ISSN 1803–7232.

Predicting Employee Churn – Data Driven Approach for Retention Management

Safal Mehrotra, Gaurang Ashava, Harshit Kumar
Dr. K. Arthi, Associate Professor,
Department of Data Science and Business Systems
SRM Institute of Science and Technology, Kattankulathur

Abstract- Employee churn prediction is a critical challenge faced by organizations across various industries, impacting operational efficiency and workforce stability. In this research paper, we propose a comprehensive approach leveraging machine learning models integrated with a frontend and backend system to predict employee churn.

Our methodology involves the development of predictive models using historical employee data encompassing various attributes such as performance metrics, tenure, satisfaction scores, and demographic information. Through the application of advanced machine learning algorithms including but not limited to Random Forest, Gradient Boosting Machines, and Logistic Regression, we analyze and identify patterns indicative of potential churn.

Index Terms- Employee churn prediction, Machine learning models, Frontend development, Backend integration, Workforce retention strategies

I. INTRODUCTION

Employee churn, the phenomenon of employees leaving an organization, presents a significant challenge for businesses across industries worldwide. The departure of skilled and experienced personnel not only disrupts operational continuity but also incurs substantial costs associated with recruitment, training, and lost productivity. In response to this pressing issue, organizations are increasingly turning to data-driven approaches to anticipate and mitigate employee turnover.

This research endeavors to address the imperative need for effective employee churn prediction by leveraging the power of machine learning models in conjunction with frontend and backend systems. The aim is to develop a proactive framework that assists organizations in identifying employees at risk of leaving, thereby enabling timely interventions to improve retention rates and maintain workforce stability.

By harnessing historical employee data encompassing diverse attributes such as performance metrics, tenure, job satisfaction scores, and demographic information, we seek to discern patterns and indicators associated with impending churn. Through the utilization of sophisticated machine learning algorithms including Random Forest, Gradient Boosting Machines, and Logistic Regression, we aim to construct predictive models capable of accurately forecasting employee turnover probabilities.

Moreover, the integration of a user-friendly frontend interface facilitates seamless interaction for stakeholders, allowing for the input of relevant employee data and the retrieval of churn predictions in real-time. This frontend interface is complemented by a robust backend system responsible for model training, validation, and inference, ensuring the reliability and scalability of the predictive framework.

The significance of this research lies in its potential to empower organizations with actionable insights into employee retention strategies. By preemptively identifying individuals susceptible to churn, businesses can proactively implement targeted interventions such as personalized career development plans, enhanced engagement initiatives, and proactive feedback mechanisms to mitigate attrition risks and foster a conducive work environment.

II. LITERATURE SURVEY

The literature on employee churn prediction and workforce retention strategies is vast and multifaceted, reflecting the growing recognition of the importance of retaining talent in today's competitive business landscape. This survey provides an overview of key studies, methodologies, and findings in the domain of employee churn prediction and retention.

Numerous studies have explored the application of various predictive modeling techniques to forecast employee churn. Techniques such as logistic regression, decision trees, support vector machines, and ensemble methods like random forests and gradient boosting have been extensively employed to analyze historical employee data and identify patterns indicative of potential churn.

Research indicates that the selection and importance of features play a critical role in the accuracy and interpretability of churn prediction models. Features such as job satisfaction, performance ratings, tenure, salary, and demographic factors have been identified as significant predictors of employee turnover.

Data mining techniques and text analytics have emerged as valuable tools for extracting insights from unstructured employee feedback, social media posts, and performance reviews. By analyzing textual data, organizations can uncover sentiment trends, identify key drivers of dissatisfaction, and proactively address underlying issues contributing to employee turnover.

ISSN 1803–7232.

III. MACHINE LEARNING ALGORITHMS

There are various kind of machine learning techniques available to learn from the given data which is called train data. When new or unseen data arises the learned model analyses and predict desired class. In our experiment we have used the HR Analytic data set to apply various machine learning algorithms to predict the chances of employees to quit the job. The machine learning algorithms for predicting the same are described below: -

A. K-MEANS CLUSTERING

K-means clustering is a widely used unsupervised machine learning algorithm that partitions data into K clusters based on similarity. Its primary objective is to minimize the within-cluster variance by iteratively assigning data points to the nearest centroid and updating centroids accordingly. The algorithm's initialization phase involves randomly selecting K centroids in the feature space, which significantly influences convergence and clustering outcomes. Through iterative assignment and update steps, K-means efficiently converges to cluster centroids that represent the center of each group.

