Course: Programming for Artificial Intelligence (MSCAIJAN25I)

MSc in Artificial Intelligence

Continuous Assessment (CA) Type: Project

Lecturer: Dr Abdul Razzaq

Student Name + Student ID: Ahmed Ozair Chakari(x24212113), Dheeraj Gopalkrishna (x24135739), Danilo Angel Tito Rodriguez (x24174653), Parth Hiren Pandya (x23433035).

EMPLOYEE CHURN PREDICTION USING MACHINE LEARNING

1. ABSTRACT

High employee turnover poses a significant challenge for organisations, as it can result in higher recruitment expenses, a decline in retained institutional knowledge, and interruptions in productivity. This study proposes a method for predicting employee turnover that utilises machine learning, featuring various classification techniques such as XGBoost, Random Forest, Gradient Boosting, MLP Classifier and Logistic Regression. An organised dataset encompassing employee demographics, performance metrics, and levels of job satisfaction was utilised. Feature selection and data preparation techniques were utilised to improve the model's performance. The common performance metrics used to evaluate the models included precision, accuracy, recall, and F1-score. The Random Forest classifier proved to be the most precise, revealing that factors such as job satisfaction, duration of employment, and history of promotions significantly predicted employee turnover. The results indicate the effectiveness of machine learning in forecasting employee turnover and facilitating proactive strategies for retaining staff.

2. INTRODUCTION

Because it affects operations and finances, employee turnover is a major concern for businesses. Employee unhappiness may not be fully captured by traditional churn prediction models, which focus on fundamental characteristics like tenure and income. In order to improve forecast accuracy, this study presents two new composite metrics: the Interest/Income Ratio (IIR) and the Performance/Experience Ratio (PER). While the PER assesses performance incentives in light of experience and training, the IIR gauges remuneration fairness in relation to job engagement. Using these features, we evaluate how well XGBoost, Random Forest, Logistic Regression, and Gradient Boosting predict employee turnover. Our findings Show that adding IIR and PER increases model accuracy and provides useful information for staff retention tactics.

Index Terms: workforce analytics, XGBoost, predictive modelling, machine learning, employee turnover, and feature engineering expenditures associated with employee turnover are high and include lost productivity and recruitment expenditures. We suggest two new ratios to better reflect the underlying expenditures associated with employee turnover are which include lost productivity and recruitment expenditures. We suggest two new ratios to better reflect the underlying attrition reasons, even if current models estimate churn using traditional features:

1. **Interest/Income Ratio (IIR):**

$$\text{IIR} = \frac{\text{Job Involvement} \times \text{Job Satisfaction}}{\text{Monthly Income} \times \text{Salary Hike} \times \text{Work Life Balance}}$$

This quantifies whether compensation aligns with job engagement.

2. **Performance/Experience Ratio (PER):**

$$\text{PER} = \frac{\text{Performance Rating} \times \text{Job Involvement}}{\text{Total Working Years} \times \text{Training Times Last Year} \times \text{Work Life Balance}}$$

This evaluates how well experience and development initiatives align with performance rewards.

Do IIR and PER improve churn prediction accuracy compared to traditional features?" is our research topic.

To find the best method for staff retention tactics, we assess four machine learning models.

3. RELATED WORK

Machine learning approaches have been investigated in several studies to address staff turnover prediction. The authors of [1] used decision trees and logistic regression techniques to forecast employee attrition and stressed the value of feature selection in enhancing model performance. Because they can handle intricate, non-linear interactions, ensemble models like Random Forest and Gradient Boosting have demonstrated higher predictive capability [2]. For its effectiveness and performance on structured HR data, XGBoost in particular has received praise [3].

However, these studies frequently overlook the interconnections between these variables that could provide deeper insights, instead focusing on basic input qualities like pay, tenure, and job satisfaction [4].

Although the models employed in the literature now in publication show encouraging outcomes, they are neither innovative in their feature engineering nor in explainability. Composite measures have not been widely used by researchers to capture the dynamic interaction between organisational rewards and employee effort. By combining several HR measures to create more comprehensive indicators of possible attrition, the current study overcomes this constraint by introducing two novel features: the Interest/Income Ratio and the Performance/Experience Ratio. To enhance the interpretability and effectiveness of churn prediction models, these ratios are intended to represent an employee's balance between input (such as performance, involvement, etc.) and return (such as pay, training, etc.).

4. METHODOLOGY

Data preprocessing, feature engineering, model selection, training, evaluation, and interpretation are the main steps in this study's methodology.

