Predictive Analytics for Employee Attrition and Strategic Retention Interventions

Ajinkya Gajananrao Chintawar
x23113561@student.ncirl.ie
School of Computing
National College of Ireland

Shrey Sanjay Kekade
x23194316@student.ncirl.ie
School of Computing
National College of Ireland

Thanmayee Mandava
x23204192@student.ncirl.ie
School of Computing
National College of Ireland

***Abstract*— Employee turnover presents significant challenges for many organizations. It reduces productivity, morale, and profitability while driving up recruiting & training costs. For this project, we are going to predict employee retention and understand what factors contribute most to an employee leaving the company by using statistical techniques from historical HR data. We also developed a predictive model using the Random Forest algorithm with under-sampling to consider class imbalance and predict accurately which employees are at risk of leaving the organization. Consequently, these findings may be hints at how they could input interventions around ways through which employee satisfaction and retention levels are increased to reduce turnovers in achieving a sustained workforce-company wealth that will enable them to meet their productivity targets.**

Keywords*— *Employee Attrition, HR Analytics, Predictive Analytics, Human Resource Management, Employee Retention

I. INTRODUCTION

Employee attrition is a major issue within organizations, it causes disturbances in operations and processes, low morale among employees as well as increased costs on new recruitment of employee turnover. A sign of high turnover rates can include job dissatisfaction, not enough pay or perks that come with the role and minimum room for growth as well as unhealthy practices by management. Solving these issues is key to ensuring a stable workforce that enables business continuity and growth.

In this research, we work on recognizing crucial influencing factors for employee turnover and develop a model that forecast at risk employees based on historical HR data by using predictive analytics. The aim is to furnish actionable insights to increase employee retention, drive drop-in satisfaction ratings, and lower turnover. The attrition of employees affects not only the organization immediately but also strategically as it leads to loss of knowledge and stands growth prospects at innovation. It can also ruin an organization's image and so it will be a challenge to get good employees. This proves the importance of data-driven human resource management. Predictive analytics helps organizations assess how employees perform and why they leave so that effective, personalized retention strategies can be implemented.

Problem Statement:

- *Operational Impact:* Loss of skilled employees leads to knowledge drain and operational disruptions.
- *Financial Burden:* High turnover increases recruitment and training costs.
- *Reputational Damage:* Frequent attrition harms the company's reputation, making it difficult to attract top talent.
- *Strategic Setbacks:* Employee turnover hinders innovation and long-term growth.
- *Need for Precision:* Traditional methods lack the precision to effectively predict and prevent employee departures, necessitating a more accurate and data-driven approach.

II. LITERATURE REVIEW

A. Exploration of Predictive Techniques for Employee Attrition

Predictive analytics methods are widely used to study the phenomenon of employee attrition & retention. These studies have used a variety of machine-learning algorithms to analyze factors that affect employee churn.

Machine Learning Models for Churn Prediction:

Research has shown that building predictive models for churners is important to a business to know who will be going so the company can start retention strategies appropriately. Hoang Tran et al. conducted a study to investigate the influence of customer segmentation over churn prediction with machine learning models in the banking arena. They applied their feature selection method to the dataset and noticed that Random Forest might give a good classification performance in which it reached 97.25% [1]. Similarly, Dias et al. explored machine learning for customer churn prediction in retail banking and observed that stochastic boosting was the top performer, with a combination of the total value of bank products [2] combined with the number of transactions during those months as key predictors.

Class Imbalance Handling in Churn Prediction:

Class imbalance in churn prediction datasets Sun et al. showed the benefit of balancing datasets by using Synthetic Minority Oversampling (SMOTE) and Under-sampling Techniques, whose results are remarkable as creating synthetic samples for minority classes significantly increases performance [3]. Adoption of this method is widely used for making predictive models more stable in imbalanced prediction [4].

Survival Analysis Models:

Churn behaviours have also been analyzed using survival analysis models. Mavri and Ioannou applied proportional hazard models to study customer churn in the banking sector. with observations that reinforced the notion that better knowledge of switch behaviours can offer winning signals for retention strategies [5]. Survival analysis has been combined with Random Forest techniques by Larivière and Van den Poel to forecast retention as well as profitability [6].

Customer Lifetime Value and RFM Variables:

Glady et al. suggested attributing churners in retail financial services using customer lifetime value as a measure. In another study, they used decision trees, neural networks, and logistic regression models to predict churn using RFM (Recency-Frequency-Monetary) variables by predicting that these techniques allow the identification of potential high-risk customers [7].

