Human Resource Analytics - Report

Team Members

Mohan Krishna Chagalamarri - A20392859
Kavya Goli - A20392402

1. Introduction:

"The goal is to understand the factors that contribute most to employee attrition and create a model that can predict if a certain employee will leave the company or not"

Companies that maintain a healthy organization and culture are always a good sign of future prosperity. Recognizing and understanding what factors that were associated with employee turnover will allow companies and individuals to limit this from happening and may even increase employee productivity and growth. These predictive insights give managers the opportunity to take corrective steps to build and preserve their successful business.

2. Related Work:

In recent years, around 900+ people worked on this dataset in Kaggle website. Randy Lao and Zhibo Yang did amazing research on this dataset to predict whether certain employee will leave the company or not. Randy Lao used AdaBoost, Decision Tree and Logistic Regression Models to predict. Though the accuracy of base model is very low (0.5). The accuracy of the model increased rapidly to 0.78 with Logistic Model,0.93 with AdaBoost, 0.94 with Decision trees.

Zhibo Yang used K- Nearest neighbors along with the Logistic Regression and Decision Trees techniques and achieved around 0.96 of accuracy.

3. Approach:

The approach includes Standard Machine Learning pipeline of Knowing the problem, Obtaining, Preprocessing, Exploration, Modelling, Iterating and Insights.

1. Loading the Data:

The data set is used from Kaggle website named as Human Resource Analytics. The data is representative of real world as it covered most of the features required to solve the problem but the data is very clean as no much cleaning and preprocessing is required. In the real world the data might not be as clean as this data set.

2. Data Cleaning:

Generally, cleaning the data is important preprocessing step in data mining. This dataset is clean but we have examined the dataset to make sure that everything is readable and the observation values match the feature names appropriately.

a. There is no Unique Identifier for each employee. So, the new column named "ID" is added and assigned the unique value to each employee.
b. The column name "sales" is replaced with the meaningful name as "Department" as the data is representing the departments.
c. Checked for the NA and missing values in the data. The data is clean with zero NA values.

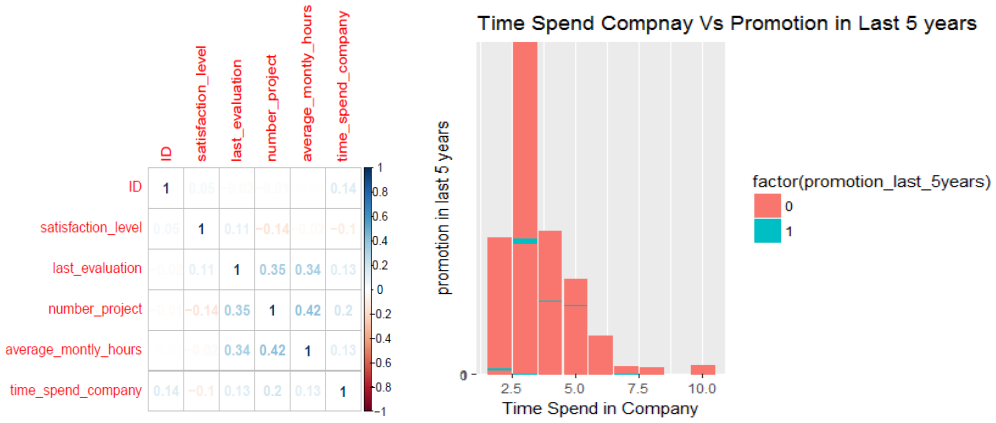
3. Exploring the Data and Preprocessing:

a. Converting the variables to proper data type – The features like left, salary, work_accident, department and promotion_last_5years are converted to categorical features.

b. Converted the Salary feature to Ordinal feature as (“high”,” medium”,” low”)

c. Finding the distribution for numeric variables and descriptive statistics.

d. Correlation matrix is build to understand the correlation between the features.



Summary of Data:

i. The dataset has about 15000 employees with 10 features and the turnover rate of company is about 24% and Mean satisfaction of the employee is 0.61

ii. The Correlation Matrix states that number_project vs last_evaluation ratio is 0.35, number_project vs average_montly_hours ratio is 0.42, average_montly_hours vs last_evaluation ratio is 0.34.

iii. People who spend 6 or more years and who spend 2 years at the company are less likely to go. People are more likely to leave when they have spent 3-5 years here.

iv. An interesting group: 5-year-group. People who are in this group are more likely to leave than stay. When the years people spent in the company lies in 3-5: the more they've been here, the more likely they leave.

v. Very less people got promoted even though they are spending more time in the office. Also, low and medium income people are leaving the company.