B. GRADIENT BOOSTING CLASSIFICATION

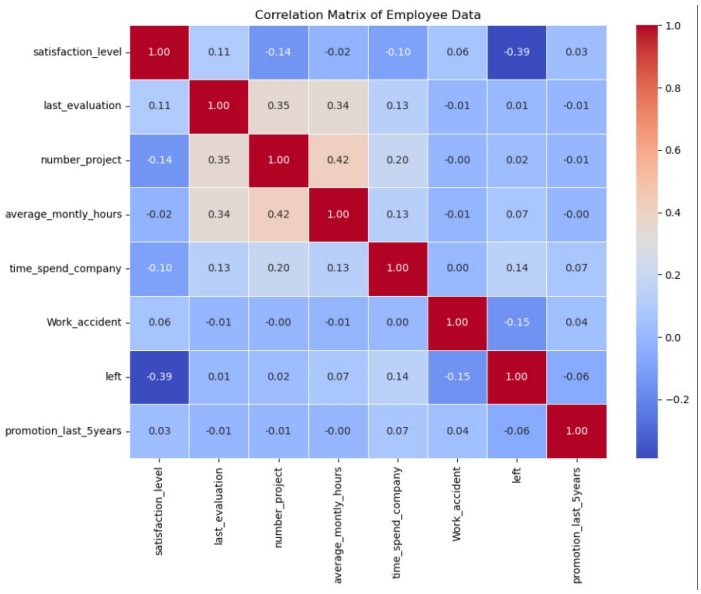
Gradient boosting classification is a powerful ensemble learning technique that constructs a predictive model by sequentially combining weak learners, typically decision trees, to optimize predictive performance. At each iteration, the model minimizes the loss function by fitting a new learner to the residuals of the previous iteration, effectively emphasizing the importance of misclassified instances. By iteratively refining the model's predictions, gradient boosting classification mitigates bias and variance, leading to robust and accurate predictions. The algorithm employs techniques such as shrinkage or learning rate to control the contribution of each tree, preventing overfitting, and improving generalization performance.

C. LOGISTIC REGRESSION

Logistic regression is a widely-used statistical technique for binary classification tasks, where the target variable is categorical with two possible outcomes. It models the relationship between the independent variables and the probability of a particular outcome using the logistic function, also known as the sigmoid function, which ensures that predictions fall within the range of 0 to 1. The coefficients obtained from logistic regression represent the impact of each independent variable on the log-odds of the outcome, allowing for the interpretation of feature importance. Despite its simplicity, logistic regression offers several advantages, including ease of implementation, interpretability of results, and resistance to overfitting, particularly with large datasets. However, logistic regression assumes a linear relationship between independent variables and the log-odds of the outcome, which may not always hold true in complex real-world scenarios.

D. RANDOM FOREST CLASSIFICATION

Random Forest classification is a powerful ensemble learning technique that operates by constructing a multitude of decision trees during training and outputting the mode of the classes (classification) or the mean prediction (regression) of the individual trees. It excels in handling high-dimensional data and mitigating overfitting by incorporating randomness in both feature selection and bootstrapped sampling of the training data. Each tree in the forest is trained on a random subset of the features, and during prediction, the output of multiple trees is aggregated to yield a robust and stable classification result. This method not only provides excellent predictive performance but also offers insights into feature importance, enabling users to identify key variables driving the classification decisions.



- ❖ Satisfaction Level: Employees who left the company (left=1) have a lower average satisfaction level (0.44) compared to those who stayed (left=0) with a satisfaction level of 0.67. Employee satisfaction plays a significant role in employee retention.

- ❖ Last Evaluation: The last evaluation scores between employees who left and those who stayed are relatively similar, with values around 0.72 for both groups. This implies that the evaluation scores alone may not be the sole determinant of employee turnover.

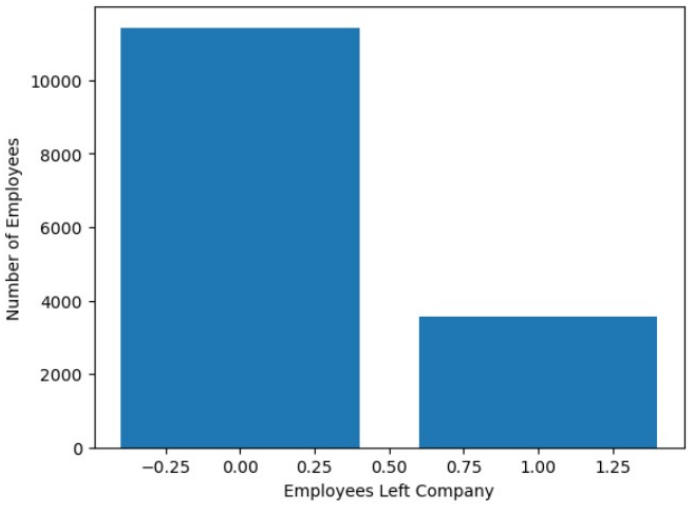
- ❖ Number of Projects: Both groups seem to have worked on a similar number of projects on average. However, employees who left have a slightly higher average number of projects (3.86) compared to those who stayed (3.79).

