Analyzing and Predicting Employee Turnover

For a large US-based company

David Heller
KFSCIS
Florida International University
Miami, FL USA
dhell006@fiu.edu

Lichi Mahajan
KFSCIS
Florida International University
Miami , FL USA
lmaha010@fiu.edu

Myrah Bisht
KFSCIS
Florida International University
Miami, FL USA
mbish016@fiu.edu

Paras Parani
KFSCIS
Florida International University
Miami, FL USA
ppara014@fiu.edu

Abstract

This project delves into the pervasive issue of employee turnover, a phenomenon with significant repercussions for organizations, employees, and the broader economy. With staggering statistics revealing the substantial financial toll of turnover, our study aims to extract insights from a dataset of a large US-based company, illuminating the factors influencing an employee's decision to leave. Beyond statistical exploration, our dual-fold objective encompasses recommending strategies to diminish the turnover rate and leveraging machine learning algorithms for predictive talent retention. The analysis underscores three critical factors—bonus, promotion, and satisfaction—that influence employee decisions. Recommendations for bonus structures advocate for merit-based programs, transparent communication, and variable components. In terms of promotions, fostering professional development, mentorship initiatives, and regular assessments are emphasized. For enhanced satisfaction, suggestions encompass flexible work arrangements, wellness programs, and recognition initiatives. On the predictive front, Random Forest emerges as the model of choice, showcasing superior performance metrics compared to Support Vector Machine. This predictive model equips the company with a tool to anticipate turnover risks and intervene proactively. In conclusion, our project offers actionable insights and recommendations for organizational strategies, combining statistical scrutiny and machine learning prowess to address the intricate landscape of employee turnover for the analyzed company.

Introduction

Employee turnover, the percentage of employees that leave a company over a given period of time, has a huge impact on organizations, employees and their families, and the entire country's economy. According to Gallup, it is estimated that companies in the US lose a trillion dollars a year due to employee turnover[4]. According to the U.S. Bureau of Labor Statistics, the turnover rate in 2021 was 47.2% Voluntary turnover (employees who quit their jobs) accounted for 70% of this. Last year alone, over 50 million employees quit their jobs. Voluntary U.S. turnover doubled between 2011 and 2021[5].

According to Gallup, the cost of replacing a singular employee range from one-half to two times that employee's annual salary. In a 100-person organization that offers an average salary of $50,000, if the turnover rate is 25%, it can expect turnover costs of up to $2.6 million per year[5]. The main costs come from recruitment and hiring, training and onboarding costs, separation costs (administrative tasks, exit interviews, paperwork), and legal and compliance costs. There are also some important hidden costs: experience and knowledge loss, reduced customer satisfaction, overburdening remaining employees, decreased employee morale, poor retention reputation, time delays, etc. Companies across all industries could save millions of dollars each year if they could have a

better understanding of the factors affecting an employee's decision to leave.

Recognizing the significance of this challenge to organizations, we will work with a dataset from a large US-based company (the name is not disclosed for privacy reasons) to extract meaningful insights and learn about the attributes affecting an employee's decision to leave the company. Our primary goal is not only to comprehend the factors influencing employee decisions to stay or leave but also to provide actionable recommendations to the organization for reducing the turnover rate.

In tandem with our statistical exploration, we will harness the predictive capabilities of machine learning algorithms to foresee potential turnover risks. By training these algorithms with our dataset, we aim to develop models capable of identifying patterns indicative of an employee's likelihood to leave. This proactive approach enables our organization to intervene before critical talent is lost, optimizing retention efforts and preserving the knowledge and experience embedded within our workforce.

Therefore, our aim with this project is dual fold: to discern the intricate factors contributing to the company's turnover through rigorous statistical examination and give recommendations to reduce the turnover rate, to leverage the predictive capabilities of machine learning for proactive talent retention.