Traditional Models: Logistic Regression and Decision Trees

Logistic regression and decision trees are two popular techniques used in churn prediction studies. Nie et al. determined that for the task of predicting credit card churn in Chinese banks, logistic regression performed better than decision trees [8]. Lin et al. The proposed approach properly resembles the behaviour detection of credit card fraud and was applied to discover rules with a rough set theory for explaining customer churn [9].

B. Selected Methodology for Predicting Employee Attrition

These techniques reaffirm the necessity of deploying a strong predictive model for employee attrition. Using the Random Forest algorithm and under-sampling for balancing data, we want a model that predicts future employee attrition rates to provide actionable steps toward improving employee satisfaction/retention. This method leverages previous work and skilfully targets the unique hurdles that we face with employee churn, rendering it a star in our effort to combat high business costs due to turnover making it an ideal choice for study.

III. METHODOLOGY

The CRISP-DM framework is one of the most common methodologies for data mining projects, and this project adheres to it. It is still a systematic process with the following six phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. This framework will help in making sure that all around systematical approach is achieved during the entire employee attrition analysis and employee churn predictive model development.

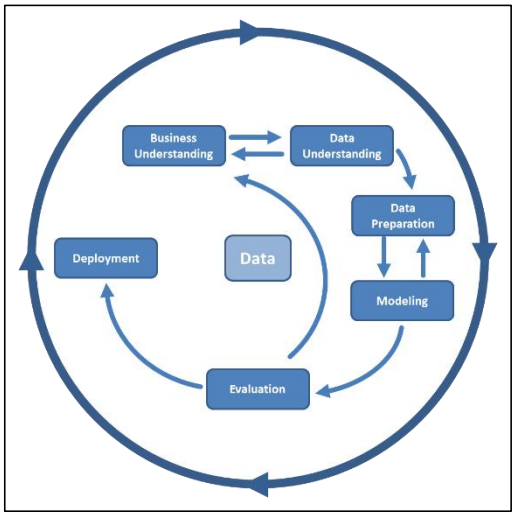

Figure 1: CRISP-DM Framework

A) Business Understanding

The first phase involves Understanding the context of the business and its objectives. Regardless of the reason, turnover at any level is a fundamental issue for an organization that on one hand brings productivity to low levels by default and kills morale and profitability massively accordingly while vastly psychosis expensive recruitment returns up together with their training. The project's purpose should be to build a predictive model that is able to predict whether the employee will leave from organization or not. We will learn the top things that lead to attrition and shall also discuss useful suggestions on how you can save your employees for longer term and eventually reduce turnover rates as well.

B) Data Understanding

This is the phase where we inspect and discover more about our dataset. The dataset is taken from Kaggle and contains records of 1470 so each employee will have 35 features. Some of the key attributes are age, monthly income, job role and whether an employee has left the organization or not. Some of the initial data exploration includes Summarizing Data, Discovering Missing values, and EDA (Explorative Data Analysis - Distribution of features). Examples of the analysis included calculating descriptives to inform central tendencies and variance and producing visualizations to identify patterns or outliers.

C) Data Preparation

Data cleaning is a process that cleans the data to maximize output, and usually occurs right before Data Transformation. This part involves dealing with missing values, encoding categorical variables, and scaling numerical features. Use some under-sampling to balance the classes in the dataset. This implies that while working with this model, there will be no bias in the majority class.

Steps in data preparation included:

i. *Handling Missing Values*: Ensuring no missing values were present in the dataset.
ii. *Encoding Categorical Variables*: Converting categorical variables into numerical format using Label Encoding.

iii. *Scaling Numerical Features*: Normalizing numerical features using Standard Scaling.
iv. *Addressing Class Imbalance*: Under sampling the dataset to equalize the counts of each target value in both Classification Algorithms for this reason.

Figure 2 below shows the class distribution before under-sampling where we can see that dataset is imbalanced. Figure 2 presents the class distribution post-under-sampling, showing a well-supported dataset.

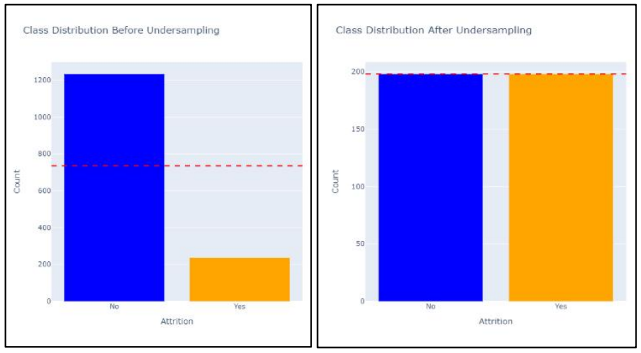

Figure 2: Class Distribution Before and Under Sampling

D) Modelling

During the modeling, we use different machine learning algorithms to create a predictive model. Among many algorithms we worked with Random Forest, it is stronger and more accurate in handling large datasets. Performed the hyperparameter tuning using GridSearchCV to maximize model accuracy. Furthermore, feature importance analysis was applied to discover the most important predictors of employee attrition.