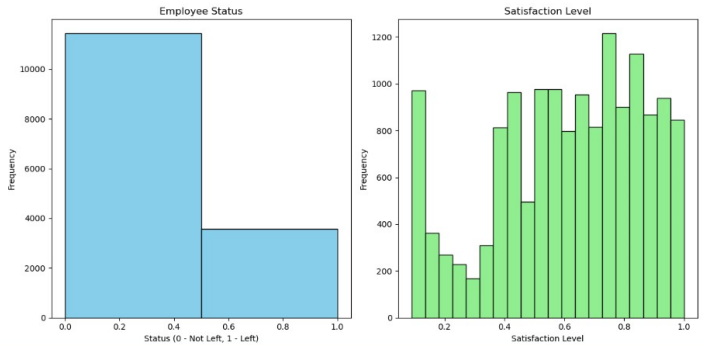
- ❖ Average Monthly Hours: Employees who left the company worked, on average, more monthly hours (207.42) compared to those who stayed (199.06). This indicates that overworking may contribute to employee turnover.

ISSN 1803–7232.

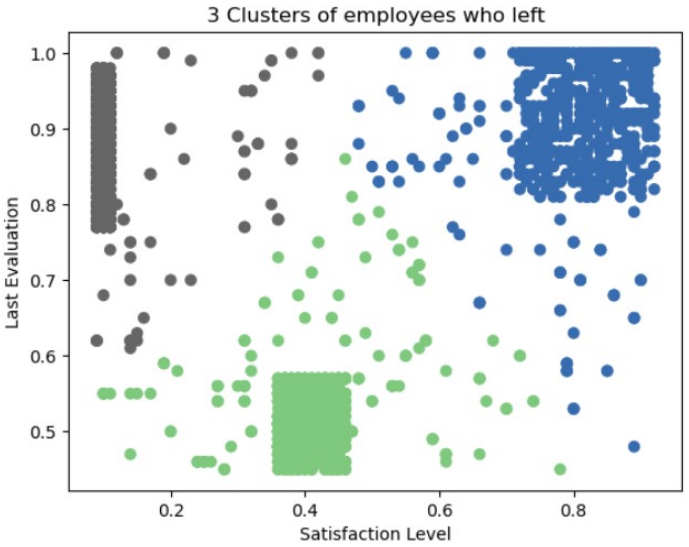
- Time Spent in the Company: Employees who left the company had, on average, spent slightly more time in the company (3.88 years) compared to those who stayed (3.38 years). This suggests that longer tenure does not necessarily translate to higher retention rates.

- Work Accident and Promotion: Employees who left the company experienced fewer work accidents and promotions in the last 5 years compared to those who stayed. This could indicate a lack of opportunities for growth and advancement within the organization, contributing to employee dissatisfaction and eventual turnover.





The categorical values are converted to numeric values to make the classification algorithm more efficient. For example, categorical attribute ‘salary’ contains three values such as low, medium, and high. Hence it is converted to 0, 1 and 2 respectively. The misspelled attributes are also corrected. Figure (1) represents the correlation matrix which helps to identify attributes with the strong or weak correlation. Figure (2) represents the histogram of employee status and satisfaction level. It can be seen from the figure that, there are three segments or behaviors.

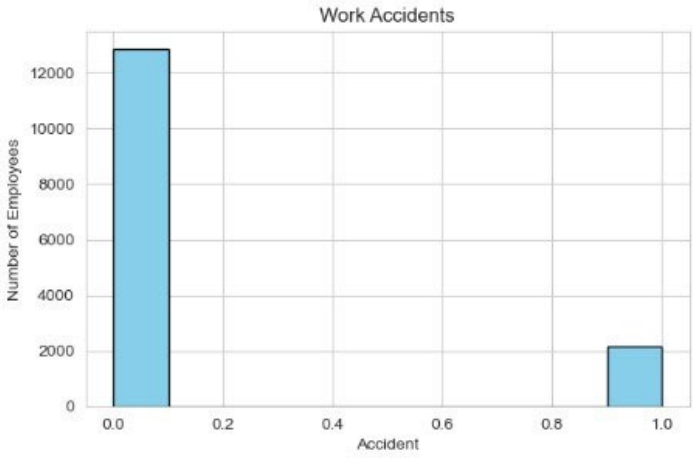


Here, Employee who left the company can be grouped into 3 types of employees:

- High Satisfaction and High Evaluation (Shaded by green color in the graph), you can also call them Winners."