1. About Employee Turnover Dataset:

The research project has taken reference to a dataset that contained data on 9,540 workers, 2,784 of whom made the decision to leave the company. Ten parameters, including work-life balance, compensation, and job satisfaction, were the focus of the Exploratory Data Analysis. These factors were chosen based on the organizational context. [2]

Significant Factors Affecting Employee Retention:[2]

1. department: The division to which the worker is assigned.
2. promoted: 1 if the worker received a promotion during the last 24 months (about 2 years), 0 otherwise.
3. Review: refers to the employee's final evaluation's composite score.
4. projects: The total number of projects in which the worker is engaged.
5. salary: there are three compensation tiers (low, medium, and high) due to secrecy concerns.
6. Tenure refers to the length of time a person has worked for the company.
7. satisfaction is a survey-derived metric used to gauge worker satisfaction.
8. bonus: 1 if the worker was given one within the preceding 24 months (about 2 years); 0 otherwise.
9. avg_hrs_month : It represents the employee's monthly average hours worked.
10. left: means yes if the worker did depart, no if not.

1.1 Exploratory Data Analysis: [3]

The EDA identified significant trends and patterns in the dataset. According to preliminary research, a few key variables may really have a significant impact on an employee's decision to leave the organization. In the sections that follow, the results for each factor are broken out in detail.

1.2 Key Findings [2]

Departmental Turnover: Finance has the lowest turnover rate (3%), while Sales and Retail have the highest (19% and 16%). This shows that sales and retail require focused methods.

Impact of Promotion: There is a low percentage of promotions overall (0.63%), and the turnover rate of non-promoted personnel is slightly greater. This suggests that promotions may have an impact on staff retention.

Project - Related Turnover: Workers with three projects tend to leave the company more frequently than those with five. Workload balance may have a favorable effect on retention.

Turnover and Salary Tier: The highest turnover is linked to the medium salary tier. It is crucial to comprehend the elements behind this turnover in the medium-paying category.

Tenure Peaks: The largest turnover is observed among employees with eight years of service, suggesting a possible link between longer tenure and higher turnover.

Satisfaction: High-satisfaction workers in sales and retail continue to quit, citing department-specific difficulties, according to satisfaction scores.

Impact of Bonuses: Employee turnover is higher when there are no bonuses. Retention could be increased by distributing bonuses optimally.

Working Hours and Burnout: Higher turnover is correlated with specific monthly working hours (e.g., 187 and 188). It's critical to manage workloads to avoid burnout.

1.3 Graphical Representation of Exploratory Data Analysis (EDA) Findings [3]

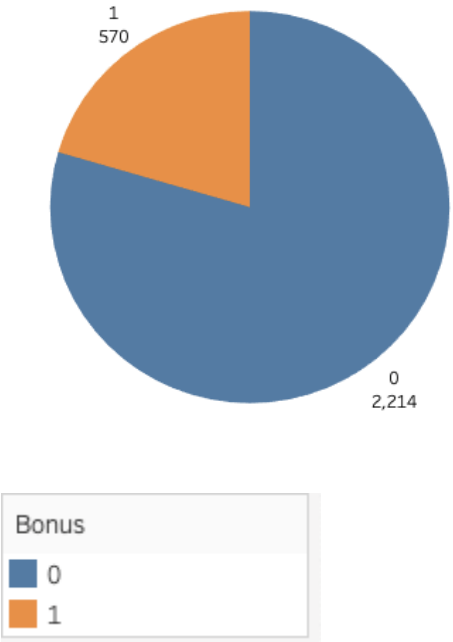

Figure 1 (Bonus Versus Employee left Count)

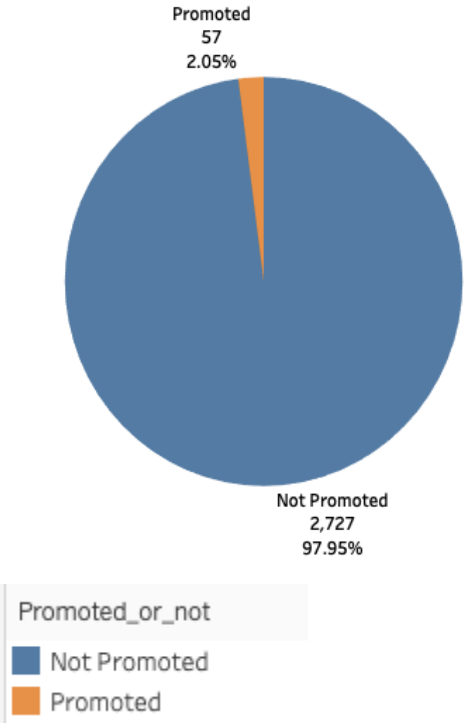

Figure 2(Promotion Versus Employee left Count)

