regression because it showed the highest accuracy.

2. Business Problem

- Many fortune 500 companies suffer huge losses due to low employee retention. Nowadays, successful companies are those that are the most adept at attracting, developing and retaining individuals who can drive a global organization. Thus, the challenge for organizations is making sure they have the capability to find, assimilate, develop, compensate and retain such talented individuals.
- We conduct this project for the benefit of the stakeholders such as the owners and employees of the organization.
- The problem is amenable to be solved by analytics because we have adequate data and the desired correlations between the different parameters so that we can analyze our decision variable.
- Even though the organization would like to retain all the good employees there will always be an economic constraint leading the decision makers to select the best performing employees.
- The analysis will help the HR team of an organization to make better judgements about an employee's future based on their performances and other parameters. It will also help the organization to increase the retention rate of good employees.
- We have the agreement from all the stakeholders for the project.

3. Analytics Problem

- Using exploratory data analytics techniques, we determine that we can select few significant parameters which affect the retention rate and analyze the correlation between them.
- The analysis will help us to determine that how the variation in the selected parameters will affect our output which is the employee retention rate.
- In the long run if the good employees are retained, better projects can be undertaken with better execution and this will lead to increased profits. Also, if a good employee is about to leave the organization we can identify the parameter due to which he is leaving.

4. Data

- The data the team has taken involves 10 different parameters like satisfaction level, average monthly hours worked, employee satisfaction, which department does the employee belong etc.
- We have developed a correlation matrix from the different parameters and then identify which parameters will greatly influence our outcome.
- According to the correlation matrix we have identified the 3 main parameters that will influence the outcome. They are- satisfaction level of the employee, Last evaluation of the employee and the average number of monthly hours worked by the employee.

5. Data Exploration

```
  Employee.ID    Satisfaction.Level Last.Evaluation  No.of.Projects  Avg.monthly.hours Time.spent.on.job Work.Accident
Min.   :    1   Min.   :0.0900     Min.   :0.3600   Min.   :2.000   Min.   : 96.0     Min.   : 2.000    Min.   :0.0000
1st Qu.: 3750   1st Qu.:0.4400     1st Qu.:0.5600   1st Qu.:3.000   1st Qu.:156.0     1st Qu.: 3.000    1st Qu.:0.0000
Median : 7500   Median :0.6400     Median :0.7200   Median :4.000   Median :200.0     Median : 3.000    Median :0.0000
Mean   : 7500   Mean   :0.6128     Mean   :0.7161   Mean   :3.803   Mean   :201.1     Mean   : 3.498    Mean   :0.1446
3rd Qu.:11250   3rd Qu.:0.8200     3rd Qu.:0.8700   3rd Qu.:5.000   3rd Qu.:245.0     3rd Qu.: 4.000    3rd Qu.:0.0000
Max.   :14999   Max.   :1.0000     Max.   :1.0000   Max.   :7.000   Max.   :310.0     Max.   :10.000    Max.   :1.0000

      Left        Promotion.in.last.5yrs       Department      Salary
Min.   :0.0000   Min.   :0.00000        sales       :4140   high  :1237
1st Qu.:0.0000   1st Qu.:0.00000        technical   :2720   low   :7316
Median :0.0000   Median :0.00000        support     :2229   medium:6446
Mean   :0.2381   Mean   :0.02127        IT          :1227
3rd Qu.:0.0000   3rd Qu.:0.00000        product_mng: 902
Max.   :1.0000   Max.   :1.00000        marketing   : 858
                                        (Other)     :2923
```

Figure 1: Summary of the database

This table describe the statistics of each parameter of the dataset.
We can see that the mean attrition rate of employees left is approximately 24%
Satisfaction level is around 61%
Performance average is around 72%.
On average, people work on 3 - 4 projects a year and about 200 hours per month.

The figure below presents the correlations between each variable. The size of the bubbles reveals the significance of the correlation, while the color presents the direction (either positive or negative).