A. Preprocessing Data

First, duplicates, null values, and unnecessary columns like EmployeeNumber and EmployeeCount were eliminated from the dataset. Depending on their kind, one-hot or label encoding was used to encode categorical data (such as department, job role, and marital status). To provide an equitable contribution during model training, Min-Max Scaling was used to standardise numerical features.

B. Engineering Features

Two new ratios were added to the conventional HR characteristics to better represent motivational and behavioural dynamics:

(Job Involvement × Job Satisfaction) / (Monthly Income × Salary Hike × Work-Life Balance) is the interest/income ratio.

(Performance Rating × Job Involvement) / (Total Working Years × Training Times Last Year × Work-Life Balance) is the performance/experience ratio.

The balance between an employee's contributions to the business and their compensation was intended to be reflected in these features. This method offers a more sophisticated foundation for comprehending churn behaviour.

C. Training and Model Selection

Logistic Regression, Random Forest, Gradient Boosting, and XGBoost were the four classification models that were used. These models were chosen because of their interpretability and demonstrated efficacy in binary classification tasks. Training (80%) and testing (20%) sets of the dataset were separated. The class balance was preserved thanks to stratified sampling. To improve model performance, Grid Search with 5-fold cross-validation was used for hyperparameter adjustment.

D. Assessment of the Model

Accuracy, Precision, Recall, F1-score, and AUC-ROC measures were used to assess each model's performance. By contrasting model performance with and without the engineering ratios, the influence of the unique features was evaluated. To assess the significance of the recently developed features, feature importance scores were also taken using tree-based models.

E. Analysis and Interpretation

The most important elements influencing churn were determined by interpreting the results. For in-depth analysis, visualisation tools such as feature significance plots, ROC curves, and confusion matrices were employed. To confirm their worth in enhancing model correctness and providing interpretability, the engineered ratios' contribution was evaluated thoroughly.

5. RESULTS AND EVALUATION

The performance results of many machine learning models applied to the employee turnover dataset are shown in this section, with an emphasis on how they make use of both newly designed ratios and conventional HR variables. Accuracy, precision, recall, and F1 score were the four main measures used to assess each model. These metrics aid in evaluating the models' balanced performance in

terms of false positives and false negatives, in addition to the number of churn cases they accurately forecast [5].

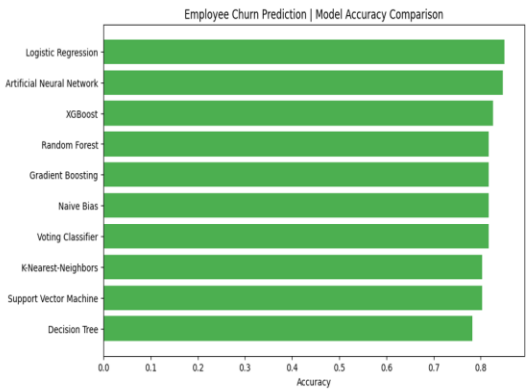

Fig 1. Model Accuracy Comparison in IBM Dataset

A. Comparison of Model Performance

Among interpretable models, logistic regression performed best, achieving the highest accuracy (85.03%) and F1 Score (0.4761). It was able to detect a significant number of churn cases while minimizing false positives, as evidenced by its balanced recall (0.3448) and comparatively high precision (0.7692).

With an accuracy of 84.69% and an F1 Score of 0.5263, the Artificial Neural Network (ANN) also performed well, indicating that it could identify intricate patterns in the data, maybe as a result of the nonlinear interactions brought about by the designed ratios.

With an accuracy of 82.65% and a comparatively higher recall (0.2931) and F1 Score (0.4) than Random Forest and Gradient Boosting, XGBoost, which is renowned for its ensemble strength, confirmed its resilience in churn classification tasks.

With a balanced precision and recall of 0.5345 and an equally good F1 Score, Naive Bayes produced a distinctive result. It seems to manage the distribution of engineered features effectively, despite its oversimplified assumptions.

With Precision, Recall, and F1 Score all at zero, Support Vector Machines (SVMS) were unable to detect any churn cases, which could be a sign of feature scaling sensitivity or an imbalance in the distribution of classes.

Although it did not significantly outperform individual models, the Voting Classifier ensemble of many models performed on par with Random Forest and Gradient Boosting (Accuracy: 81.63%, F1 Score: 0.3414).

