Understanding and Predicting Employee Turnover

Vinay Kiran Raju Meesaraganda
Pace University, New York , NY 10038, USA
VM84055N@pace.edu

Abstract

Employee turnover is a significant problem for businesses as it can cause a great deal of disruption and financial loss. Predicting and understanding the factors that contribute to employee turnover is important for businesses to address these issues and implement measures to retain valuable employees. This analysis used a dataset containing information about employee job satisfaction, and performance evaluations. Exploratory data analysis revealed several interesting patterns and correlations. The project implemented various machine learning algorithms to predict employee turnover, including logistic regression, random forest. The random forest algorithm had the highest accuracy of 96% and AUC score of 0.95.This project demonstrated the effectiveness of machine learning algorithms in predicting employee turnover and the value of analysing employee data to identify potential factors contributing to turnover. By implementing measures to address these factors, businesses can improve employee retention and reduce the costs and disruptions associated with employee turnover.

1.Introduction

Employee turnover is a critical challenge faced by organizations across various industries. The loss of experienced and skilled employees can result in significant costs related to recruitment, training, and productivity. Thus, it is essential for organizations to understand the factors that contribute to employee turnover and develop strategies to retain employees. In this paper, we present an analysis of employee turnover using a dataset from Kaggle . The dataset contains information on various factors that could potentially impact employee turnover, such as employee satisfaction, salary, work accidents, and performance evaluations. Our analysis focuses on exploring the relationship between these factors and employee turnover. Specifically, we use exploratory data analysis (EDA) techniques to gain insights into the dataset and build predictive models to identify the key factors that contribute to employee turnover.

The analysis presented in this paper is intended to provide valuable insights for organizations looking to reduce employee turnover. By identifying the key factors that contribute to turnover, organizations can develop targeted strategies to retain employees. Additionally, the methods and techniques presented in this paper can be applied to other datasets to gain insights into employee turnover and other human resource-related challenges.

The layout of this paper is organized as follows. Section 2 provides an overview of the dataset used in this analysis and the methods used to pre-process and prepare the data for analysis. Section 3 presents the results of our EDA, focusing on the relationships between different factors and employee turnover. In Section 4, we build predictive models using machine learning algorithms to identify the key factors that contribute to turnover. Finally, we discuss the implications of our findings and provide recommendations for organizations looking to reduce employee turnover in Section 5.

2.Dataset and Pre-Processing

The dataset used in our study is a human resource dataset containing information about employees who have left or are still working for a company. The dataset includes 14,999 observations with 10 features, such as employee satisfaction, number of projects, work accidents, promotion, salary, performance rating, department, and time spent at the company.

To prepare the data for analysis, we first inspected the dataset to identify any missing or erroneous data. We found that there were no missing data points in the dataset, but there were some outliers present in the data that could have a significant impact on the analysis. Therefore, we performed an outlier analysis on the dataset to detect any outliers present in the data.

Next, we analysed the distribution of the dataset to identify any patterns or trends in the data. We found that the dataset was skewed towards employees who were still working for the company. There were 11,428 observations of employees who were still working for the company, while only 3,571 observations represented employees who had left the company. We also analysed the distribution of the dataset by department and found that the sales department had the highest number of observations, followed by technical and support departments.

we converted categorical data into numerical data using Label Encoder. We have observed class Imbalance in the data set and handled the classes imbalance using smote technique after splitting the data into training and testing. After pre-processing the data, we performed exploratory data analysis (EDA) to gain insights into the data.

3.Exploratory Data Analysis

These analyses were carried out to identify patterns, trends, and relationships in the data that could provide insights into why employees left the company.

The analysis on of satisfaction versus evaluation. Three distinct clusters of employees who left the company were identified. The first cluster (Hard-working and Sad Employee) had satisfaction below 0.2 and evaluations greater than 0.75, which could indicate that employees who left the company were good workers but felt horrible at their job. The second cluster (Bad and Sad Employee) had satisfaction between about 0.35~0.45 and evaluations below 0.58, which could mean employees who "under-performed." The third cluster (Hard-working and Happy Employee) had satisfaction between 0.71.0 and evaluations were greater than 0.8, which could mean that employees left because they found another job opportunity.

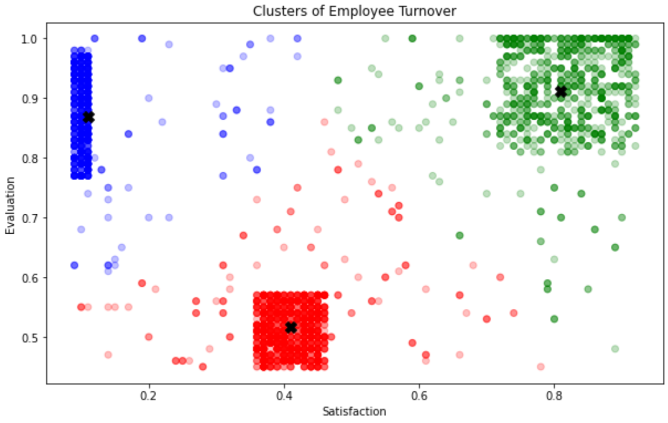


Figure 1. satisfaction versus evaluation

The analysis focused on employee satisfaction, and it was observed that there was a tri-modal distribution for employees that left the company. Employees who had low satisfaction levels (0.3-0.5) left the company more often than those with higher satisfaction levels.

