

Predictive Analytics for Employee Attrition: A Machine Learning and Survival Analysis Approach

Priyanshi Sunil Saparia¹

¹Student, Dept. of Computer Engineering, Vidyavardhini College of Engineering and Technology, Mumbai, Maharashtra, India

Abstract - Employee attrition continues to be one of the most pressing operational issues in modern organizations, leading to great financial losses as organizations have to spend money on recruiting new staff and training them. The current paper proposes an evidence-based approach towards identifying potential attrition factors and estimating employee turnover with the help of data analytics based on a database provided by IBM containing information about 1,470 employees described in 35 variables. Specifically, Random Forest algorithm was used to predict whether employees will leave their positions or not using ROC-AUC score, while Kaplan-Meier analysis helped to visualize differences in attrition among different departments within the company. Cox Proportional Hazards regression was implemented in order to calculate the effect of independent variables on time-to-event in question. It was established that overtime, level of satisfaction with the position, average salary per month, and proximity to a workplace were among the key factors determining the likelihood of employee attrition. Using those variables identified as attrition factors, the cost-benefit analysis was conducted in order to assess the ROI of implementing a retention strategy.

Key Words: Employee Attrition, People Analytics, Random Forest, Survival Analysis, Kaplan Meier, Cox Regression, Human Resource Analytics, Predictive Modelling, Workforce Retention, Cost-Benefit Analysis

1. INTRODUCTION

Employee turnover is an ever-present concern in businesses worldwide, leading to huge financial losses for companies. Studies confirm that replacement costs associated with one employee vary from 50% to 200% of their annual income, which includes expenditures on recruitment, training, and lost productivity over the period of transition [1]. The loss caused by attrition reaches many crore per annum for big corporations having several hundred employees.

The development of people analytics has helped HR departments overcome their limited capabilities to deal. Studies using statistical methods in modeling attrition behavior have traditionally used logistic regression and survival analysis. In an empirical study using

with attrition by offering ways to prevent it using machine learning algorithms to recognize the signs of potential resignations and undertake preventive actions.

The research focuses on the following hypotheses. First, what are the strongest predictors of employee attrition? Second, how do employee survival chances differ between departments depending on their seniority in an organization? Third, what is the return on investment (ROI) on employee retention activities?

The rest of the paper is structured as follows. Section 2 is devoted to literature review. In section 3, we describe data and methods used. Results are presented in section 4, followed by policy recommendations in section 5. Section 6 contains conclusions and recommendations for future work.

1.1 Motivation

While there is data related to HR available in most firms, it is common practice for attrition-related decisions to be based on intuition instead of analytics. It is clear that through the use of publicly available data and the built-in functionalities in Python, an entire analytics pipeline can be created, making such a process feasible even for SMEs that do not have data scientists working for them..

2. LITERATURE REVIEW

Previous studies in employee turnover can be classified into three categories: theories of turnover, statistical methods, and machine learning.

One of the pioneering theories on voluntary turnover was put forth by Mobley (1977). According to his model, voluntary turnover results from a cognitive process beginning with job dissatisfaction, followed by resignation [2]. The theory has been instrumental in motivating future studies to examine factors related to satisfaction, such as job engagement, work-life balance, and remuneration.

Cox regression to estimate the hazards ratios of employee retention in terms of position and pay grade, Taplin & Winterton (2007) found that both factors significantly

predicted employee turnover [3]. Survival analysis has been advantageous in human resource management research due to its ability to handle censored cases—employees who have not yet left the organization when the data is collected.

However, the use of machine learning algorithms for HR analytics has seen a significant increase in the last decade. For example, Alao and Adeyemo (2013) tested decision tree and naive Bayes algorithms for employee attrition and found that ensemble models outperformed in dealing with the skewed nature of attrition data sets [4]. In a more recent study, Fallucchi et al. (2020) used Random Forest and gradient boosting algorithms on the IBM HR data set and obtained ROC-AUC scores higher than 0.80, where overtime hours and job level were identified as key features [5].

On the other hand, the connection between machine learning algorithm results and the business cost model has not been well-explored in previous studies. This research aims to address this gap by incorporating predictive model results into a well-formulated cost-benefit model.

3. METHODOLOGY

3.1 Dataset

The IBM HR Analytics Employee Attrition and Performance dataset which is publicly accessible on Kaggle was utilized in this project. The dataset consists of employee data comprising 1,470 observations with 35 variables which include demographic features (age, gender, marital status), employment features (department, job role, job level), compensation features (monthly income, stock option level), and job satisfaction features (job satisfaction, work-life balance, environmental satisfaction). Target variable of the study is binomial attrition (Yes/No) having class distribution of about 16.1% attrition and 83.9% retention. There are no missing values within the dataset for any variable.