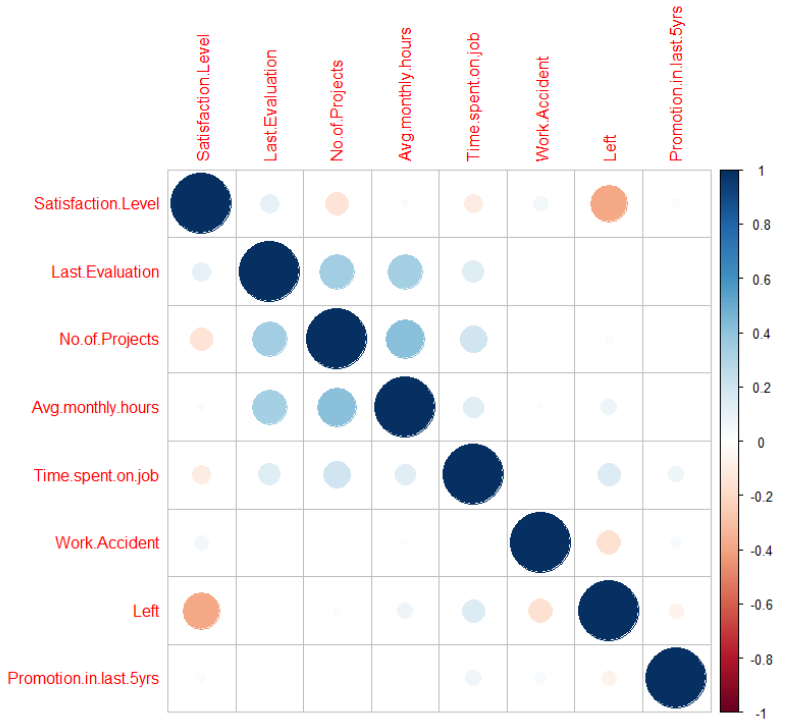


Figure 2: Correlation table for all employees

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

5.1 Study of the employees who have left

The study of the frequency distribution of the parameters of the employees that have left the company shows why we don't want to retain everybody. Some people don't work well as we can see from their evaluation, but clearly there are also many good workers that leave.