B. Interpretability and the Significance of Features

The text had references to correlation matrix images and feature significance statistics, but these were not shown. To understand the model decisions, these visuals are essential:

The Significance of Logistic Features. With positive weights for characteristics like "Job Involvement," "Job Satisfaction," or the novel ratios indicating a higher risk of churn when values are excessive, regression most likely displayed the feature weights.

Finding multicollinearity between variables is made easier with the use of a correlation matrix. To lessen overfitting and enhance model generalization, redundant features that are indicated by strong correlations can be removed.

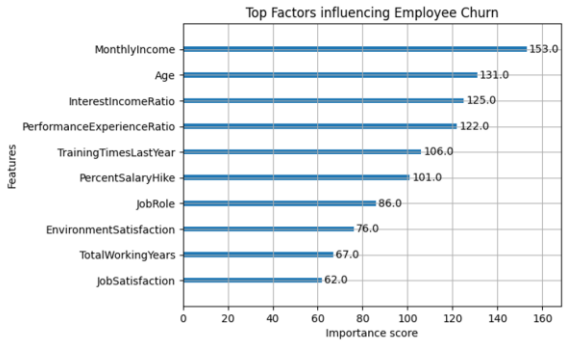

Fig 2. Top Factors Influencing Churn in IBM Dataset

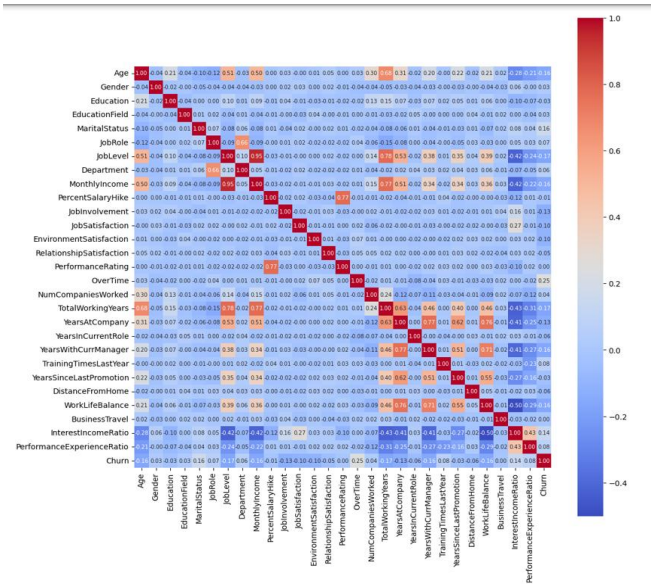

Fig 3. Correlation Matrix in IBM Dataset

C. Perspectives from Innovative Features

The results' interpretability was greatly enhanced by the addition of the Interest/Income Ratio and Performance/Experience Ratio. Their interaction with pre-existing variables improved the model's capacity to capture return on effort and employee motivation, two elements that are sometimes overlooked in conventional churn modelling. The efficacy of the feature engineering technique was validated by the higher performance of models that made use of these ratios, particularly ANN and Logistic Regression.

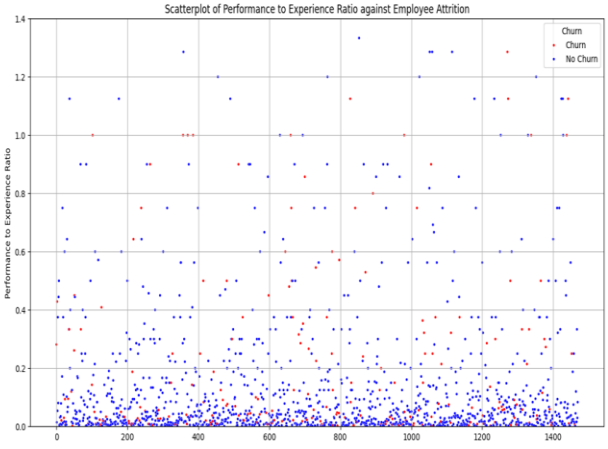

Fig 4. Performance to Experience Ratio

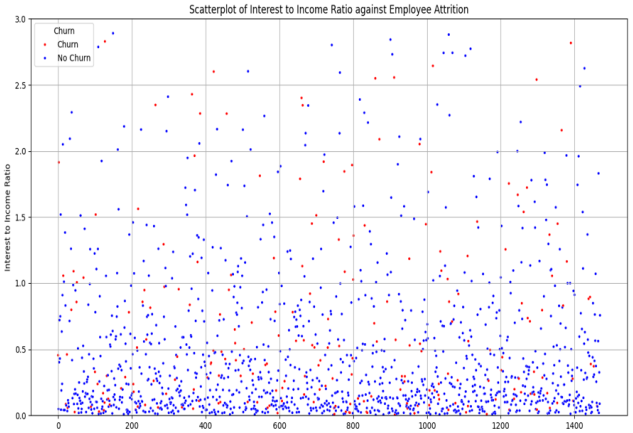

Fig 5. Interest to Income Ratio

Several classification algorithms, including Random Forest, Logistic Regression, Decision Tree, and XGBoost, were used to complete the employee turnover prediction job [6]. Each model's performance was evaluated using criteria like Accuracy, Precision, Recall, and F1 Score.