3.2 Exploratory Data Analysis

Data analysis started with calculations of attrition ratios based on department, overtime working status, and job satisfaction levels. Correlation was calculated based on all numeric variables against binary target attrition variable. This was to assess the presence of any linear relationship. Positive correlations show association with higher likelihood of attrition.

3.4 Survival Analysis

Tenure of employees (YearsAtCompany) was considered the time-dependent variable while employee attrition is an event. Survival analysis techniques were implemented using non-parametric estimator - the Kaplan-Meier method. Non-parametric survival curve was constructed

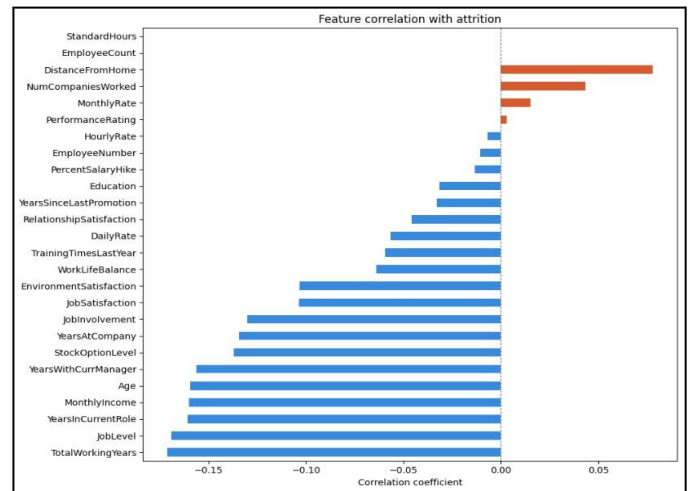


Figure 1 : Feature correlation with attrition

3.3 Classification Model

For predicting attrition rate, a Random Forest classifier was trained. Prior to training, categorical features were encoded using label encoding technique. In order to solve the problem of imbalance in classes, class_weight was set equal to "balanced", thus, sample weights were determined inversely proportional to class frequency without performing oversampling. Data were split into 80 percent training and 20 percent testing sets based on stratified random sampling with the same seed. Model performance was assessed using ROC-AUC score, classification report, and confusion matrix.

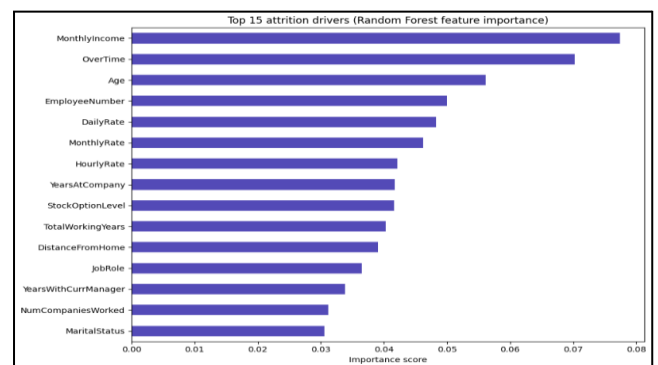


Figure 2 Top 15 attrition driver

for each department to demonstrate the probability of remaining employed in the company with respect to time. Log-rank test was conducted to test the null hypothesis that no difference exists between survival functions. Cox proportional hazards regression model was developed to determine the impact of multiple covariates on hazard

rate.

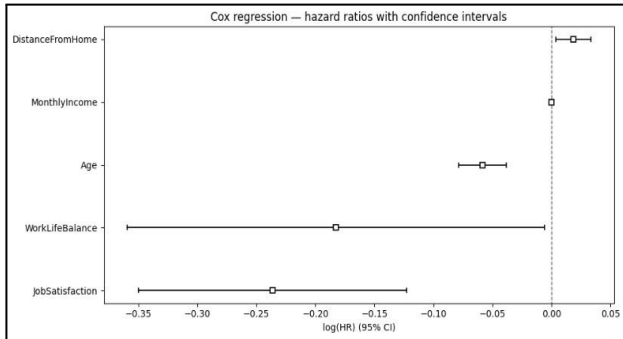


Figure 3: Cox regression — hazard ratios with confidence intervals

Distance and overtime had the most significant positive correlations with attrition.