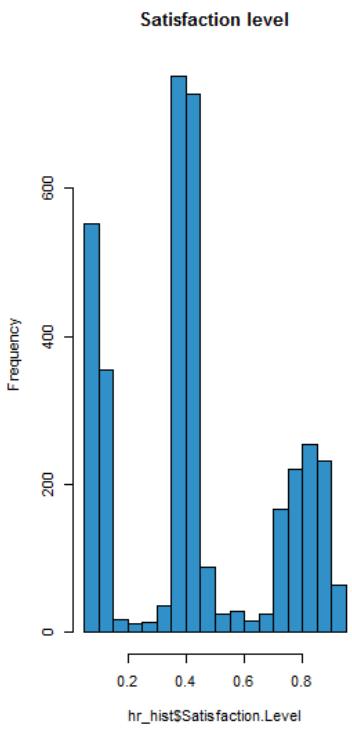


Fig2: Frequency of Satisfaction level

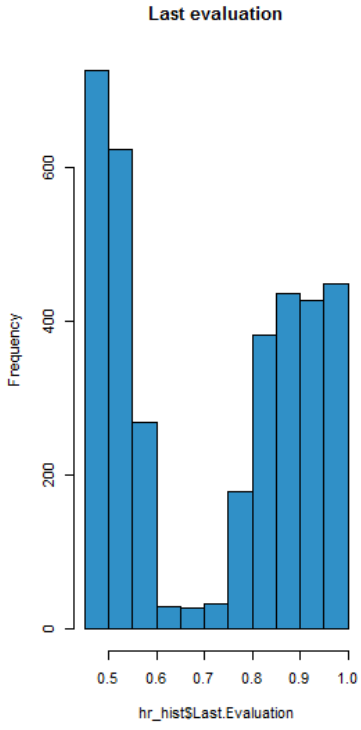


Fig3: Frequency of Last Evaluation

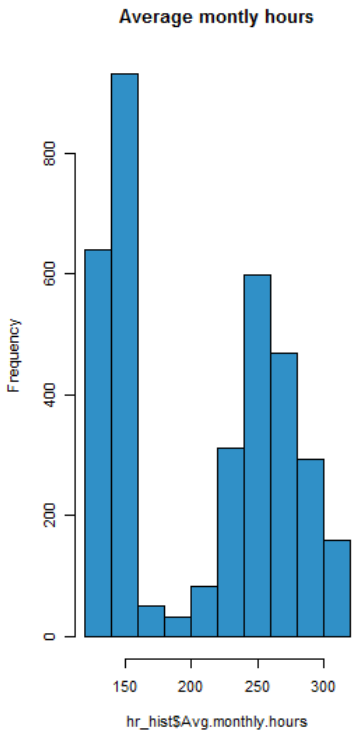


Fig4: Frequency of Average monthly hrs

The study of the frequency distribution of the parameters of the employees that have left the company shows that majority of the employees that left had low satisfaction level, low evaluation why we don't want to and spent less time on the job. But clearly there are also many good workers that leave.

More problematic, here are the total of employees that received an evaluation above average or were working on more than 4 projects at the same time and still have left the company. These are the people the company should have retained. In the total of 15000 employees that compose our database, 3571 people left out of which 2020 employees were above average. But, due to constraints we cannot retain all the 2020 employees. Hence identification of the most valuable employees is necessary.

5.2 Study of valuable employees that left

The most valuable employees i.e. the employees having above average - last evaluation, time spent in company, number of projects helps us understand why they tend to leave.

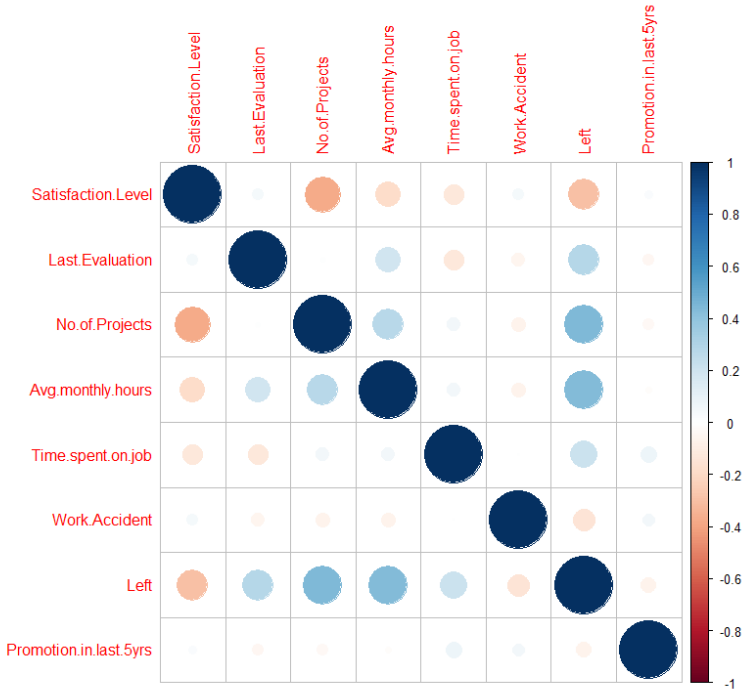


Figure 5: Correlation table for valuable employees

The above correlation table confirms that on average, valuable employees that leave are not satisfied, work on many projects, spend many hours in the company each month and aren't promoted.

Figure 6: Ratio of valuable employees who left (2020 out of 3571)

6. Methodology and Model Building

The objective is to predict which valuable employee will leave next. For this the database is filtered to include the most valuable employees i.e. the employees having above average - last evaluation, time spent in company, number of projects. This database contains 10394 employees.