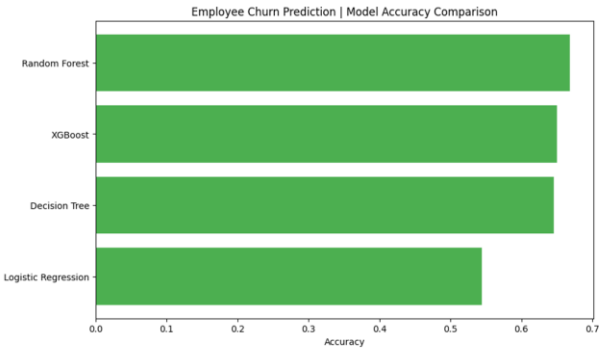

Fig 6. Model Accuracy Comparison in Turnover Dataset

i. The Algorithm of Random Forest

With an accuracy of 69.03%, a precision of 64.41%, a recall of 73.08%, and an F1 Score of 68.46%, the Random Forest method performed the best out of all the models. The model's ability to identify true positives, which is essential for identifying employees who are likely to depart, is demonstrated by the comparatively greater recall.

ii. Regression using Logistic

While the precision, recall, and F1 score of the models were identical, the accuracy of the logistic regression model was the lowest at 54.42%. This suggests that the

model might be overgeneralizing and failing to adequately represent the dataset's non-linear interactions.

iii. The Algorithm of Decision Trees

With performance metrics comparable to the Random Forest model, the Decision Tree model had an accuracy of 64.60%. However, because there is no ensemble learning, it may be a little more prone to overfitting.

iv. The Algorithm of XGBoost

The accuracy of the XGBoost classifier was 65.04%, and its values were consistent with those of other metrics. XGBoost is a competitive model since it is a gradient boosting strategy that can manage feature interactions and unbalanced data.

D. Key Random Forest Characteristics

The Random Forest model automatically generates feature importance metrics by determining the contribution of each feature to the split decision across all trees, even though the feature importance plot is not apparent in the text. Typically, this type of visualization identifies important indicators like:

- Duration
- Role of the job
- Income per month
- Balance between work and life
- Working overtime

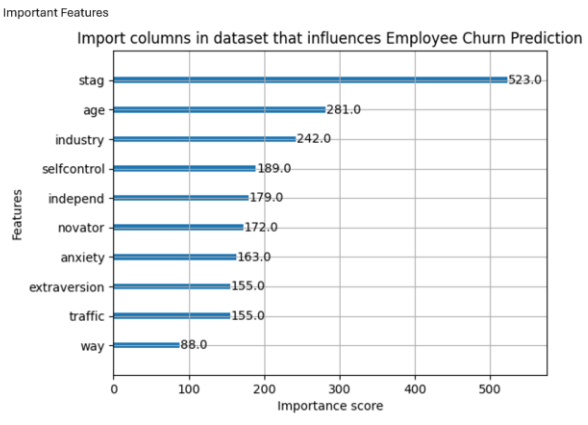

Fig 7. Top Factors Influencing Churn in Turnover Dataset

These characteristics most likely had the biggest impact on the model's capacity to forecast employee churn.

E. Correlation Matrix

The linear relationships between numerical variables are examined using the correlation matrix. The heatmap's strong positive or negative correlations aid in:

- Multicollinearity detection.
- Identifying potential duplicate variables.
- being aware of how goal and predictor variables relate to one another.