E) Evaluation

This is a stage where evaluation of the model is taken on the different metrics such as accuracy, precision, recall, F1 score, and ROC-AUC. The former includes confusion matrices and ROC curves whereas for the latter, we already have an answer. At the heart of achieving reductions in human resources and days at risk, predictions of employee attrition by the final model have been crucial to formulating plans for HR interventions.

A systematic guided approach guarantees that developers get through all the stages of the CRISP-DM framework one after other without fail ultimately resulting in a reliable and explainable predictive model to determine Employee attrition.

IV. RESULTS

This study's results are quantitative and qualitative, allowing for a comprehensive understanding of employee turnover as well as actionable insights into an organization on how they can increase retention.

A) Quantitative Findings

Overall, it was 75% accurate in correctly predicting the attrition status of most employees.

Table 1: Key performance metrics for the model

Accuracy	Precision	Recall	F1 Score	ROC-AUC
0.75	0.84	0.75	0.78	0.73

These metrics indicate a tradeoff of precision, which is about ensuring that all who will leave are identified correctly in exchange for losing some others from the right answers. ROC-AUC score indicates that the model has good diagnostic capabilities to well distinguish between employees who are going to stay vs those at risk of leaving [Figure 2].

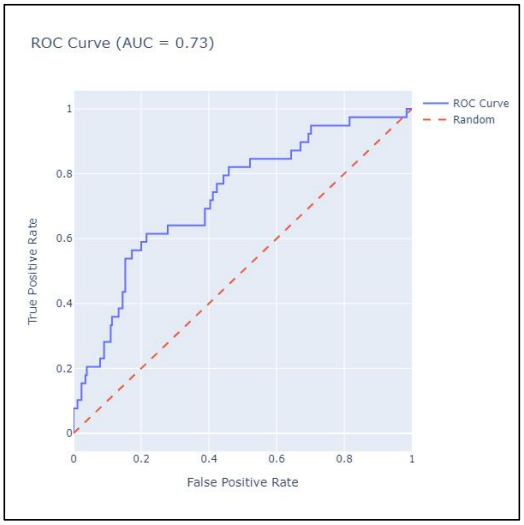

Figure 3: ROC Curve (AUC = 0.73)

The ROC curve (Figure 3) with an AUC score of 0.73 is approximately for measuring the separability or preciseness with which model output classes can be distinguished, i.e., how well positive rates are predicted as true in comparison to false. In a ROC curve, the true positive rate is plotted as a function of the false positive rate for different cut-off points. A score of 0.73 or greater indicates a good ability to distinguish between employees who will leave and those who do not, but it is not a perfect model with the highest possible AUC. This score suggests that the model does not have a fine-grain discriminatory power.

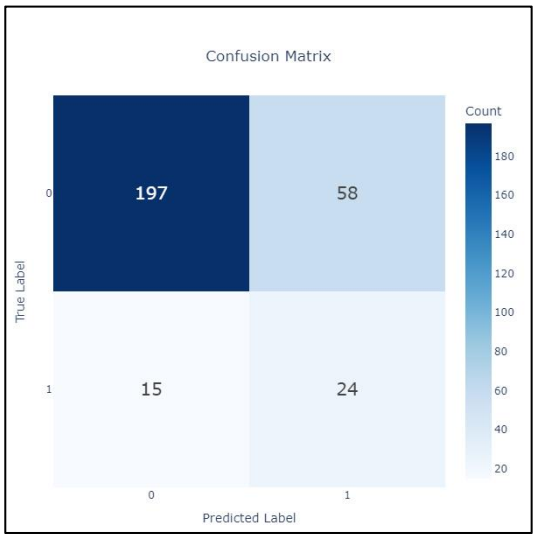

Figure 4: Confusion Matrix

Figure 4 illustrates the Confusion matrix, which shows how well our model has performed in terms of true positives (attrition correctly predicted), and vice versa for false negatives - attrition wrongly predicted. With this matrix, we can get a more precise look at how well the model is predicting as compared to the distribution of errors, which on one hand tells us about the recall and the other side of precision. If there are many true positives and negatives, the model is reliable whereas if false detection rates arise because of an incorrect classification by the algorithm were real churners classified incorrectly as non-churners or vice versa.