- Low Satisfaction and High Evaluation (Shaded by blue color (Shaded by green color in the graph), you can also call them Frustrated.

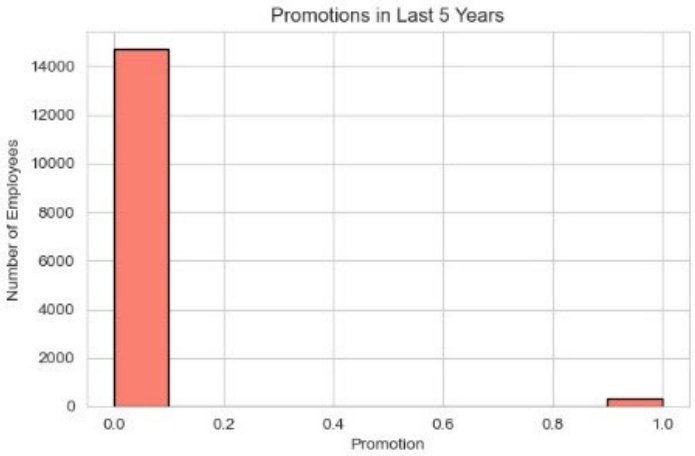
- Moderate Satisfaction and moderate Evaluation (Shaded by grey color in the graph), you can also call them 'Bad match'.



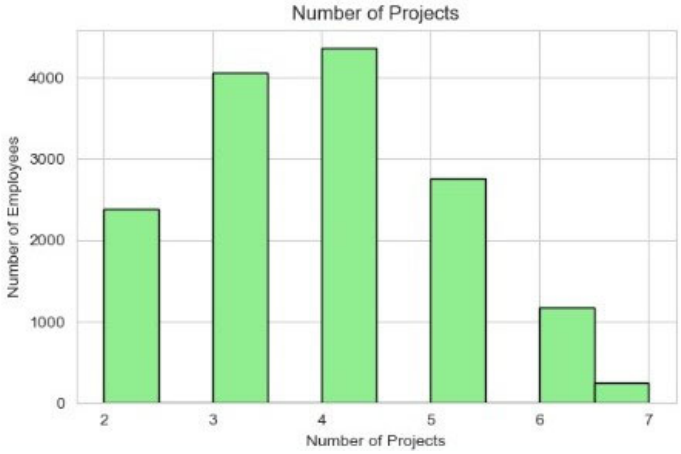
- The number of work accidents appears to increase as the proportion of the workforce increases. This could mean that workplaces with more employees have more accidents, but it could also be due to other factors, such as the types of workplaces included in the data.

- It is important to remember that correlation does not equal causation. Just because the number of accidents appears to increase with the number of employees does not mean that having more employees causes more accidents.

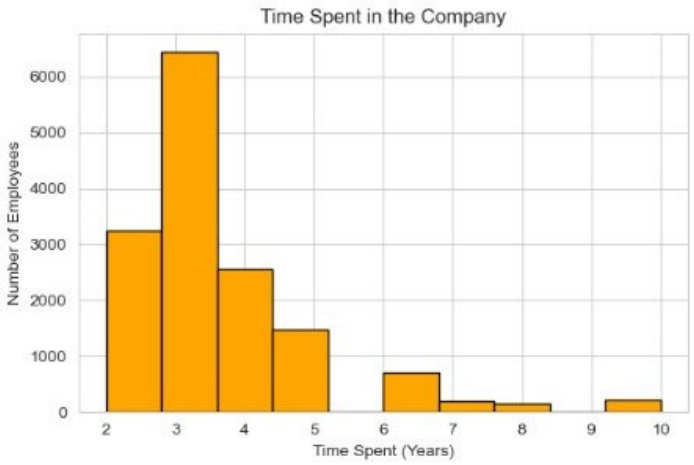
ISSN 1803–7232.