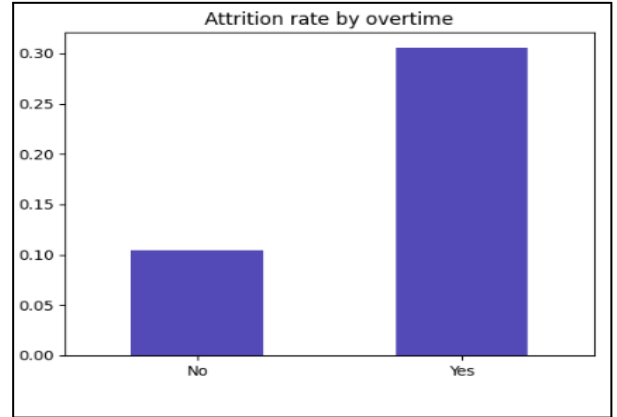


Figure 4 : Attrition Rate by department

3.5 Cost-Benefit Model

Attrition Cost Estimation:

The financial effects of attrition were analyzed using the cost model approach. The cost associated with filling the vacant position due to attrition is assumed to be 75% of annual salary as recommended by studies [6]. Annual attrition cost = (Number of departing employees) * (Average salary) * (Replacement cost). The expected saving due to 25% reduction in attrition, which can be obtained through specific retention strategies identified from the model's results, was estimated and compared against the cost of implementation to determine the ROI.

4. RESULTS

4.1 Exploratory Findings

A study on attrition rate by departments showed notable differences between departments. Employees who were required to do overtime had a much higher attrition rate than other employees, which is consistent with the assumption that workload was the major reason for employees leaving the job voluntarily. Job satisfaction proved to be negatively correlated with attrition, and employees with lower levels of job satisfaction had much higher attrition rates.

The study found that monthly income, work tenure, and job status had the most significant negative correlation with attrition; therefore, employees who earned more money, had many more years of experience, and worked in senior positions left the company at a smaller rate than others.

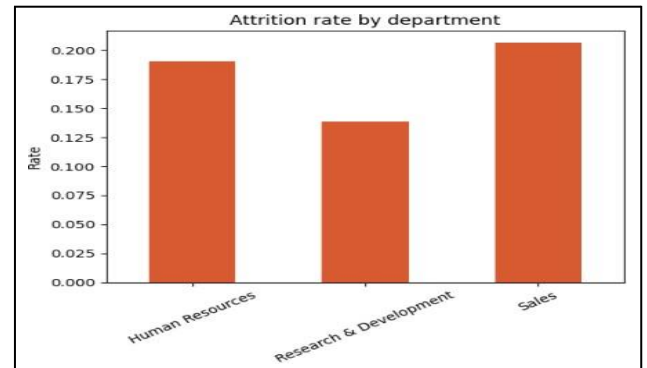


Figure 5 : Attrition Rate by overtime

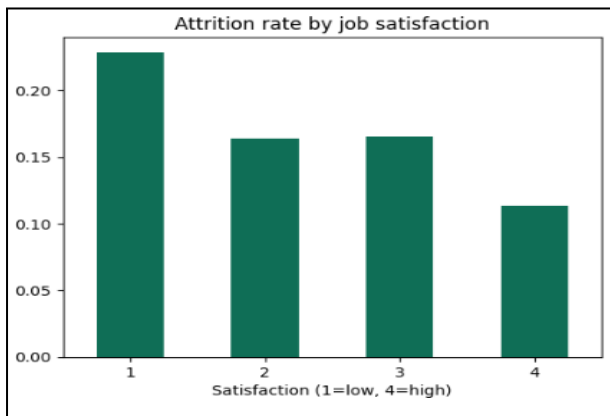


Figure 6 : Attrition Rate by Job satisfaction

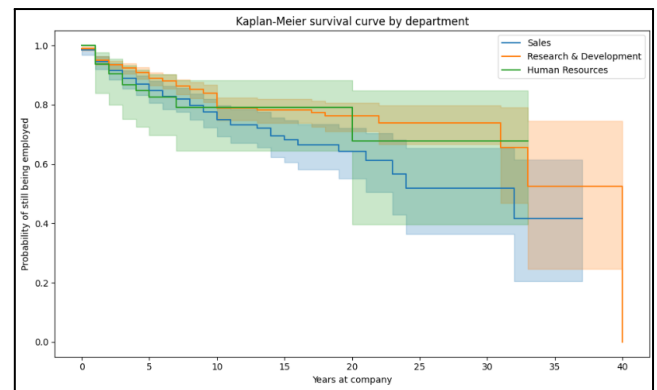


Figure 7: Kaplan-Meier survival curve by department

4.2 Classification Model Performance

In this context, the ROC-AUC for the Random Forest classifier was 0.763, which suggests that this model has high discrimination capacity to distinguish between employees who are at risk of leaving and those who are not. From the feature importance plot, the three important attrition factors were:

1. MonthlyIncome (importance = 0.077)
2. OverTime (importance = 0.070)
3. Age (importance = 0.056)

These results are supported by existing literature where remuneration and work overload have been found to be the primary predictors of attrition.