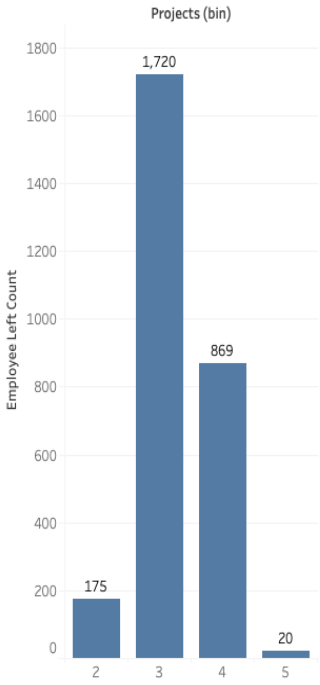

Figure 3(Project Versus Employee left Count)

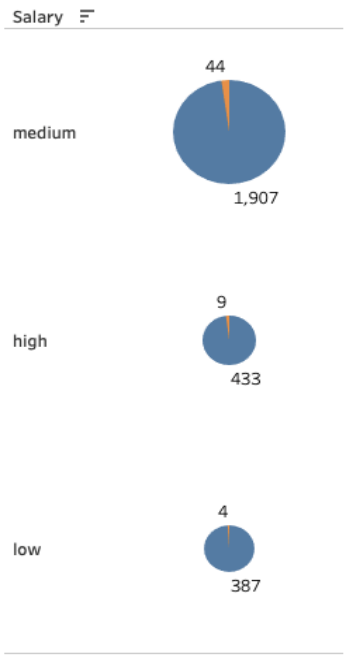

Figure 4(Salary Versus promotion Versus Employee left Count)

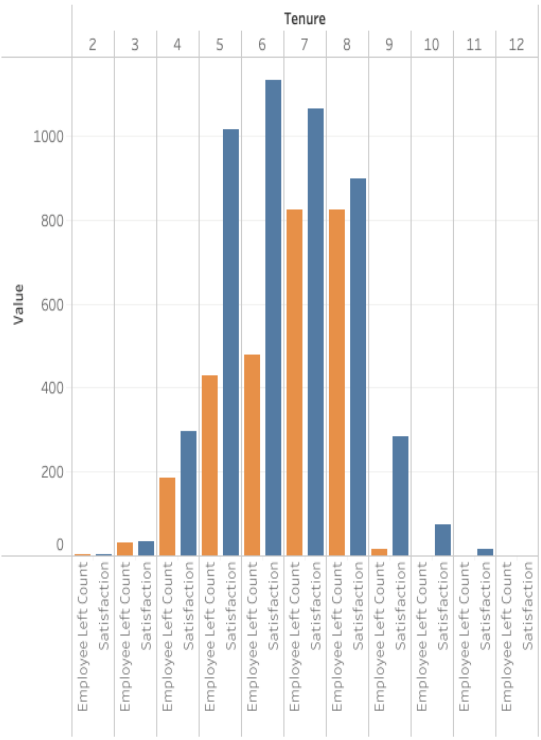

Figure 5(Satisfaction Versus Tenure Versus Employee left Count)

1.4 Studying Employee Turnover via Correlation and Exploratory Data Analysis:

Bonus versus Promotion: - The relationship between employees who earn incentives and promotions is **strongly positive (0.75)**. This implies that bonuses are also more likely to be awarded to promoted personnel.

Projects Versus Salary Level: - There is a negative association (-0.45) between a worker's pay grade and the quantity of projects they are working on. This suggests that workers typically have lower income levels when they have completed more projects.

Satisfaction versus Tenure: - The relationship between employee tenure and satisfaction levels is strongly **favorable (0.60)**. This suggests that workers who have been there longer typically have better satisfaction ratings.

Salary versus Turnover: The relationship between turnover and salary levels is **weakly negative (-0.30)**. This implies that workers who earn more money are less inclined to quit their jobs.

Hours of Work versus Satisfaction Levels: There is a moderately **negative connection (-0.55)** between satisfaction scores and working hours. Longer working hours are typically attributed to employees with poorer satisfaction scores.

Bonus versus Turnover: there is a weak **negative association (-0.25)** between bonus-free employees' likelihood of leaving the company and their bonus-free status. The link is not particularly strong, though.