Who are Valuable Employees?

The evaluation criteria and Monthly hours spend in the company are considered for evaluating the valuable employees. Here we are not considering the promotion because very less people got promoted in last 5 years. For our analysis we are finding the average time an employee spent on each project. Then, we converted the variable into 3 levels.

In general, an employee must work for 160 hours per month. We have splitted this variable into 3 levels and then according to the level we have given categories as [0,1,2]

Rule for Finding Valuable Employees:

Valuable employees can be decided based on their department statistics. Variables Used to decide Valuable employees: average_montly_hours, last_evaluation while deciding threshold we have taken mean of each variable for each department.

Total valuable Employees according to above rule: 4754

Total number of employees :14999.

Decide who all are valuable employees:

	Department	number_of_employees	number_of_employees_left	percent
1	accounting	767	204	3.759804
2	hr	739	215	3.437209
3	IT	1227	273	4.494505
4	management	630	91	6.923077
5	marketing	858	203	4.226601
6	product_mng	902	198	4.555556
7	RandD	787	121	6.504132
8	sales	4140	1014	4.082840
9	support	2229	555	4.016216
10	technical	2720	697	3.902439

From the above table we can say that "management' and 'R and D' people are leaving more compared to other departments. The 'management' people are staying for long time and working for more hours in the company. The 'management' and 'R and D' people have more functional knowledge compared to other departments. So, we are considering these people as valuable.

4. Model Fitting and Selection:

a. Model Selection:

Model Selection is made using the Cp Model, LASSO, Forward and Backward Selection methods. From these feature selection techniques, we decided to we are selecting all variables from dataset because each have given importance to all 9 variables. The Null Hypothesis and T-test says that we cannot reject any of the features as all the features are important for predicting “left”.

b. Sampling Dataset Using Stratified Sampling:

Stratified sampling is used to divide the dataset into Train and test Data. Stratified sampling will use the strata for dividing the dataset. Stratified Sampling is used to divide the dataset because it reduced the bias and gave good trade off between bias and variance.

c. Base Line:

24% of the dataset contained 1's (employee who left the company) and the remaining 76% contained 0's (employee who did not leave the company). The Base Rate Model would simply predict every 0's and ignore all the 1's. The base rate accuracy for this data set, when classifying everything as 0's, would be 76% because 76% of the dataset are labeled as 0's (employees not leaving the company).

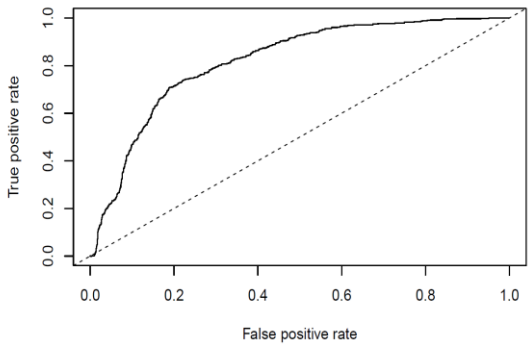
d. Logistic Regression:

Below are Confusion Matrix and ROC curve details of Logistic Regression Model:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2197  573
         1   88  141

               Accuracy : 0.7796
                 95% CI : (0.7643, 0.7943)
    No Information Rate : 0.7619
    P-Value [Acc > NIR] : 0.0117
```

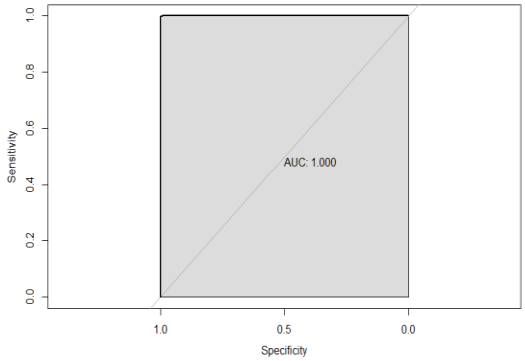


The Accuracy obtained using the Logistic regression is around 0.78 which is more than the base line accuracy. The ROC curve which represents the relation between true positive rate and false positive rate would be better compared to the base line.

e. Random Forest:

Below are Confusion and ROC curve details of Random Forest Model:

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction    0    1
##          0 2285    5
##          1    0  709
##
##                Accuracy : 0.9983
##                  95% CI : (0.9961, 0.9995)
##     No Information Rate : 0.7619
##     P-Value [Acc > NIR] : < 2e-16
```

