Developing solutions for Employee Retention Management

Bhargava Vidiyala (WT1976)

Sai Vishwa Gaddam (BF6887)

Vigneswaran Anandan (BX6811)

Abstract

Employee turnover is a critical issue that companies face, as it not only leads to the need for new recruitment but can also be costly. The process of replacing an employee with a new one can be particularly expensive, especially when hiring for senior positions or those with unique expertise. HR agencies have noted that a significant portion of the new employee's salary is often allocated towards recruitment service charges. Additionally, there is also a time lapse between the departure of the old employee and the arrival of the new one, resulting in potential productivity losses for a few months or quarters. Moreover, the recruitment process, including interviews and other formalities, can be a significant burden for the organization. Despite the efforts to ensure good working conditions, employees still tend to leave companies, making it imperative for organizations to find ways to retain existing employees.

1.Introduction

Employee attrition implies a reduction in employment caused by employees who willingly quit or resign from a company. Employee turnover is defined as the number of existing employees who are replaced by new employees during a specific time period. Using past data stored in human resources (HR) departments, analysts can build and design a machine learning model that can predict which employees will leave the company.

The reasons for employees leaving the job should be known so that appropriate corrections can be made. The advanced knowledge regarding when employees may leave the job can equip HR to plan their activities with respect to

(I) solving the problem on hand so that retention can be stopped.

(ii) possible new hiring.

The solution is expected to address the problem so that employee retention can be encouraged and attrition can be kept at a very minimum. So that HR can use the information to provide ambiance and ensure high motivational levels in the organization.

The solution should specifically address the questions like

1) Possibility or likelihood of an existing employee leaving the company.

2)Key pointers of an employee leaving the company.

3)Actions, policies, or strategies that can be adapted based on the results to improve employee retention

2.DATASET

The HR dataset used in this study was obtained from Kaggle and comprises data collected from both employees and employers of a company. It contains 14999 observations and 10 variables which are described below.

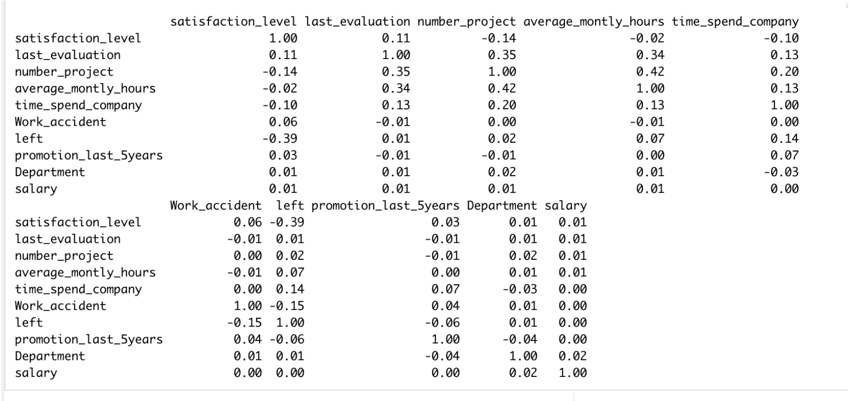
Table 2.1: Variables and their Description

	Variables	Description
1	Satisfaction_level	Tells the rate of satisfaction by the employee (0-1).
2	Last_Evaluation	Evaluation the performance on basis of last project (0-1).
3	Number_Project	Number of projects completed
4	Average_monthly_hours	Number of hours worked on average in a month
5	Time_spend_company	Total time spent in the company as an employee (in years)
6	Work_accident	Any accident at the time of work (0-No,1-Yes)
7	Left	Did he left the company or not (0-No,1-Yes)
8	Promotion_last_years	Whether he got the promoted in the last 5 years (0-No,1-Yes)
9	Department	Name of department in which employee is working
10	Salary	Classified into three factors(1-low,2-Medium,3-High)

Predictor Variables: Last_Evaluaiton, time_spent_company, number_project, average_monthly_hours, left,work_accident,promotion_last_years,Department,Salary.