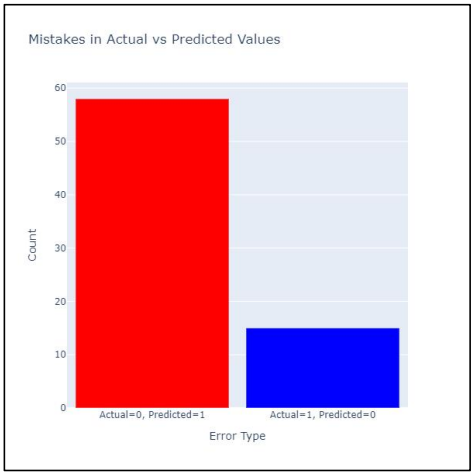

Figure 5: Mistakes in Actual vs Predicted Values

Figure 5 illustrates the distribution of mistakes in actual vs. predicted values. The model made 221 correct predictions and 73 mistakes, resulting in a ratio of correct predictions to mistakes of approximately 3.03. This indicates that for every mistake, the model made about three correct predictions, highlighting its overall effectiveness in forecasting employee attrition while still leaving room for improvement.

Key Predictors:

The feature importance analysis revealed several critical factors influencing employee turnover:

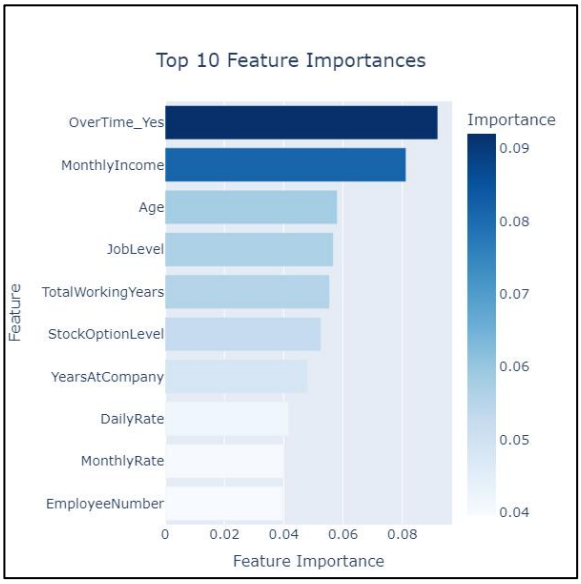

Figure 6: Top 10 Feature Importances

- *Overtime:* For employees working an extensive number of extra hours, it is crucial to progress in their career thus managing workloads and striving for a balanced/work life.
- *Monthly Income:* Sufficient income will be required to make employees stay as low salaries lead more likely to attrition.
- *Age:* Younger employees at age tend to have a higher turnover rate, often because they are still viewing career mobility and exploration.
- *Job Level:* High-level job negatively affects Attrition. Since the more career advancement opportunities, the longer employees tend to remain.
- *Total Working Years:* The longer an employee, the more stable career progression is and results are of course important.
- *Stock Option Level*: Higher share options are a good thing for retention because they offer long-term financial incentives.
- *Years at Company:* Loyalty and longevity are key to growing a company, so we focused on the desire for people to stay with us.

B) Business Implications and Qualitative Analysis

A qualitative takeaway of the insights was made through feedback by employees and survey results which highlighted that career development possibilities along with personal interventions were major push factors in retaining employees. Themes emerging from qualitative analysis included perceptions of:

- *Career Development:* High desire for visibility in terms of career progression and professional growth opportunities. Offering structured career plans is an opportunity to keep employees satisfied and engaged. With training programs, mentorship opportunities, and different levels of career ladders, companies can create a path to success that is aligned with ambition. That investment is not only good for retention but also results in a higher-skilled and capable workforce.
- *Work-Life Balance:* More than half of the workers who have already left focused on excessive overtime and a lack of work-life balance. Flexible work regulations, such as working remotely and habitual breaks with flexible hours will aid in minimizing this. This results in more productive, less stressed employees who are ultimately satisfied with their jobs. Not only does this blunt-force solution lead to lower burnout and turnover for a healthier, more productive workforce.

Personalized Interventions: A retention strategy aimed at each employee uniquely enables companies to solve unique issues and requirements hence making the individual feel valued and understood so his/her experience is personalized. Providing HR with powers of prediction, to be able to identify which employees are at risk has given them the advantage of extending a helping hand, whether that means unique benefits for everyone, job development within your area and industry, or an employee wellness program. A personalized approach does have the potential to vastly improve employee morale and retention as the company

caring so much about what is best for each individual shows itself.