The data exploration identifies that the top three parameters that affect a valuable employee are satisfa ction level, Last evaluation and No. of projects.

```
  Employee.ID    Satisfaction.Level Last.Evaluation  No.of.Projects  Avg.monthly.hours Time.spent.on.job Work.Accident
Min.   :    2   Min.   :0.0900     Min.   :0.3600   Min.   :2.000   Min.   : 96.0     Min.   : 2.000    Min.   :0.000
1st Qu.: 4020   1st Qu.:0.5000     1st Qu.:0.7100   1st Qu.:3.000   1st Qu.:171.0     1st Qu.: 3.000    1st Qu.:0.000
Median : 7588   Median :0.6800     Median :0.8200   Median :4.000   Median :217.0     Median : 4.000    Median :0.000
Mean   : 7600   Mean   :0.6224     Mean   :0.7883   Mean   :4.209   Mean   :211.3     Mean   : 3.838    Mean   :0.152
3rd Qu.:11220   3rd Qu.:0.8300     3rd Qu.:0.9100   3rd Qu.:5.000   3rd Qu.:252.0     3rd Qu.: 5.000    3rd Qu.:0.000
Max.   :14998   Max.   :1.0000     Max.   :1.0000   Max.   :7.000   Max.   :310.0     Max.   :10.000    Max.   :1.000

      Left         Promotion.in.last.5yrs       Department      Salary     left
Min.   :0.0000   Min.   :0.00000        sales       :2809   high  : 896   0:8374
1st Qu.:0.0000   1st Qu.:0.00000        technical   :1913   low   :4944   1:2020
Median :0.0000   Median :0.00000        support     :1560   medium:4554
Mean   :0.1943   Mean   :0.02319        IT          : 858
3rd Qu.:0.0000   3rd Qu.:0.00000        product_mng: 616
Max.   :1.0000   Max.   :1.00000        marketing   : 588
                                        (other)     :2050
```

Figure 7: Summary of the filtered database

After setting the cross-validation the next step is building and comparing different predictive models. The first one uses a tree model, the second is naives bayes and the third is logistic regression.

7. Predictive Modelling

The valuable employee database is split into two parts namely – Training and Testing. 75% data goes into training and the prediction is done based on this dataset. Based on the data exploration the drivers of attrition are identified and incorporated into the model
Left = Satisfaction level + Last evaluation + Avg. monthly hours

The results for the three predictive models after setting up the cross validation are shown below

Parameter	Tree Learning	Naives Bayes	Logistic regression
Accuracy	0.9574	0.9288	0.9573
Sensitivity	0.98	0.9878	0.9894
Kappa value	0.8586	0.76	0.8632
P - value	2.8 e-07	< 2.2e-16	< 2.2 e-16

Table 1: Results for the three predictive models

The confusion matrix and the accuracy figures of the different models 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 give the best results lay out actionable insights.

7.1 Estimating drivers of attrition

a) Variable Importance: To assess the relative importance of individual predictors, we look at the absolute value of t-statistic for each model parameter.
Satisfaction level = 0.2467297
Average monthly hours = 0.4234960
Last evaluation = 0.3297742
This suggests that satisfaction level is the most important factor followed by average monthly hours and last evaluation.

b) Wald Test: Wald test is used to evaluate the statistical significance of each coefficient in the model
The idea is to test the hypothesis that the coefficient of an independent variable
in the model is significantly different from 0.
For evaluating this, the R function used was ***regTermTest.***
Satisfaction level:
F= 338.352 p-value<- <2.22e-16
Average monthly hours
F= 604.4484 p-value= <2.22e-16
Last evaluation
F= 996.8379 p-value= <2.22e-16

From the Wald Test, we verify that the satisfaction level is the most important factor followed by average monthly hours and last evaluation.

