EMPLOYEE CHURN PREDICTION

Kola Sai Chakradhar
skola01@syr.edu
Syracuse, USA

Sai Karthik Kosuri
skosuri@syr.edu
Syracuse, USA

Rachana Munagla
rmunugal@syr.edu
Syracuse, USA

Abstract - A successful organization's workforce is its most important asset, and employee engagement and retention are critical. The phenomena of workers quitting their employment, or employee churn, is a complicated topic that reflects deeper problems with workplace settings, management styles, and employee satisfaction. This study uses a broad dataset that combines operational, performance-related, and emotional characteristics to explore the human factors influencing employee turnover. Using sophisticated machine learning methods like Random Forest classifiers and Gradient Boosting, our investigation reveals the complex narratives surrounding employee exits. We find that work-life balance, job happiness, and recognition have a big impact on churn, especially for technical workers who feel underappreciated in performance reviews or have heavy workloads. This study uses predictive analytics and visual storytelling to show not just who is going but also why. The results advocate for an organizational culture that values its people, recognizes their needs, and creates a space where each team feels heard, recognized, and valued. They act as a lighthouse for organizational change. In addition to exploring employee churn, this study serves as a call to action for companies looking to improve retention tactics and reignite employee relationships to strengthen organizational culture and secure a more stable future.

Keywords- *Gradient Boosting, Random Forest, Data Visualization, Employee Engagement, Job Satisfaction, Work-Life Balance, Employee Retention, Workforce Analytics, Predictive Modeling, Machine Learning, and Employee Churn.*

I. INTRODUCTION

In the ever-changing business environment of today, keeping great talent is just as important as bringing it in. Turnover, often known as employee churn, is a major problem for businesses of all kinds since it can result in higher operating expenses, less output, and lower morale among staff members. Thus, it has become strategically necessary to comprehend and anticipate employee churn to ensure sustained business growth. The pace at which current employees leave a company and are replaced by new hires is known as employee churn. Elevated turnover rates may indicate more serious problems inside the corporation, like insufficient remuneration, subpar management techniques, restricted career advancement opportunities, or even a negative company culture. On the other hand, low turnover rates are frequently a sign of a productive workplace and successful employee engagement initiatives.

Using machine learning techniques, this research seeks to improve the accuracy of employee turnover predictions, allowing firms to take proactive steps to hold onto their most valuable employees. This study uses a variety of predictive models, such as Random Forest and Gradient Boosting classifiers, to find hidden patterns and factors that influence employee turnover. By using these insights, the research will make a valuable contribution to the current discussion on personnel management and offer workable solutions to reduce attrition. This paper's goal is to identify the critical factors influencing employee decisions to leave as well as to forecast the probability of such actions. Business executives will gain the information necessary to create more resilient companies where retaining and gratifying employees is of utmost importance.

II. METHODOLOGY

A.Data Description

The HR data utilized in this study came from a Kaggle source that is openly accessible. It consists of documents that documents both the people who have departed and the people who have stayed with the company. The ten attributes that make up the dataset's structure offer a wide range of information, from job-related facts to personal performance measures. Every characteristic is intended to provide information about potential influencing elements on an employee's decision to remain with the organization or depart. A closer examination into each quality is provided below:

Degree of Satisfaction: a number that represents the employee's overall level of job satisfaction and ranges from 0 to 1. Since higher churn rates are generally correlated with lower satisfaction levels, this is an important predictor of turnover.

Last Evaluation: This characteristic, which also has a range of 0 to 1, indicates the employee's score from their most recent performance review. It aids in determining whether performance indicators correspond with outcomes related to employment (staying/leaving).

Number of Projects: The total number of projects that a worker has been allocated is shown by this integer characteristic. It offers information on workload and how it could affect employee churn. Average Monthly Hours: This is a continuous variable that indicates how many hours an employee works on average each month. It is employed to measure the amount of work and operational demands made on staff members.

Duration of Stay at Company: This attribute, which is expressed in years, shows how long a worker has been with the company. While lower tenure may indicate possible problems with job fit or engagement, longer tenure can be a sign of devotion, contentment.

Work Accident: An attribute that is binary (yes or no) that indicates if an employee has experienced an accident at work. This might have an impact on a worker's decision to quit because they are worried about their safety at work.

Promotion in the Last 5 Years: This binary feature indicates if the worker has received a promotion in the last five years. A worker's motivation and propensity to stick with the organization can be greatly impacted by this information.

Departments: A categorical characteristic that designates the division in which a worker is employed. This makes it possible to analyze the turnover rates among the organization's various functionalareas.

Salary: This factor, which is divided into three categories (low, medium, and high), aids in understanding how pay affects employee retention.

Left: A binary result variable that indicates if the worker has quit from the organization (1) or not (0). This study's dependent variable is what predictive models are trained on the technologies .