4.3 Survival Analysis Results

In another aspect, the Kaplan-Meier survival plot indicates that the probability of survival (employee staying within the organization) is sharply declining in the first three years of tenure regardless of department. Department-specific survival functions suggest significant differences, with the Sales department experiencing the sharpest drop-off in survival probability. Cox regression reveals that the variable with the maximum hazard ratio is JobSatisfaction (hazard ratio = -0.236), implying that every increment in job satisfaction significantly decreases the hazard of attrition.

4.4 Cost-Benefit Analysis

On the basis of the organizational variables modeled in this research paper, the annual cost incurred by employee turnover is estimated to be Rs. 10,62,00,000. With a retention program focused on the major factors driving turnover, which includes managing overtime, compensation, and job satisfaction, turnover can be decreased by 25 percent, thereby saving an additional 59 employees per year. The gross savings from the above intervention would be Rs. 2,65,50,000, whereas the cost of implementation would be Rs. 15,00,000, providing a net profit of Rs. 2,50,50,000.

Table 1: Cost-Benefit Summary

Parameter	Value
Number of Employees	1470
Attrition Rate at Present	16.10%
Number of Employees Who Leave	336
Attrition Cost per Employee	75% of Yearly Pay
Total Annual Attrition Cost	Rs. 10,62,00,000
Expected Attrition Reduction	25%
No. of Employees to be Saved	59
Gross Savings	Rs. 1,85,50,000

5. DISCUSSION

The research findings have numerous practical implications for HR policies. Three evidence-based recommendations are suggested based on the model outputs.

Firstly, monthly income was determined to be the most important predictor with a score of 0.077. Organizations are advised to engage in continuous monitoring of the salary trends within the company compared to market averages and develop clear salary advancement models. Furthermore, the survival analysis showed that the first three years of the working relationship pose the highest risk of departure and that attractive initial salaries are essential.

Secondly, overtime commitment was found to be the second most important variable (importance: 0.070). Companies should analyze those departments with high levels of overtime – in particular, the Sales department with the highest level of employee attrition (20.6%) – and establish some mechanisms for compensating the workers through more flexible work schedules, extra days off, or even transferring part of the responsibility to other employees.

Thirdly, job satisfaction was discovered to be the best individual indicator that prevents employees from leaving with a hazard ratio of -0.236. Relatively cheap solutions like regular performance feedback and rewarding systems can greatly increase employees' job satisfaction rates.

Even a conservative decrease in turnover rate by 25 percent yields a net ROI of Rs. 2,50,50,000, which is approximately seventeen times the investment required to implement the project, thus rendering the proactive infrastructure of people analytics economically sound.

6. CONCLUSIONS

This paper provides a comprehensive approach involving machine learning classification, survival analysis, and financial modeling to tackle the organizational challenge of attrition. The results of training the Random Forest classifier based on the IBM HR analytics dataset indicated a high degree of predictability in terms of a high ROC-AUC value. Furthermore, it has been shown that the problem of employee attrition is highly correlated with the first year of employment, being associated with overtime hours worked, pay, and job satisfaction. Finally, the financial model shows that retention initiatives have the potential to produce a positive ROI.

The most significant limitation associated with this work is based on the fact that the employed dataset is not real but rather hypothetical and developed by IBM. In spite of the realistic nature of the dataset and its popularity in academic research, its application to solving the problem in real organizational settings needs to be further validated using proprietary HR datasets. Future research directions may involve investigating the performance of more advanced deep learning models such as LSTM networks and employing NLP-based sentiment analysis of exit interviews.

ACKNOWLEDGEMENT

The author thanks Vidyavardhini College of Engineering & Technology, Mumbai for offering an academic setting wherein this study was carried out. The IBM HR Analytics dataset used in this study is from Kaggle's open data repository.

REFERENCES

- [1] Society for Human Resource Management, *Retaining Talent: A Guide to Analyzing and Managing Employee Turnover*, SHRM Foundation, Alexandria, VA, 2008.
- [2] W. H. Mobley, "Intermediate linkages in the relationship between job satisfaction and employee turnover," *Journal of Applied Psychology*, vol. 62, no. 2, pp. 237-240, 1977.
- [3] I. Taplin and J. Winterton, "Understanding Labour Turnover in a Labour Intensive Industry," *Journal of Management Studies*, vol. 44, no. 7, pp. 1132-1150, 2007.
- [4] D. Alao and A. B. Adeyemo, "Analysing Employee Attrition Using Decision Tree Algorithms," *Computing, Information Systems, Development Informatics and Allied Research Journal*, vol. 4, no. 1, pp. 17-28, 2013.
- [5] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. W. De Luca, "Predicting Employee Attrition Using Machine Learning Techniques," *Computers*, vol. 9, no. 4, p. 86, 2020.