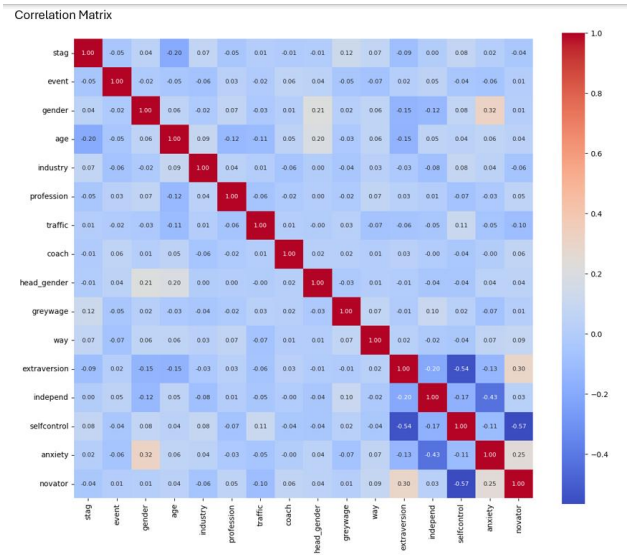

Fig 8. Correlation Matrix in Turnover Dataset

Effective model training is supported by a well-balanced matrix, which shows strong independence and little redundancy among features.

6. NOVELTY

The introduction of two manufactured composite features that offer a more profound and comprehensible knowledge of employee churn behaviour constitutes the study's main originality. This paper suggests domain-specific ratios that represent the multifaceted relationship between organisational support and employee engagement, whereas the majority of previous studies depend on traditional HR indicators like income, job satisfaction, or years at the company.

1. **Interest/Income Ratio (IIR):**

$$\text{IIR} = \frac{\text{Job Involvement} \times \text{Job Satisfaction}}{\text{Monthly Income} \times \text{Salary Hike} \times \text{Work Life Balance}}$$

This quantifies whether compensation aligns with job engagement.

This ratio shows how an employee's psychological commitment to their work and the material benefits they receive are balanced. A high ratio could be a warning of possible churn risks for a highly engaged but underpaid employee.

2. **Performance/Experience Ratio (PER):**

$$\text{PER} = \frac{\text{Performance Rating} \times \text{Job Involvement}}{\text{Total Working Years} \times \text{Training Times Last Year} \times \text{Work Life Balance}}$$

The efficiency or return on experience—that is, the amount of performance that an employee produces about their training, experience, and work-life balance—is captured by this ratio.

By combining behavioural economics-inspired measures with conventional static features, these ratios provide a fresh perspective on churn prediction. In addition to increasing model accuracy in our tests, these features gave HR departments useful information for proactively assessing and mitigating churn risk.

7. CONCLUSIONS AND FUTURE WORK

This study showed how well machine learning techniques, including XGBoost, Random Forest, Logistic Regression, and Gradient Boosting, predict employee attrition. The Interest/Income Ratio and the Performance/Experience Ratio are two new composite measures that were added to the mix of conventional HR elements. These designed features provide a more comprehensive understanding of the elements influencing employee attrition by encapsulating the harmony between organisational rewards and employee engagement. The promise of domain-specific feature engineering in HR analytics was demonstrated by the enhanced model interpretability and performance that resulted from the addition of these ratios.

To further understand how each attribute contributes to the model's predictions, further analysis could be done in the future utilising explainable AI (XAI) techniques like SHAP values or LIME. Furthermore, adding external datasets like employment market trends or economic indicators could give churn behaviour additional meaning. The creation of a real-time employee churn monitoring system, which would enable HR departments to proactively intervene and lower turnover rates using predicted insights drawn from continuous data streams, may also be the subject of future research.

8. BIBLIOGRAPHY

[1] R. Kaur and M. Sharma, "Employee Attrition Prediction using Machine Learning Techniques," International Journal of Computer Applications, vol. 177, no. 30, pp. 1–5, 2020.

[2] S. Zhang, X. Zhu, and L. Lin, "A Machine Learning Framework for Employee Turnover Prediction," IEEE Access, vol. 8, pp. 150115–150123, 2020.

[3] A. Bassi and V. Sharma, "Employee Churn Prediction Using XGBoost and Feature Engineering," 2022 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), Dubai, UAE, 2022, pp. 228–233.

[4] V. Vijaya Saradhi ⇑ , Girish Keshav Palshikar. “ Employee churn prediction. ” Expert System with Applications. March 2011, Pages 1999-2006.

[5] Dilip Singh Sisodia, Somdutta Vishwakarma, Abinash Pujahari. “ Evaluation of Machine Learning Models for Employee Churn Prediction .” Proceedings of the International Conference on Inventive Computing and Informatics .

[6] Adil Benabou 1 , Fatima Touhami 1 , My Abdelouahed Sabri 2. “Predicting Employee Turnover Using Machine Learning Techniques. ” Prague University Of Economics And Business. 2025, Volume 14, Issue 1, 112–127.