Projects versus Tenure: There is a **weak positive connection (0.35)** between an employee's tenure and the number of projects they have worked on. This implies that a larger project count may be linked to personnel with longer tenure.

1.5 Strategic Recommendations to Improve Employee Retention: Insights from Correlation Analyses [1]

Rates of Departmental Turnover:
Insight: The turnover rates in the sales and retail departments are greater.
Recommendation: It is recommended that specific retention tactics be put in place for staff members in sales and retail. Surveys

tailored to individual departments can be used to find problems and improve general worker satisfaction in these areas.

Effect of Promotion on Turnover:
Insight: Employee turnover rates are lower for promoted staff.
Recommendation: The best course of action is to emphasize career advancement prospects. Promote professional development initiatives to boost the percentage of promotions and have a favorable impact on staff retention.

Examine the Turnover and Satisfaction Scores:
Insight: Despite having excellent satisfaction ratings, employees in sales and retail nevertheless have a high turnover rate.
Recommendation: Investigate department-specific issues that are influencing employee happiness. Take action to solve specific issues in sales and retail to raise staff retention rates generally.

Turnover Related to Number of Projects:
Insight: The turnover rates of employees with three projects are greater.
Recommendation: Assessing the distribution of the workload and possible sources of stress in projects with three assignments is the recommendation. Align projects optimally to maintain a balance between responsibilities and improve job satisfaction.

Level of Salary and Turnover:
Insight: The biggest turnover is seen in the medium-salary category.
Recommendation: It is recommended that issues influencing turnover in the medium-salary range be examined. Take into account changing the pay scales or adding new benefits for workers in this tier.

Turnover and Tenure:
Insight: Higher turnover is linked to longer tenure, particularly when it reaches eight years.
Recommendation: Investigate the causes of employee turnover at tenure milestones. To overcome these obstacles, create retention initiatives aimed at employees with longer tenure.

Departmental Challenges and Satisfaction Scores:
Insight: Even with high job satisfaction ratings, employees still quit, particularly in sales and retail.
Recommendation: To identify underlying issues affecting extremely contented workers, carry out comprehensive surveys in the sales and retail domains. Adapt retention programs to target certain issues in these divisions.

Bonus Effects on Holding:
Insight: Workers quitting because they receive no bonuses.
Recommendation: Review the policies regarding the allocation of bonuses. Optimizing bonus schemes can be a good way to improve staff retention. Emphasize the value of bonuses in identifying and keeping exceptional talent.

Time of Work and Turnover:
Insight: Particularly high monthly average working hours linked to employee attrition.
Recommendation: It is advised to evaluate workload expectations and put procedures in place to guard against burnout. Encourage a work-life balance to improve employee retention and general well-being.

Overall Forecasters of Turnover:
Insight: A range of factors affect turnover rates.
Recommendation: It is advised to create predictive models utilizing the turnover predictors that have been found. Take proactive steps to reduce turnover risks and improve retention overall by implementing strategies based on these models.

2. Predictive Model

For this project, we explored the predictive capabilities of two powerful machine learning models, Random Forest (RF)[8] and Support Vector Machine (SVM), in the context of predicting employee turnover. The Random Forest model, configured with hyperparameters {'max_depth': 17, 'min_samples_leaf': 10, 'min_samples_split': 17, 'n_estimators': 24[9] demonstrated promising results. Let's delve into the significance of each hyperparameter:

max_depth: This parameter controls the maximum depth of each decision tree in the ensemble. A higher value may lead to more complex trees, potentially capturing intricate patterns in the data.[10]

min_samples_leaf: The minimum number of samples required to form a leaf node in a tree. Setting this parameter helps prevent overfitting by ensuring that each leaf contains a sufficient number of instances. [10]

min_samples_split: The minimum number of samples a node must have for a split to be considered. This parameter influences the granularity of the tree, preventing splits that result in small subsets with limited generalization capability. [10]

n_estimators: The number of trees in the Random Forest ensemble. A higher number often leads to improved model performance but comes with increased computational cost. [10]