7.2 Insights from the model

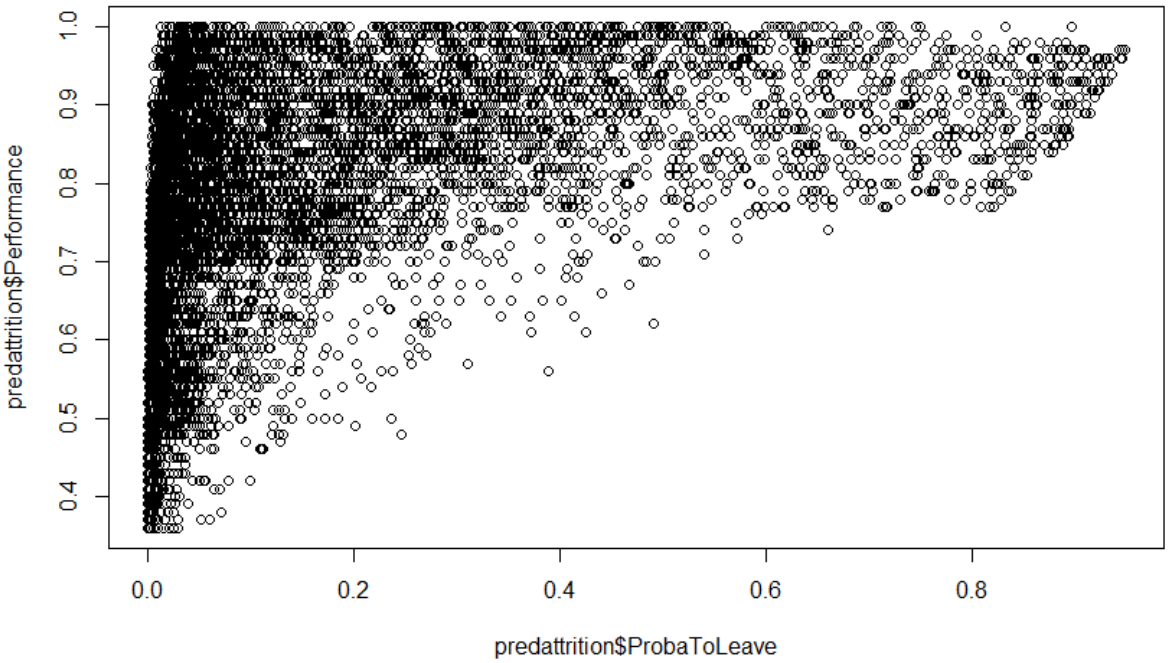


Figure 8: Probability to leave vs. Performance

Here is a plot that show the probability of the employee to leave vs their performance. We need to focus on the top right since these employees have high performance, but still have a high probability of leaving. To do that we build a data table where we rank the probability to leave found in the logistic regression model and the performance, we therefore find the priority for the company.