Implications for Business Strategy:

- *Improved Retention:* When career development, work-life balance, and Universal Interventions addressing selecting features from all three regions of need are in place, employees stay a long time. Predictability in a healthy way which allows one to plan for the long term and operate with ease.
- *Increased Employee Satisfaction:* Putting importance on career development and work-life balance directly influences employee satisfaction. When employees are happy at work, they will be more engaged and productive, it also helps with retaining them in the long term.
- *Cost Savings:* Reducing turnover, lower recruitment and training costs will be less and this amount adds up to some profound financial savings. It also prevents a brain drain of knowledge and concerns about your overall productivity standing back up to capacity.
- *Improved Employer Branding:* Those organizations that are marketing career progression, work-life balance, and well-being opportunities will win over the best talent. This service improves the reputation and competitiveness of an organization regarding jobs.
- *Proactive HR Management: It* is more proactive for HR management if retention strategies are designed on qualitative insights. The company can eliminate the source of problems before they blow up by anticipating and proactively meeting employee demands.

Overall Model Performance: The overall model considers a balanced data set and therefore we can say this data set would provide accurate predictions because there is the least bias in it. The confusion matrix was also realized, and the ROC curve further verified the model.

V. Conclusion

For any business, such findings are vital insights that can be used to create targeted HR interventions. High turnover disrupts operations, drives up costs, and impacts morale. So, using predictive analytics, businesses can not only detect which of their employees are at a greater risk of leaving but also guide them and undertake relevant retention initiatives. Improving work/life balance and salary competitiveness are important steps in reducing turnover. In addition, offering an apprenticeship scheme and developing clear structures for incentives can help keep top employees, thus creating a sustainable business.

By implementing any of these strategies, not only can turnover be reduced but also a more positive work environment is created. A reliable workforce helps in operations to continue high levels of expertise and innovation. Higher employee morale and satisfaction means higher productivity as well, and this also results in an overall reduction of future turnover. In conclusion, organizations with predictive analytics on staff turnover help to not only manage this process but also critical for long-term business success and competitive differentiation as a workforce that is resilient and motivated of employees.

REFERENCES

[1] H. Tran, N. Le, and V.-H. Nguyen, "Customer churn prediction in the banking sector using machine learning-based classification models," Interdisciplinary Journal of Information, Knowledge & Management, vol. 18, pp. 87-105, 2023. DOI: 10.28945/5086.

[2] J. Dias, P. Godinho, and P. Torres, "Machine learning for customer churn prediction in retail banking," in ICCSA 2020, LNCS 12251, O. Gervasi et al., Eds. Springer, 2020, pp. 576-589. DOI: 10.1007/978-3-030-58808-3_42.

[3] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: a review," International Journal of Pattern Recognition and Artificial Intelligence, vol. 23, no. 4, pp. 687-719, 2009.

[4] Y. Xie, X. Li, E. W. Ngai, and W. Ying, "Customer churn prediction using improved balanced random forests," Expert Systems with Applications, vol. 36, no. 3, pp. 5445-5449, 2009.

[5] M. Mavri and G. Ioannou, "Customer switching behavior in Greek banking services using survival analysis," Management of Financial Services, vol. 34, pp. 186-197, 2008.

[6] B. Larivière and D. Van den Poel, "Investigating the role of product features in preventing customer churn using survival analysis and choice modeling: the case of financial services," Expert Systems with Applications, vol. 27, no. 2, pp. 277-285, 2004.

[7] N. Glady, B. Baesens, and C. Croux, "Modeling churn using customer lifetime value," European Journal of Operational Research, vol. 197, no. 1, pp. 402-411, 2009.

[8] G. Nie, W. Rowe, L. Zhang, Y. Tian, and Y. Shi, "Credit card churn forecasting by logistic regression and decision tree," Expert Systems with Applications, vol. 38, no. 12, pp. 15273-15285, 2011.

[9] C.-S. Lin, G.-H. Tzeng, and Y.-C. Chin, "Combined rough set theory and flow network graph to predict customer churn in credit card accounts," Expert Systems with Applications, vol. 38, no. 1, pp. 8-15, 2011.

[10] W. Buckinx and D. Van den Poel, "Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting," European Journal of Operational Research, vol. 164, no. 1, pp. 252-268, 2005.

[11] K. Coussement, D. F. Benoit, and D. Van den Poel, "Improved marketing decision making in a customer churn prediction context using generalized additive models," Expert Systems with Applications, vol. 37, no. 3, pp. 2132-2143, 2010.

[12] K. H. Cho, S. G. Huh, and M. Y. Cho, "Customer churn prediction using decision tree techniques," International Journal of Information Management, vol. 21, no. 5, pp. 1-12, 2008.