The consistency and quality of the dataset were carefully examined, and the necessary preparation measures were implemented, such as encoding categorical categories for analysis and managing missing values. The foundation for using different machine learning approaches to accurately anticipate staff attrition is laid by this thorough data review.

B. Analytical Techniques

Predictive modeling and clustering techniques are combined in our investigation since they are individually better suited to identify trends and churn predictors:

Random Forest and Gradient Boosting Classifiers: These sophisticated ensemble models are chosen because to their adeptness in managing intricate feature interactions and non-linear correlations. Large feature sets and a variety of data kinds are two things that they excel at handling in our dataset.

Their capacity to conduct feature importance analysis is essential since it aids in identifying the key elements influencing employee attrition.

K-Nearest Neighbors (KNN): This tool is used to evaluate how similar employees are to one another based on their traits. It aids in grouping employees that share churn tendencies. Cross-validation was used to identify the ideal number of neighbors in order to maintain a balance between the model's sensitivity and the specificity in this KNN.

K-Means Elbow Method Clustering: K-Means is used to find naturally occurring clusters in the employee data that may be associated with various churn rates. The Elbow Method is important here because it allows us to determine the optimal number of clusters to use for maximum insight by showing the sum of squared distances from each place to its assigned cluster center and finding the point at which improvements become negligible in the prediction.

C.Evalution Metrics

In order to guarantee a thorough assessment of our models, we utilize a blend of indicators that enable us to evaluate their efficacy from several perspectives:

Accuracy: This indicator assesses how well the model predicts churn overall. The ratio of correctly predicted observations—both true positives and true negatives—to all of the dataset's observations is how it is computed. Although accuracy is an intuitive indicator of model performance, situations where an overall success rate is needed or where the classes are almost evenly distributed benefit most from its use. Relying only on accuracy, however, may be deceptive in circumstances with imbalanced datasets, as it may reflect the predominance of the majority class for the accuracy.

F1 Score: For datasets with imbalanced classes, like ours, where the percentage of departing employees may be much lower than the percentage of remaining employees, the F1 score is an essential indicator. The ratio of accurately predicted positive observations to all predicted positives, or precision, and recall, or the ratio of properly predicted positive observations to actual positives, are its harmonic means. This statistic offers a fair assessment of the precision and recall of the model and sheds light on the minority class forecast's accuracy, which is crucial information for churn prediction.

Silhouette Score: This measure, which we used in our clustering analysis, assesses how well an object fits within its cluster in comparison to other clusters. It is computed by calculating the separations between every data point, the closest cluster, and the centroid of the cluster to which it belongs. In clustering settings, it is preferable for clusters to be closely spaced and densely packed, as indicated by a higher Silhouette Score. This statistic is especially helpful in verifying that the clusters created are relevant, distinct, and have little overlap—a crucial aspect of confirming the efficiency of the clustering approach.

Through the integration of several analytical tools and careful evaluation measures, our methodology offers a comprehensive framework that can be used to both explain and anticipate employee turnover. With this multimodal approach, we are able to evaluate the efficacy of several exploratory and predictive models in addition to identifying the primary causes causing employee turnover. We make sure that our predictions are accurate and evenly distributed among classes by using the Accuracy and F1 Score metrics, and the Silhouette Score validates that our clustering technique successfully represents the underlying groups in the data. When combined, these measures provide a wealth of information about the dynamics of employee attrition and help shape the creation of focused intervention plans supported by sound analytical data in this silhouette score analysis.

III. Contextual Explanation of Algorithms

The Gradient Boosting Classifier was chosen for its ability to effectively handle complicated datasets that contain a combination of numerical and category characteristics, which is often the case in HR data. It

excels at optimizing various loss functions and handling imbalanced datasets, such as when churn occurrences are relatively scarce compared to non-churn examples.

Mathematical Explanation: The formula $Fm(x)=Fm-1(x)+ \gamma_m h_m(x)$ depicts the iterative update of the model, where Fm(x) is the prediction of the boosted model at iteration m. Each iteration incorporates a novel decision tree $hm(x)$hm(x), adjusted by a factor γm (learning rate), with the purpose of enhancing the model by diminishing mistakes in the predictions made in the previous iteration.

Drawbacks: Despite its ability to handle complicated datasets effectively, the Gradient Boosting Classifier still has limitations. Training can be computationally demanding and time-consuming, particularly when dealing with huge data sets, because of the sequential nature of boosting. Moreover, if the parameters such as tree depth and learning rate are not appropriately calibrated, the model is more likely to suffer from overfitting. The efficacy of the model is strongly dependent on the quality of the data it uses. The presence of noisy data and outliers can greatly diminish the model's performance.