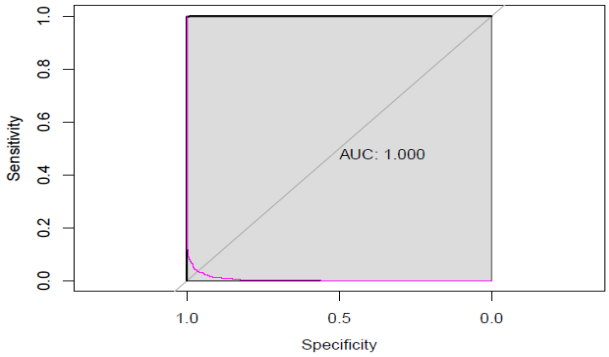


The Accuracy obtained using the Random Forest model is around 0.998 which is very accurate and high. The ROC curve is similar to right angled triangle due to high accuracy rate and good prediction.

f. Support Vector Machines:

Below are the Confusion and ROC curve details of Support Vector Machines.

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction    0    1
##          0 2255   49
##          1   30  665
##
##               Accuracy : 0.9737
##                 95% CI : (0.9673, 0.9791)
##    No Information Rate : 0.7619
##    P-Value [Acc > NIR] : < 2e-16
```

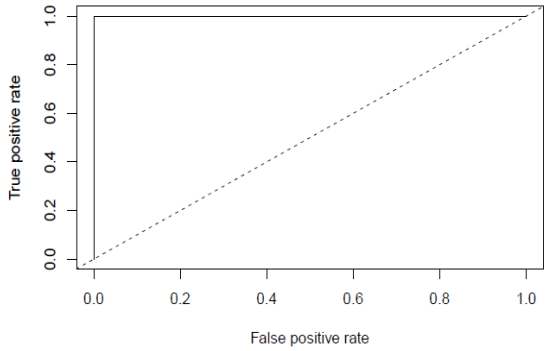


The Accuracy obtained using the Support Vector machines is around 0.97 which is good and better compared to baseline. The ROC curve is near to Right angled triangle as the accuracy is very high.

g. CART Implementation (Classification and Regression Trees):

Below are the Confusion and ROC curve details of CART Model.

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction    0    1
##          0 2285    0
##          1    0  714
##
##               Accuracy : 1
##                 95% CI : (0.9988, 1)
##    No Information Rate : 0.7619
##    P-Value [Acc > NIR] : < 2.2e-16
```



The Accuracy obtained using the CART implementation is around 0.998 which is similar to Random Forest Technique. The ROC curve is almost Right-angled triangle as the accuracy is very high.

4. Results:

S.No	Model	Accuracy (in %)
1	Base Line	76
2	Logistic Regression	78
3	Random Forest	99.83
4	Support Vector Machines	97.37
5	CART	99.8

The Baseline performance is beaten by all the models such as Logistic Regression, Random Forest, Support Vector Machines and CART. The Baseline did not consider other features into consideration for predicting the “left” predictor. The Model fits has improved the prediction on the test data and achieved the accuracy more than 97% minimum.

5. Conclusion:

The Plots and respective code for the below summary results are included in the code pdf file named (HR_Analytics_Code_A20392859_A20392402.pdf). The predictions that we can make based on the analysis of data and models are:

- Employees with either really high or low evaluations should be taken into consideration for high turnover rate.
- Employee Satisfaction plays the key role for employee turnover rate.
- Employees with 4 and 5 years of experience in the company should be taken into consideration for high turnover rate.
- Employees with Low and Medium salaries are the bulk of employee turnover.
- Employees generally left when they are underworked (less than 150hr/month or 6hr/day)
- Employees generally left when they are overworked (more than 250hr/month or 10hr/day)
- Employees that had 2,6, or 7 project counts was at risk of leaving the company.

6. Appendix:

The R file and pdf generated from R Markdown are attached along with this document in the zip.