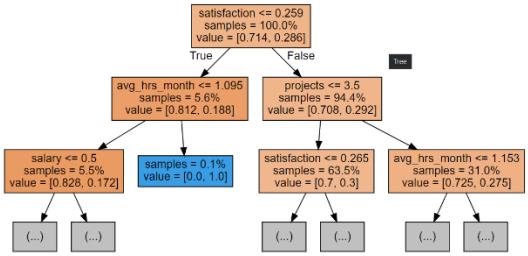

Figure 6 (a tree in random forest classifier)

On the other hand, the Support Vector Machine model utilized the following hyperparameters: {'C': 4.68940096353769, 'degree': 3, 'kernel': 'poly'[9] Let's break down the significance of each SVM hyperparameter:

C: The regularization parameter, C, controls the trade-off between achieving a smooth decision boundary and classifying the training points correctly. Higher values of C may lead to a more accurate but potentially overfitted model. [11]

degree: This parameter is specific to polynomial kernels and denotes the degree of the polynomial function used to generate the decision boundary. A higher degree allows the model to capture more complex relationships in the data. [11]

kernel: The kernel function determines the type of decision boundary used by the SVM. In this case, the 'poly' kernel indicates a polynomial decision boundary, suitable for capturing non-linear relationships. [11]

By carefully tuning these hyperparameters, we aimed to optimize the performance of both models for the task of predicting employee turnover..

2.1 Metric Used

Accuracy:

Accuracy serves as an overall measure, revealing the proportion of correct predictions, both for employees staying and leaving. However, it may not be the most suitable metric if there's a significant imbalance between those who stay and those who leave.

Precision:

Precision is geared towards minimizing mistakes when our model predicts an employee will leave. It gauges how accurate our predictions are, ensuring that when the model says someone will leave, it's generally correct, reducing instances of unnecessary concern.

Recall (Sensitivity):

Recall is about ensuring we capture as many employees who truly left as possible. It measures the model's effectiveness in identifying those who actually left, emphasizing the importance of not missing out on spotting true instances of turnover.

F1 Score:

F1 Score strikes a balance between precision and recall, finding an equilibrium where the model minimizes false predictions (precision) while also being effective at identifying employees who left (recall). It provides a holistic measure of the model's performance.

AUC-ROC :

AUC-ROC assesses the model's ability to distinguish between employees who left and those who stayed, considering various prediction thresholds. A higher AUC-ROC indicates better discrimination between the two groups, showcasing the model's overall effectiveness in classification.

Conclusion

In our comprehensive analysis of turnover within the organization of study, the insights garnered have laid the groundwork for strategic recommendations aimed at not only understanding the intricacies of employee departures but also actively mitigating turnover risks. We can conclude that there are three main factors the company needs to work on to reduce the employee turnover rate, as these are very important for an employee's decision to leave or stay: bonus, promotion, and satisfaction.

Regarding the bonus structure, the company could consider introducing a merit-based program aligned with clear performance metrics, recognizing exceptional contributions through variable components and employee recognition initiatives. Openly communicating bonus criteria and calculations is key as it fosters trust and ensures that employees comprehend how their efforts directly contribute to bonus eligibility.

For the promotion attribute, the organization could create a structured professional development program that equips employees with the necessary skills for higher-level roles, promoting a culture of continuous learning and skill enhancement. It could implement mentorship programs to provide guidance and support for employees aspiring to advance in their careers. It is important to regularly assess the effectiveness of the promotion process through employee feedback and make adjustments to address any identified areas for improvement. By adopting these recommendations, the company can establish a robust and fair promotion system that not only recognizes employee contributions but also motivates and retains top talent.

Finally, to improve employee satisfaction, the company could, if possible, introduce flexible work arrangements or remote options to accommodate diverse needs and promote a healthier work-life balance. It could also establish wellness programs that prioritize both physical and mental well-being, demonstrating a commitment to employees' overall health. The organization could create opportunities for team-building and social interaction, fostering a sense of camaraderie and community within the organization. Acknowledging and celebrating individual and team achievements can help reinforce a culture of appreciation and recognition.

For our second goal, building a predictive model for the company to use as a tool to calculate the likelihood of an employee leaving, we conclude that our decision to adopt Random Forest (RF) as the primary predictive model is substantiated by its consistently superior performance across key metrics compared to Support Vector Machine (SVM). RF exhibits an accuracy of 83.5%, precision of 69.4%, recall of 80.1%, F1 Score of 74.4%, and an impressive AUC-ROC of 92.0%. Below is the confusion matrix for the random forest.