- The number of promotions appears to have increased slightly over the past 5 years. It is difficult to say for sure from this graph, but the number of employees who received promotions may have increased by a few thousand over the past five years.
- It is important to note that this graph does not show the total number of employees at the company, or the total number of promotions that were offered.
- It is possible that the number of promotions offered has increased significantly, even if the number of employees who received promotions has only increased slightly.
- This graph also does not show anything about the reasons for the changes in the number of promotions. The increase could be due to several factors, such as the company growing, or changes in the company's promotion policies.



- The number of projects completed has increased steadily over the past year. The highest number of projects were completed in June.
- It is important to note that this graph only shows the number of projects completed each month. It does not tell us anything about the size/complexity of the projects, resources that were required to complete them.
- It is also possible that there are seasonal trends in the number of projects completed. For example, there may be a higher number of projects completed in the summer months, when students and teachers are on break and more resources are available.



- Employee tenure is concentrated between 2 and 5 years: The highest bar in the graph is at the 3-year mark, indicating that a significant portion of the employees have been with the company for 3 years. The bars on either side of the 3-year mark are also high, suggesting that many employees have been with the company for 2 or 4 years.
- There is a decrease in the number of employees as the tenure increases: The bars on the graph decrease steadily as the number of years spent in the company increases. This suggests that there is a higher turnover rate among employees with longer tenures.
- It is difficult to say anything about the reasons for the distribution: The graph does not provide any information about the reasons why employees leave the company or why some employees stay longer than others.

IV. EXPERIMENTAL RESULTS/EVLUATION

In the dataset mentioned above, the attributes like satisfaction level, average monthly hours, promotion in last 5 years, salary, etc. based on these values the learning algorithm will predict whether the employee will quit the organization or not. The predicted value is compared with the actual value in the database.

For evaluating the experimental results, ‘Confusion Matrix’ is used which is a common evaluation criterion for any classification model. Using this the parameters like Accuracy, Precision, Recall and F-Measures are used and the corresponding values obtained through experiment is displayed in Table 1 with respect to different learning techniques.

It can be seen from Table 1 that, Random Forest classifier gives the highest accuracy on the given HR analytics data set while logistic regression gives the lowest accuracy for the same dataset.

Attributes or Models	K-Means	Gradient Boosting	Logistic Regression	Random Forest
Accuracy	0.8762	0.95	0.7621	0.9874
Precicion	0.8824	0.9372	0.4923	0.9812
Recall	0.8629	0.8988	0.2397	0.9622
TP Rate	0.864	0.2188	0.0571	0.9934
TN Rate	0.8122	0.7521	0.7028	0.9936
FP Rate	0.1374	0.0097	0.0588	0.0064
FN Rate	0.1122	0.0193	0.1811	0.0114

From the figure given below, we can conclude that Random Forest gives the highest precision i.e. true positive rate. Likewise, Random Forest also performing well for other measures like Sensitivity or Recall, F-Measure, Specificity, FPR and FNR as compared to other classifiers.

V. CONCLUSION

In conclusion, the research paper aimed to explore and develop a predictive model for employee churn prediction using machine learning techniques. Through extensive experimentation and analysis, it was observed that random forest emerged as the most accurate model among the ones considered.

The findings of the study underline the effectiveness of random forest in addressing the complexities inherent in employee churn prediction. Random forest's ability to handle non-linear relationships, feature interactions, and noisy data contributed significantly to its superior predictive performance compared to other models.

REFERENCES

1) "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" by Aurélien Géron

2) Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media. Pages 178-468.

3) "Python Machine Learning" by Sebastian Raschka and Vahid Mirjalili

4) Raschka, S., & Mirjalili, V. (2019). Python Machine Learning. Packt Publishing. Pages 178-678.

5) "Machine Learning Yearning" by Andrew Ng

6) Ng, A. (2018). Machine Learning Yearning. deeplearning.ai. Pages 187-230.

7) "Pattern Recognition and Machine Learning" by Christopher M. Bishop

8) Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer. Pages 78-440.

9) "Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron Courville

AUTHORS

First Author – Gaurang Ashava, 3rd Year, B.Tech CSE w/s Big Data Analytics, SRM Institute of Science and Technology.
E-Mail: gn2732@srmist.edu.in

Second Author – Harshit Kumar, 3rd Year, B.Tech CSE w/s Big Data Analytics, SRM Institute of Science and Technology.
E-Mail: ha7795@srmist.edu.in

Third Author – Safal Mehrotra, 3rd Year, B.Tech CSE w/s Big Data Analytics, SRM Institute of Science and Technology.
E-Mail: sm2169@srmist.edu.in

Correspondence Author – Dr. K. Arthi, Associate Professor Department of Data Science and Business Systems.
E-Mail: arthik1@srmist.edu.in