The Random Forest Classifier is chosen for its capacity to handle huge datasets and its resilience against overfitting. This is particularly important in predicting churn, as overfitting can distort the identification of genuine underlying trends. Mathematical Explanation: The Gini impurity, denoted as G, is a measure of the disorder in a set used in decision trees of the Random Forest algorithm. It is calculated using the formula $G = 1 - \sum p_i^2$, where pi represents the chance of selecting an item with label i. This measure aids in effectively dividing the nodes by striving to reduce the uncertainty or 'impurity' following each division, hence maximizing the accuracy of classification at each node in the tree.

Drawbacks: Despite its robustness and effectiveness in handling huge datasets, the Random Forest Classifier nevertheless possesses certain constraints. A significant disadvantage is the complexity of its model. An abundance of trees can significantly impede the model's prediction speed, especially in real-time situations. In addition, although it exhibits superior handling of overfitting compared to several algorithms, it might still experience overfitting if the number of trees is not selected with caution. The interpretability of Random Forest might pose challenges when compared to simpler models, which can be a disadvantage in situations where model transparency is crucial.

K-Nearest Neighbors (KNN) is used because of its effectiveness in classifying instances based on a similarity function. This is particularly important in churn analysis, since identical employee profiles may have similar churn behaviors. Mathematical Explanation: The Euclidean distance formula, denoted as d(x, y), is used to compute the distance between two points x and y in an n-dimensional space. In the KNN algorithm, the distance metric is used to determine the k nearest neighbors. The classification of a sample is then determined by the majority class among these neighbors. This strategy, which relies on proximity, allows the model to make predictions by analyzing patterns that are specific to a certain location.

Drawbacks: The K-nearest neighbors (KNN) classification approach is generally simple and successful. However, it faces difficulties when dealing with high-dimensional environments because of the "curse of dimensionality". This can result in a high processing cost, as the calculation of distance becomes less significant. KNN is also susceptible to the magnitude of the data and extraneous features, which can diminish performance unless the data is appropriately pre-processed. Furthermore, the performance of the system is greatly influenced by the selection of the parameter k, and it is not effective when dealing with imbalanced datasets due to its reliance on the majority voting of the nearest neighbors.

K-Means Clustering: Context: K-Means is selected for its capacity to uncover inherent clusters within

the data, which is helpful in identifying unique kinds of employees who may be inclined to churn. Mathematical Explanation: The objective function of K-Means, $\min_S \sum_{i=1}^{k} \sum_{x \in Si} \|x - \mu i\|^2$, seeks to minimize the variance inside each cluster. In this context, x represents the data points and μi represents the centroid of cluster i. K-Means assures the compactness and distinctiveness of each cluster by minimizing the sum of squared distances between each point and its cluster's centroid. This is crucial for effective segmentation and analysis.

Drawbacks: K-Means clustering is a proficient method for categorizing data into separate groups, however it does have a few constraints. It presupposes that clusters are spherical and uniformly sized, which may not be true in numerous real-world situations. Irregular or extended shapes of the actual clusters can result in subpar performance. K-Means is very susceptible to the choice of starting centroids, and if these centroids are incorrectly initialized, it can lead to varying clustering outcomes across different runs. Moreover, it has poor performance when dealing with clusters that have different densities and sizes.

IV. Results

The Random Forest Classifier exhibited outstanding performance in forecasting employee turnover, attaining an accuracy score of 98.84% and an F1 score of 0.975. These results indicate a significant degree of accuracy and dependability in identifying both the employees who are likely to resign and those who will remain.

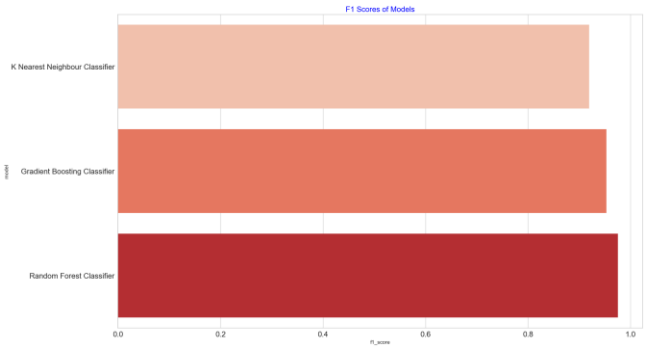

Fig1 F1 Scores of Models

The examination of feature importance indicated that the factors with the greatest impact on forecasting churn are satisfaction_level, number_project, and time_spend_company. These data indicate that employee happiness and workload closely correlate with employee retention, supporting the previous research on employee turnover.