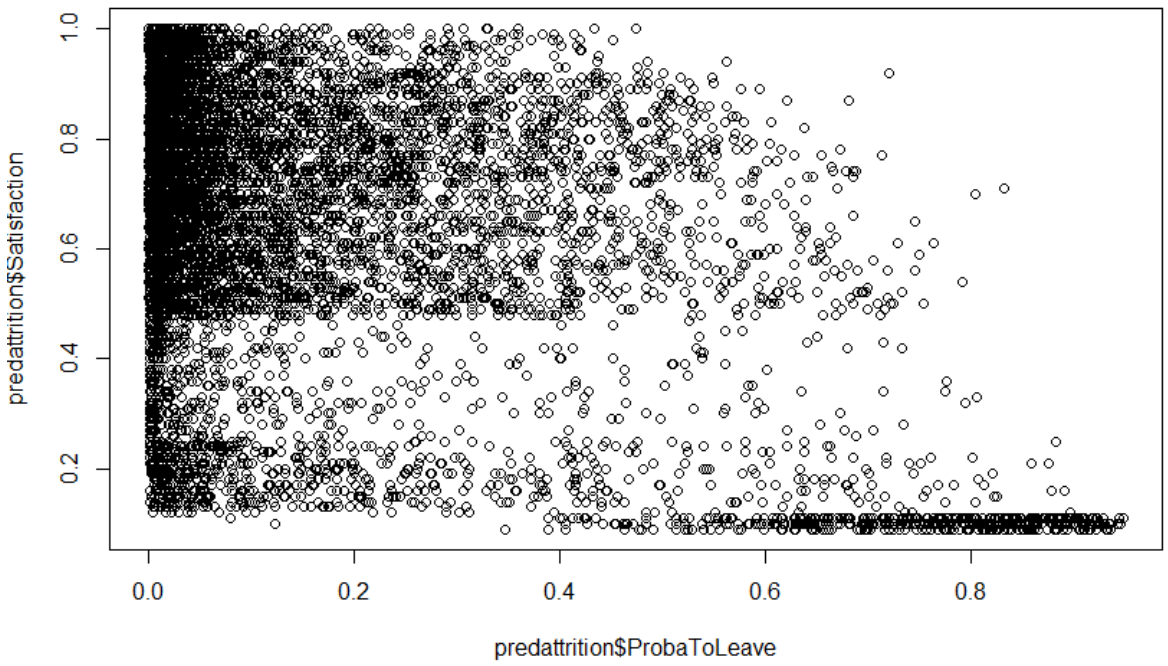


Figure 9: Probability to leave vs. Satisfaction

Here is a plot that shows the probability of the employee to leave vs. their satisfaction level. We need to focus where satisfaction levels are low. To do that we build a data table where we rank the probability to leave found in the logistic regression model and the satisfaction, we therefore find the priority for the company.

8. GUI Design and Functionality

i) The DSS helps to predict which valuable employee leaves next.
ii) It also ranks and prioritizes the valuable employees helping the employer to make decision on how many employees to retain based on the constraints.
iii) It also helps the employer to predict the probability of an employee not in the dataset by inputing his/her information

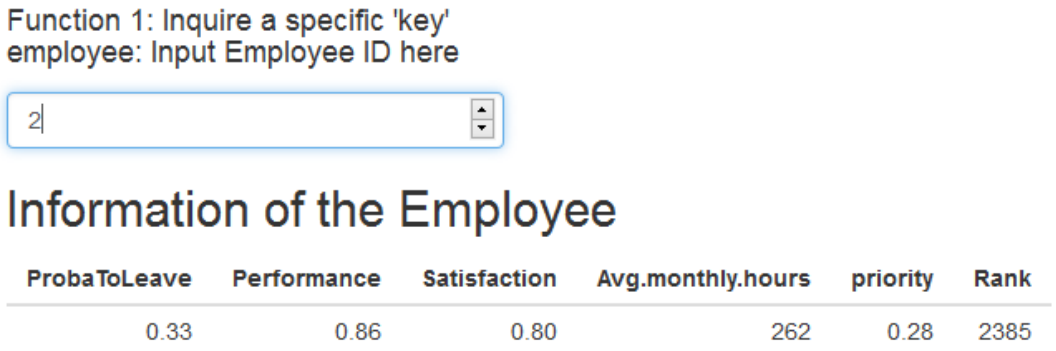
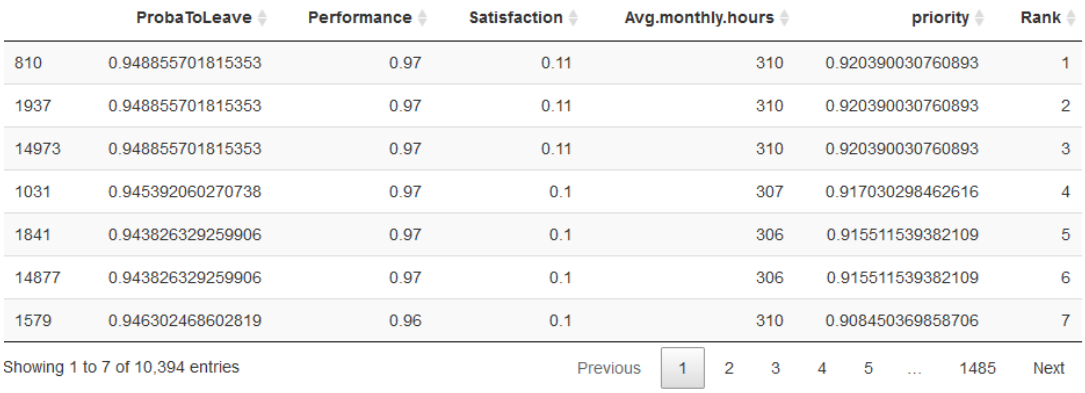