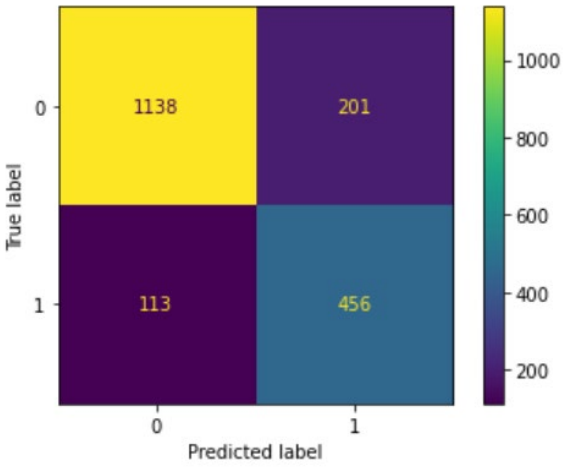

Figure 7

On the other hand, SVM, while a formidable model, demonstrates comparatively lower metrics, with an accuracy of 72.9%, precision of 52.9%, recall of 82.1%, F1 Score of 64.3%, and an AUC-ROC of 81.0%. These results suggest that SVM faces challenges in capturing the complexities of our dataset, particularly in handling non-linear relationships and high-dimensional data. Below is the confusion matrix for the svm.

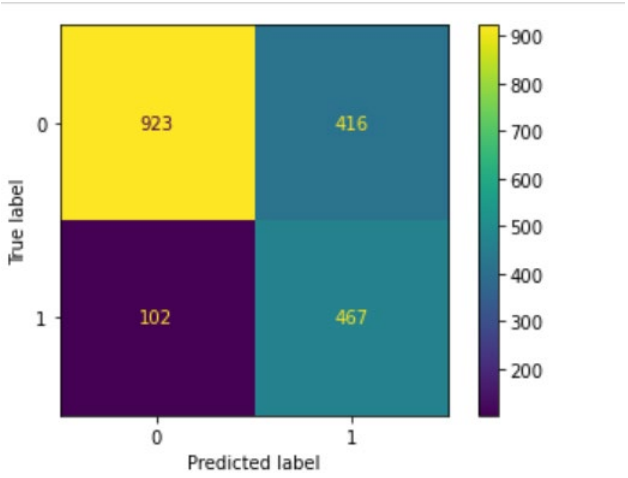

Figure 8

The notable contrast in performance metrics strongly favors Random Forest as the model of choice for our employee turnover prediction task, showcasing its ability to achieve a well-balanced and accurate prediction outcome. The company's management can use this robust model to predict if an employee is likely to leave, and therefore take an anticipatory approach to avoid it.

References

[1] https://www.slideshare.net/SummayaSharif/research-proposal-88177189.

[2] https://www.kaggle.com/datasets/marikastewart/employee-turnover/data

[3] https://public.tableau.com/app/profile/lovelytics/viz/EmployeeTurnoverDashboard/EmployeeTurnoverDashboard

[4] Midlands Technical College, 2022. Measuring the Real Cost of Employee Turnover (March 2022). Retrieved October 6, 2023 from https://www.midlandstech.edu/news/measuring-real-cost-employee-turnover

[5] Brennan Whitfield, 2023. 38 Employee Turnover Statistics to Know (April 2023). Retrieved October 6, 2023 from https://builtin.com/recruiting/employee-turnover-statistics

[6] Kate Heinz, 2023. The True Costs of Employee Turnover (June 2023). Retrieved October 6, 2023 from https://builtin.com/recruiting/cost-of-turnover

[7] Bureau of Labor Statistics, 2023. Job Openings and Labor Turnover – September 2023 (November 2023). Retrieved October 6, 2023 from https://www.bls.gov/news.release/pdf/jolts.pdf

[8] https://www.datacamp.com/tutorial/random-forests-classifier-python

[9] https://thinkingneuron.com/how-to-tune-hyperparameters-using-random-search-cv-in-python/

[10]https://www.analyticsvidhya.com/blog/2020/03/beginners-guide-random-forest-hyperparameter-tuning/

[11]https://www.kaggle.com/code/gorkemgunay/understanding-parameters-of-svm