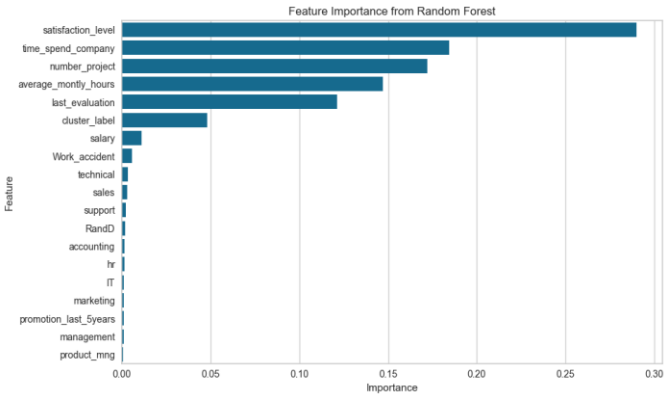

Fig 2 Feature Importance from Random Forest

The classification report of the model provides more evidence of its effectiveness, demonstrating an F1 Score of 0.97 for predicting employees who departed and 0.991 for those who stayed. These precision scores are accompanied by good recall scores. The confusion matrix demonstrates the model's strong predictive abilities, accurately identifying a substantial percentage of genuine positives and true negatives.

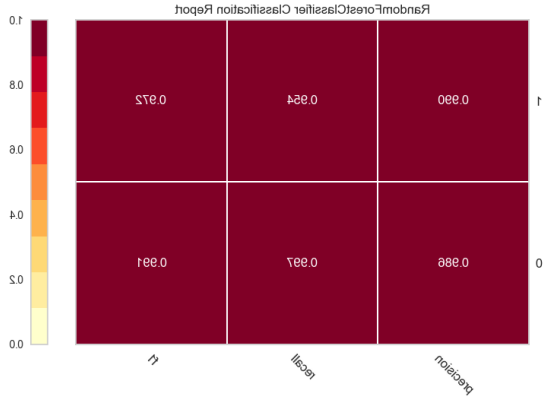

Fig 3 Classification Report

The visual representations of feature importance and model metrics provide persuasive proof of the predictive capability of our model and offer practical insights for HR policies aimed at mitigating employee turnover.

V. Future work

This research establishes a strong basis for using machine learning algorithms to accurately forecast employee turnover. Nevertheless, the ever-changing and adaptable nature of workforce analytics offers numerous opportunities for further investigation: Comparative Analysis: Subsequent research endeavors could encompass a more extensive array of machine learning algorithms, such as Support Vector Machines or sophisticated Neural Networks. This would establish a standard by which the performance of current models can be evaluated, providing valuable information on the most efficient methods for predicting customer attrition.

Feature engineering has the ability to enhance forecasting accuracy. By conducting experiments with various combinations of characteristics and developing novel ones, such as those that capture more subtle aspects of employee attitude or engagement, it is possible to obtain more accurate models.

Temporal dynamics: The departure of employees is affected by time-related factors that were not thoroughly investigated in this study. Examining the progression of employee happiness over time or the impact of seasonal business patterns on job turnover could yield a more dynamic and predictive employee behavior model.

Deploying the model into an operational HR analytics system would enable the validation of its prediction skills in real-world scenarios. Real-time prediction could also empower HR managers to promptly take actions based on the insights provided by the model.

Causal Inference: To go beyond mere correlation, future study could utilize methods such as propensity score matching to investigate the causal connections between job features and employee turnover. This could assist HR departments in not just forecasting, but also comprehending and efficiently resolving the underlying factors contributing to employee turnover.

By implementing these recommendations, future research can expand the limits of comprehending employee turnover, ultimately resulting in the creation of more advanced HR analytics tools that not only forecast but also aid in devising methods to retain employees.

VI. References

1. "Applications," vol. 38, no. 3, pp. 1999–2006, 2011.

2. D. S. Sisodia, S. Vishwakarma, and A. Pujahari, "Evaluation of machine learning models for employee churn prediction," in 2017 International Conference on Inventive Computing and Informatics (ICICI), IEEE, 2017.

3. A. D. Ekawati, "Predictive analytics in employee churn: A systematic literature review," Journal of Management Information and Decision Sciences, vol. 22, no. 4, pp. 387–397, 2019.

4. "Churn Rate Prediction - DataScience Dojo," available at: https://datasciencedojo.com/blog/churn-rate-prediction/

5. A. Aysbt, "Employee Churn Analysis," available at: https://medium.com/@aaysbt/employee-churn-analysis-be94751e4df5

6. S. H. Dolatabadi and F. Keynia, "Designing of customer and employee churn prediction model based on data mining method and neural predictor," in 2017 2nd International Conference on Computer and Communication Systems (ICCCS), pp. 74-77, IEEE, July 2017.

7. "HR Analytics Employee Churn Prediction," available at: https://www.kaggle.com/code/ktakuma/hr-analytics-employee-churn-prediction/input

8. Requirements - Hardware: Device - Ideapad, Processor - 12th Gen Intel(R) Core(TM) i7-1255U, 1.70 GHz.

9. Requirements - Software: Visual Studio Code, Python, Python Libraries - Pandas, Numpy, sklearn, Seaborn.