Figure 10: Which valuable employee leaves next



	ProbaToLeave	Performance	Satisfaction	Avg.monthly.hours	priority	Rank
810	0.948855701815353	0.97	0.11	310	0.920390030760893	1
1937	0.948855701815353	0.97	0.11	310	0.920390030760893	2
14973	0.948855701815353	0.97	0.11	310	0.920390030760893	3
1031	0.945392060270738	0.97	0.1	307	0.917030298462616	4
1841	0.943826329259906	0.97	0.1	306	0.915511539382109	5
14877	0.943826329259906	0.97	0.1	306	0.915511539382109	6
1579	0.946302468602819	0.96	0.1	310	0.908450369858706	7

Showing 1 to 7 of 10,394 entries Previous 1 2 3 4 5 … 1485 Next

Figure 11: Retaining decision based on rank

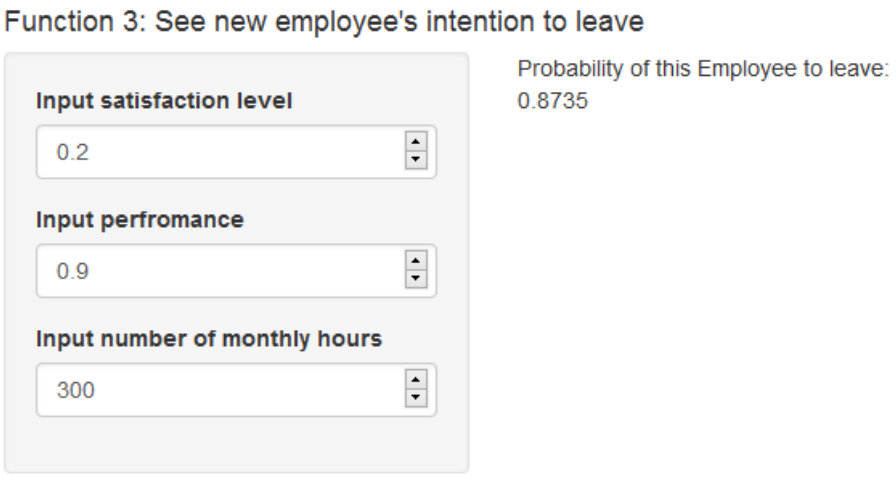


Figure 12: Predict probability of an employee not in the dataset

The packages used are tabulated as follows:

DT	R data objects (matrices or data frames) can be displayed as tables on HTML pages, and DataTables provides filtering, pagination, sorting, and many other features in the tables.
dplyr	It provides a flexible grammar of data manipulation.
caTools	Contains several basic utility functions including: moving (rolling, running) window statistic functions, fast calculation of AUC, LogitBoost classifier, round-off error free sum and cumsum, etc.
lattice	It is a powerful and elegant high-level data visualization system with an emphasis on multivariate data. It is designed to meet most typical graphics needs with minimal tuning.
ggplot2	A system for 'declaratively' creating graphics, based on "The Grammar of Graphics". You provide the data, tell 'ggplot2' how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.
caret	The caret package is a set of functions that attempt to streamline the process for creating predictive models. The package contains tools for: data splitting, pre-processing, feature selection, model tuning using resampling, variable importance estimation etc.
magrittr	Provides a mechanism for chaining commands with a new forward-pipe operator, %>%. This operator will forward a value, or the result of an expression, into the next function call/expression.