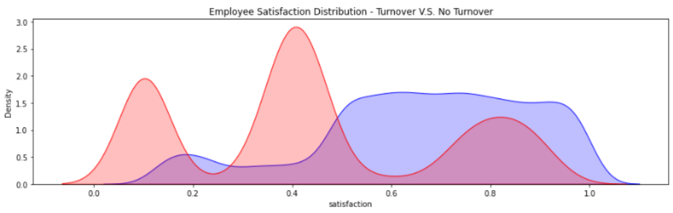


Figure 2. employee satisfaction levels

The EDA was on employee project count. It was observed that more than half of the employees with 2, 6, and 7 projects left the company. Majority of the employees who did not leave the company had 3, 4, and 5 projects. All the employees with 7 projects left the company, and there was an increase in employee turnover rate as project count increased.

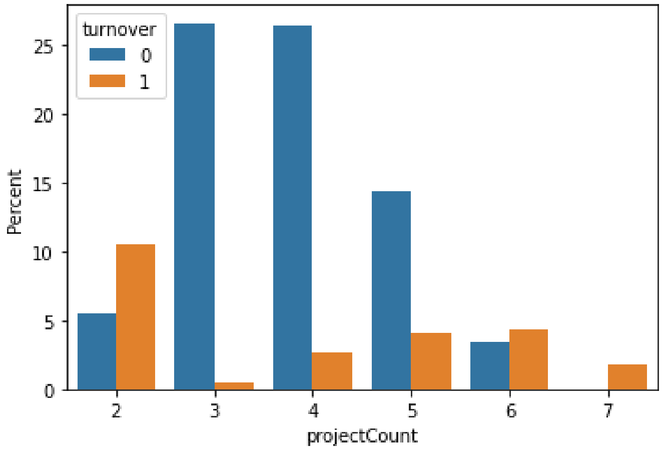


Figure 3.Employee project count Vs Turnover

while the other EDA examined the average monthly hours worked by employees. It was observed that employees who left the company generally worked either too few or too many hours.

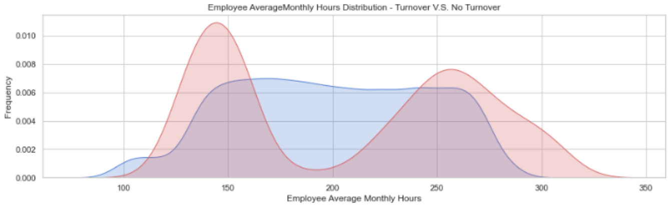


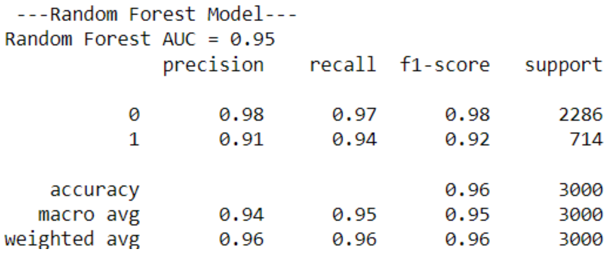
Figure 4. Average monthly hours Vs Turnover

The exploratory data analysis performed on the employee turnover dataset revealed several insights that could be valuable for understanding why employees left the company. These insights could be used to develop strategies for improving employee retention and satisfaction, such as offering more flexible working hours and reducing workloads for employees with high project counts.

4.Experimentation and Results

The objective of this section is to build a classification model that can accurately predict whether an employee is likely to leave the company or not. The dataset was split into training and testing sets using an 80-20 split. The following machine learning algorithms were evaluated for their performance in predicting employee turnover: Logistic regression and Random Forest. The performance is tested by running the models against test datasets to obtain Accuracy, AUC ROC score, precision, recall, F-1 Score.

The Random Forest classifier has achieved a highest accuracy score of 96% with a precision score of 98%, recall score of 98% , F-1 score of 0.97% and AUC-roc score of 95%.


```
 ---Random Forest Model---
Random Forest AUC = 0.95
              precision    recall  f1-score   support

           0       0.98      0.97      0.98      2286
           1       0.91      0.94      0.92       714

    accuracy                           0.96      3000
   macro avg       0.94      0.95      0.95      3000
weighted avg       0.96      0.96      0.96      3000
```

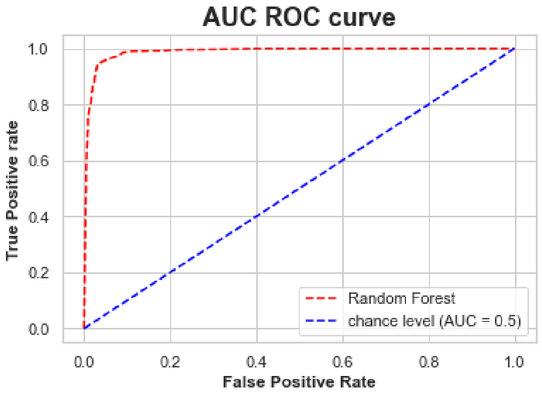


Figure 5. AUC ROC Curve

The feature importance was analysed using a random forest classifier. The results showed that project count, average monthly hours and evaluation were the three most important features for predicting employee turnover, followed by department, salary, work accident, and promotion.