Response Variable: Satisfaction Level

Correlation Matrix: Based on the data in Fig 2.1, it appears that the values are sufficient in demonstrating the absence of any correlation between the variables.



```
                      satisfaction_level last_evaluation number_project average_montly_hours time_spend_company
satisfaction_level                  1.00            0.11          -0.14                -0.02              -0.10
last_evaluation                     0.11            1.00           0.35                 0.34               0.13
number_project                     -0.14            0.35           1.00                 0.42               0.20
average_montly_hours               -0.02            0.34           0.42                 1.00               0.13
time_spend_company                 -0.10            0.13           0.20                 0.13               1.00
Work_accident                       0.06           -0.01           0.00                -0.01               0.00
left                               -0.39            0.01           0.02                 0.07               0.14
promotion_last_5years               0.03           -0.01          -0.01                 0.00               0.07
Department                          0.01            0.01           0.02                 0.01              -0.03
salary                              0.01            0.01           0.01                 0.01               0.00
                      Work_accident  left promotion_last_5years Department salary
satisfaction_level             0.06 -0.39                  0.03       0.01   0.01
last_evaluation               -0.01  0.01                 -0.01       0.01   0.01
number_project                 0.00  0.02                 -0.01       0.02   0.01
average_montly_hours          -0.01  0.07                  0.00       0.01   0.01
time_spend_company             0.00  0.14                  0.07      -0.03   0.00
Work_accident                  1.00 -0.15                  0.04       0.01   0.00
left                          -0.15  1.00                 -0.06       0.01   0.00
promotion_last_5years          0.04 -0.06                  1.00      -0.04   0.00
Department                     0.01  0.01                 -0.04       1.00   0.02
salary                         0.00  0.00                  0.00       0.02   1.00
```

Fig:2.1

3.Models and Results

Diagnostic Plots:

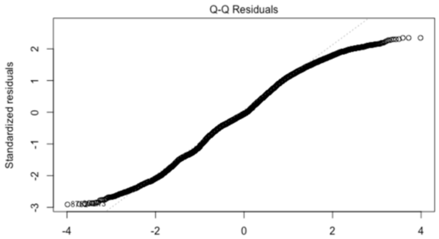


Fig 3.1

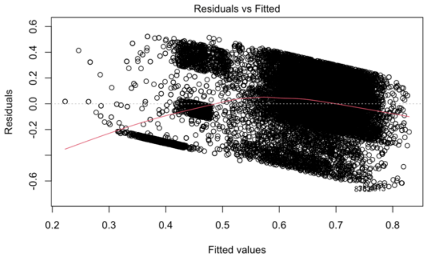


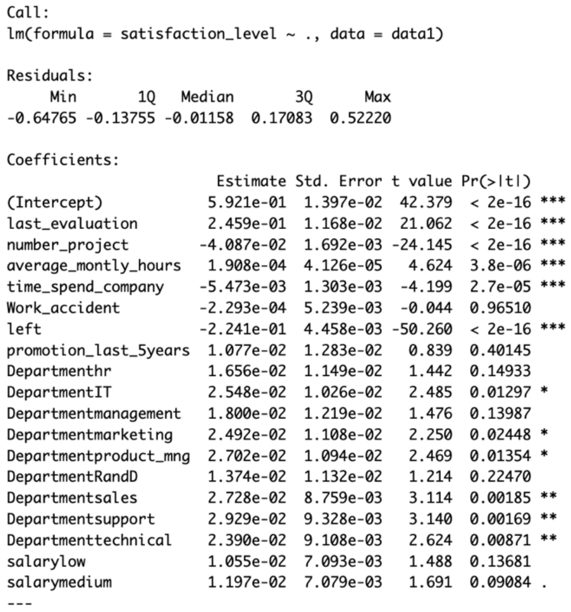
Fig 3.2

<u>Inference for Diagnostic plot:</u>

The figures generated in the analysis provide insights into the distribution and behavior of the dataset. In Figure 3.1, the QQ plot reveals a deviation in the right tail, indicating that the distribution is not entirely normal. On the other hand, in Figure 3.2, the residual plot suggests that there is a little lack of constant variance, although this behavior can be attributed to the large size of the dataset.