Table 2: Package information

9. Conclusion

The main objective of our Shiny app was to help the HR department of a company to retain good employees who are about to leave the company in the future. The Correlation results of the good employees showed us that that the three main parameters affecting the attrition rate were satisfaction level of the employee, last evaluation of the employee and the average number of monthly hours worked by the employee, with satisfaction level being the most important one.

Hence, in order to predict the probability of a good employee leaving the firm in the near future, we used Logistic Regression to build our predictive model which gave us the list of employees ranked based on their probability of leaving as well as the priority of their retention according to the three parameters selected. Our predictive model was extended to tell the probability to leave of a new employee joining the company based on the inputs of the selected three parameters.

The Shiny App was designed in such a way that it could tell us the information about a specific employee, the ranked list of good employees about to leave, provision to find the probability of leaving of the new employees joining the firm as well as the visualizations of our predictive models.

We think that our app will help the HR department to figure out what decisions they need to make from the descriptive as well as the predictive data provided by our Shiny DSS app so that they can increase the retention rate of the current as well as the future employees. This will not only lead to the long-term profits for the company, but will also help in proper motivation of the talented pool of employees who are an asset to the firm.

10. References

1. Madsen, Dag Øivind and Slåtten, Kåre, The Rise of HR Analytics: A Preliminary Exploration (January 2, 2017). Global Conference on Business and Finance Proceedings, Volume 12, Number 1, pp. 148-159. Available at SSRN: https://ssrn.com/abstract=2896602
2. L. Bassi and D. McMurrer, "A Quick Overview of HR Analytics: Why, What, How, and When?" Association for talent development, March 04, 2015
3. L. Smeyers, "7 Benefits of Predictive Retention Modeling (HR analytics)." iNostix (May 6, 2013)
4. J Fitz-enz and J. R. Mattrox II, "Predictive Analytics for Human Resource." Wiley Publication, SAS Institute Inc., Cary, North America, USA, 2014 pp. 2-3
5. L. Bassi, "Raging Debates in HR Analytics", People & Strategy, Vol. 34, Issue 2, 2011 "Applying Advanced Analytics to HR Management Decisions," James C. Sesil, Pearson Publication, New Jersey, March 2014, pp. 13-25
6. Brian S. Everitt ,Torsten Hothorn,(2009),"Logistic Regression and Generalised Linear Models: Blood Screening, Women's Role in Society, and Colonic Polyps", A Handbook of Statistical Analyses Using R, Second Edition
7. "Data Mining - Predictive Analysis". Retrieved from https://stackoverflow.com/questions/2714289/data-mining-predictive-analysis
8. "Correlation and Regression". Retrieved from https://www.datacamp.com/courses/correlation-and-regression
9. "Human Resource Analytics". Retrieved from https://www.kaggle.com/ludobenistant/hr-analytics
10. "Function reference • ggplot2". Retrieved from http://ggplot2.tidyverse.org/reference/
11. "Some example graphs in corrplot 0.60", Retrieved from https://www.r-bloggers.com/some-example-graphs-in-corrplot-0-60/
12. "How to start with Shiny". Retrieved from https://www.rstudio.com/resources/webinars/how-to-start-with-shiny

11. Links

1. Github link: https://github.com/ykothari/HR-Analytics_UR4A-Project
2. App link: https://ykothari.shinyapps.io/HR-Analytics_UR4A-Project/