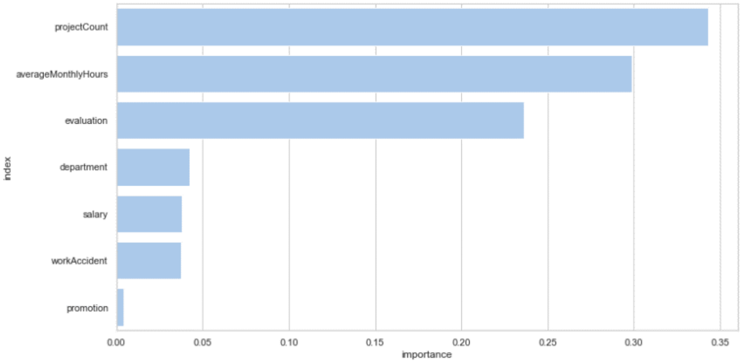


Figure 6. Feature Importance

5.Conclusion

In our employee retention problem, rather than simply predicting whether an employee will leave the company within a certain time frame, we would much rather have an estimate of the probability that he/she will leave the company. We would rank employees by their probability of leaving, then allocate a limited incentive budget to the highest probability instances.

Consider employee turnover domain where an employee is given treatment by Human Resources because they think the employee will leave the company within a month, but the employee does not. This is a false positive. This mistake could be expensive, inconvenient, and time consuming for both the Human Resources and employee but is a good investment for relational growth.

Compare this with the opposite error, where Human Resources does not give treatment/incentives to the employees, and they do leave. This is a false negative. This type of error is more detrimental because the company lost an employee, which could lead to great setbacks and more money to rehire. Depending on these errors, different costs are weighed based on the type of employee being treated. For example, if it’s a high-salary employee then would we need a costlier form of treatment? What if it’s a low-salary employee? The cost for each error is different and should be weighed accordingly.

The solution to this is we can rank employees by their probability of leaving, then allocate a limited incentive budget to the highest probability instances or we can allocate our incentive budget to the instances with the highest expected loss, for which we'll need the probability of turnover and develop learning programs for managers. Then use analytics to gauge their performance and measure progress.

References

- *T. Juvitayapun, "Employee Turnover Prediction: The impact of employee event features on interpretable machine learning methods," 2021 13th International Conference on Knowledge and Smart Technology (KST), Bangsaen, Chonburi, Thailand, 2021, pp. 181-185, doi: 10.1109/KST51265.2021.9415794.*
- *J. Yuan, "Research on Employee Turnover Prediction Based on Machine Learning Algorithms," 2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 2021, pp. 114-120, doi: 10.1109/ICAIBD51990.2021.9459098.*
- *D. M. Raza and F. Hasan, "Employee Engagement and Turnover utilizing Logistic Regression," 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), Dehradun, India, 2021, pp. 1-6, doi: 10.1109/UPCON52273.2021.9667566.*
- *Y. Duan, "Statistical Analysis and Prediction of Employee Turnover Propensity Based on Data Mining," 2022 International Conference on Big Data, Information and Computer Network (BDICN), Sanya, China, 2022, pp. 235-238, doi: 10.1109/BDICN55575.2022.00052.*
- *P. T. Burnes, "Voluntary employee turnover: why IT professionals leave," in IT Professional, vol. 8, no. 3, pp. 46-48, Jan.-Feb. 2006, doi: 10.1109/MITP.2006.78.*
- *H. Zhang, L. Xu, X. Cheng, K. Chao and X. Zhao, "Analysis and Prediction of Employee Turnover Characteristics based on Machine Learning," 2018 18th International Symposium on Communications and Information Technologies (ISCIT), Bangkok, Thailand, 2018, pp. 371-376, doi: 10.1109/ISCIT.2018.8587962.*
- *D. S. Sisodia, S. Vishwakarma and A. Pujahari, "Evaluation of machine learning models for employee churn prediction," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, India, 2017, pp. 1016-1020, doi: 10.1109/ICICI.2017.8365293.*
- *R. Chakraborty, K. Mridha, R. N. Shaw and A. Ghosh, "Study and Prediction Analysis of the Employee Turnover using Machine Learning Approaches," 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), Kuala Lumpur, Malaysia, 2021, pp. 1-6, doi: 10.1109/GUCON50781.2021.9573759.*