<u>Multilinear Regression Model:</u>

MLR (Multiple Linear Regression) is a statistical technique used to model the relationship between two or more predictor variables and a response variable. It is commonly used in machine learning to make predictions or analyze the relationship between variables. MLR assumes a linear relationship between the predictors and the response and aims to find the best-fit line through the data to make accurate predictions. It is a widely used and versatile technique in statistical analysis and data modeling.



```
Call:
lm(formula = satisfaction_level ~ ., data = data1)

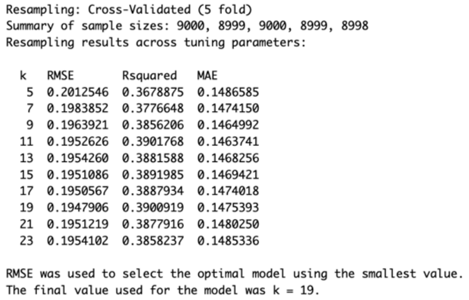
Residuals:
     Min       1Q   Median       3Q      Max 
-0.64765 -0.13755 -0.01158  0.17083  0.52220 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            5.921e-01  1.397e-02  42.379  < 2e-16 ***
last_evaluation        2.459e-01  1.168e-02  21.062  < 2e-16 ***
number_project        -4.087e-02  1.692e-03 -24.145  < 2e-16 ***
average_montly_hours   1.908e-04  4.126e-05   4.624  3.8e-06 ***
time_spend_company    -5.473e-03  1.303e-03  -4.199  2.7e-05 ***
Work_accident         -2.293e-04  5.239e-03  -0.044  0.96510    
left                  -2.241e-01  4.458e-03 -50.260  < 2e-16 ***
promotion_last_5years  1.077e-02  1.283e-02   0.839  0.40145    
Departmenthr           1.656e-02  1.149e-02   1.442  0.14933    
DepartmentIT           2.548e-02  1.026e-02   2.485  0.01297 *  
Departmentmanagement   1.800e-02  1.219e-02   1.476  0.13987    
Departmentmarketing    2.492e-02  1.108e-02   2.250  0.02448 *  
Departmentproduct_mng  2.702e-02  1.094e-02   2.469  0.01354 *  
DepartmentRandD        1.374e-02  1.132e-02   1.214  0.22470    
Departmentsales        2.728e-02  8.759e-03   3.114  0.00185 ** 
Departmentsupport      2.929e-02  9.328e-03   3.140  0.00169 ** 
Departmenttechnical    2.390e-02  9.108e-03   2.624  0.00871 ** 
salarylow              1.055e-02  7.093e-03   1.488  0.13681    
salarymedium           1.197e-02  7.079e-03   1.691  0.09084 .  
---
```

Inference for Multilinear Regression Model:

Since we have a regression model, we cannot rely solely on the adjusted R-squared value. Instead, we are also calculating the RMSE value. We obtained an RMSE value of 0.2207, which means that the root mean square of the differences between the actual and predicted values is 0.2207.

KNN MODEL:



```
Resampling: Cross-Validated (5 fold) 
Summary of sample sizes: 9000, 8999, 9000, 8999, 8998 
Resampling results across tuning parameters:

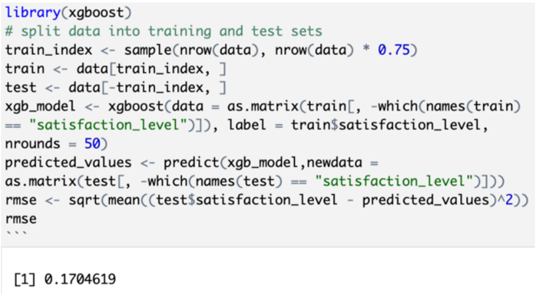
  k   RMSE       Rsquared   MAE      
   5  0.2012546  0.3678875  0.1486585
   7  0.1983852  0.3776648  0.1474150
   9  0.1963921  0.3856206  0.1464992
  11  0.1952626  0.3901768  0.1463741
  13  0.1954260  0.3881588  0.1468256
  15  0.1951086  0.3891985  0.1469421
  17  0.1950567  0.3887934  0.1474018
  19  0.1947906  0.3900919  0.1475393
  21  0.1951219  0.3877916  0.1480250
  23  0.1954102  0.3858237  0.1485336

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was k = 19.
```

The KNN model is a machine learning algorithm that uses the k-nearest neighbors' approach to predict the values of new data points. In this case, the KNN model was tuned with a length of 19, which means that the model was trained with a sample size of 19. After training the model, an RMSE (Root Mean Squared Error) value of 0.1910 was obtained, which represents the average difference between the predicted values and the actual values of the target variable. This RMSE value can be used to evaluate the performance of the KNN model, with lower values indicating better accuracy.

XGBOOST MODEL:

This paper employs the XGBoost algorithm to identify the factors contributing to employee attrition by analyzing basic data statistically and developing a performance evaluation model. XGBoost is a machine learning technique that is often used to address regression and classification problems. It employs an ensemble learning approach that combines decision trees and is primarily trained using bagging and bootstrapping methods to handle complex issues.
The algorithm uses a combination of regularization techniques, such as L1 and L2 regularization and dropout, to prevent overfitting and improve generalization performance.
XGBoost is known for its efficiency and speed due to its use of parallel processing and the ability to cache data in memory. This makes it well-suited for large-scale datasets with high-dimensional features. Additionally, XGBoost has several hyperparameters that can be tuned to optimize the model's performance for a specific problem.



~~~~
library(xgboost)
# split data into training and test sets
train_index <- sample(nrow(data), nrow(data) * 0.75)
train <- data[train_index, ]
test <- data[-train_index, ]
xgb_model <- xgboost(data = as.matrix(train[, -which(names(train)
== "satisfaction_level")]), label = train$satisfaction_level,
nrounds = 50)
predicted_values <- predict(xgb_model,newdata =
as.matrix(test[, -which(names(test) == "satisfaction_level")]))
rmse <- sqrt(mean((test$satisfaction_level - predicted_values)^2))
rmse
```

 [1] 0.1704619
~~~~

We can see that the RMSE value of XGBoost model is 0.1704.

4.Conclusion: We are comparing the RMSE values of the models in order to know the best fit model.

Model	MLR	KNN	XGBoost
RMSE	0.2207	0.1910	0.1704

The table shows the RMSE values for their respective models, and we know that the model with the lowest RMSE value is the most accurate. Among the models considered, XGBoost showed the lowest RMSE value, making it the best fit for the data. Thus, we can conclude that XGBoost is the best algorithm for this model. XGBoost is a powerful and reliable tool for regression problems, and its effectiveness in producing low RMSE values makes it a strong candidate for any application that requires high accuracy.

As the IT industry continues to expand rapidly, employee turnover has become a significant challenge. Replacing departing employees and training new ones is a time-consuming and stressful task. To address this issue, there is a growing need to develop an early warning system to detect potential employee departures. This research proposes a cost-effective XGBoost architecture for identifying employees who may leave. The proposed approach has been shown to outperform existing employee departure detection methods through rigorous experimentation.

References:

- Khaled Alshehhi, Safeya Bin Zawbaa "Employee retention prediction in corporate organizations using machine learning methods" Oct 2019.
- Ggaliwango Marvin, Majwega Jackson, MdGolam Rabiul Alam "A Machine Learning Approach for Employee Retention Prediction" Aug 2021.
- Tanmay Prakash Salunkhe "Improving Employee Retention by Predicting Employee Attrition using Machine Learning Techniques" 18/08/2018.

Code Appendix: https://github.com/vishwa1924/stat_632_final_project_5