Employee Churn Analysis

BEM2031 Introduction To Business Analytics - Final Report

Student Number - 720029134

Table of contents

GenAI Statement

Responsible and Ethical Use of GenAI Tools in the Business School

Within the Business School, we support the responsible and ethical use of GenAI tools, and we seek to develop your ability to use these tools to help you study and learn. An important part of this process is being transparent about how you have used GenAI tools during the preparation of your assignments.Information about GenAI can be found [here] and guidance on the responsible use of GenAI tools can be found [here]. The declaration below is intended to guide transparency in the use of GenAI tools and to assist you in ensuring the appropriate citation of those tools within your work.

GenAI Declaration

I *have* used GenAI tools in the production of this work.

The following GenAI tools have been used: *[please specify]* ChatGPT*

I have used GenAI tools for the following purposes:

☐ I have used GenAI tools to assist with research or gathering information.

☐ I have used GenAI tools to help me understand key theories and concepts.

☐ I have used GenAI tools to help me analyse data.

☐ I have used GenAI tools to create code.

☐ I have used GenAI tools to suggest a plan or structure of my assessment.

- [Yes] I have used GenAI tools to give me feedback on a draft.

☐ I have used GenAI tools to generate images, figures, or diagrams.

☐ I have used GenAI tools to generate creative content for my work.

☐ I have used GenAI tools to proofread and correct grammar or spelling errors.

- [Yes] **Other** *(please specify)* Debugging Code

Declaration of Citation

- [Yes] I declare that I have referenced the use of GenAI tools and outputs within my assessment in line with the University guidelines for referencing GenAI in academic work.

1. Business Understanding

This report aims to understand the reasons why employees leave an organisation, using historical human resources (HR) data to identify patterns in employee behaviour that may indicate dissatisfaction or disengagement. Employee churn presents significant challenges to businesses, especially when high-performing individuals leave unexpectedly. The associated financial, operational, and cultural costs can be extensive. By analysing structured HR data, this project aims to support the development of more effective and targeted retention strategies, shifting HR decision-making from reactive to proactive.

The dataset consists of 15,000 individual employee records, each with ten features covering both behavioural and employment-related metrics. These include *satisfaction_level*, *last_evaluation*, *number_project*, *average_monthly_hours*, *time_spend_company*, *Work_accident*, *promotion_last_5years*, *sales*, *salary*, and the target variable *left*. The report being reviewed applies a structured analytical approach, combining exploratory data analysis, clustering, and classification models (Decision Tree and Random Forest) to invesitgate churn patterns. It identifies key variables that correlate strongly with employee departures, particularly low satisfaction, workload extremes, and limited progression.

While the analysis is well executed and the findings are insightful, the report does have limitations that should be considered when applying its conclusions. The dataset is confined to internal company data and does not account for external factors such as job market competitiveness, macroeconomic conditions, or alternative career opportunities. In addition, it does not distinguish between voluntary and involuntary churn, an important distinction in practice. Despite this, the data remains useful for decision-making, especially when interpreted within a broader organisational context and alongside qualitative insights such as employee surveys or exit interviews.

There are clear benefits to this type of analysis. It provides HR and leadership teams with actionable insights into which employees may be at risk of leaving and why, supporting more tailored retention initiatives. These might include interventions around workload, job satisfaction, or career development. However, there are also risks. Misapplication of model predictions could lead to unjustified monitoring or assumptions about employees, particularly if the data contains underlying biases. There are also ethical implications associated with using predictive analytics in HR, especially when decisions are made without transparency or employee involvement.

In terms of impact, this analysis stands to benefit HR professionals, team managers, and employees provided it is used responsibly. When integrated thoughtfully into HR practice, it has the potential to improve employee engagement, reduce unnecessary turnover, and contribute to a more stable and satisfied workforce. However, care must be taken to ensure that models are validated, results are contextualised, and decisions are guided not just by data, but by ethical considerations and human judgement. Used in this way, the analysis becomes a powerful support tool for evidence-based HR strategy.

2. Data Understanding

The dataset used within Ghouzam's paper provides an adequate representation of the required features that could have an impact on whether employees would leave the company. Features like *satisfaction_level*, *last_evaluation*, *average_monthly_hours*, and *number_project* allow the analysis to easily identify employee happiness, performance, and workload. However, salary being a categorical variable with only three levels reduces the relevance of this variable and could be improved. Furthermore, the data doesn't capture the reasons for employee churn within the company; the distinction between voluntary and involuntary attrition is a vital feature that could provide the company with valuable insights. Also, the lack of demographic data could reduce the efficacy of this analysis as these features could uncover deep patterns in employee churn. Other data which could improve the applications of this modelling approach within the business consist of external factors, for example, market trends or competitor offers, which could have clear relationships with whether an employee is going to leave or not.

Beyond the raw data, the analysis in the paper did a good job of exploring some useful visualisations. The radar chart, for instance, was a helpful way to summarise mean differences between employees who stayed and those who left. It showed that the main difference was *satisfaction_level*, with employees who left typically being less satisfied. However, the use of average values in radar plots can hide important subgroups in the data. This becomes even more relevant when you consider that not all employees who leave do so for the same reason.

To analyse this further, dimensionality reduction techniques like PCA and Isomap were used. While PCA didn't reveal much separation between those who stayed and those who left, likely attributed to its linear nature, Isomap was more successful. The Isomap projection made it clearer that there were distinct groups within the dataset. This was important in suggesting that different types of employees were leaving for combination of reasons, rather than attrition being caused by a single factor. However, *left* (the target varible) was not removed from the dataset before conducting PCA and the Isomap analysis, this may have introduced label leakage into the visualisation, exaggerating separation between classes.

This idea was further backed up through KMeans clustering, which identified three different clusters of employees who left. The first group were high-performing but clearly overworked—employees who had high evaluation scores and workloads, but still ended up leaving. The second group were unhappy and underperforming, which might point to a lack of motivation or poor fit within the company. The third cluster included employees who had both high satisfaction and performance, but still left, potentially due to external opportunities or a lack of internal growth. This kind of segmentation shows just how important it is to understand subgroups in the data rather than treating all employee churn as one problem.

The radar plots were then revisited using these clusters, which helped to visualise the differences between each group. This made the analysis much more compelling and actionable, since

it suggested tailored solutions for each type of employee. For example, overworked employees might benefit from workload management, while disengaged employees might need better support or clearer progression paths.

While the dataset lacks depth in key areas like employee motivation, external influences, and demographics, the analysis makes good use of what's available. The visualisations, especially when combined with clustering, go beyond surface-level summaries and start to build a deeper understanding of the problem. However, to take the analysis to the next level, the dataset would need to be expanded. Including feedback data (like survey responses or exit interviews) and contextual business data would allow for a more accurate and actionable understanding of why employees leave.

3. Data Preparation

Ghouzam's original report presents only a superficial level of data preparation, overlooking several essential steps required for trustworthy modelling. Although no missing values were reported, the dataset contains a large number of duplicate rows — a critical issue that was neither identified nor addressed. These duplication can inflate model confidence, particularly when identical cases with the same outcome dominate the data. This can lead to overfitting and poor generalisation (Raza et al., 2022).

Outlier detection was also omitted. While tree-based models like Random Forest and XGBoost are generally easily scalable, they can still be influenced by extreme values when those values reflect non-representative or duplicated behaviour. For example, rare patterns in *average_monthy_hours* or *number_project* may cause unstable splits or introduce misleading decision thresholds. Studies have shown that handling such outliers improves model reliability and decision boundary consistency (Marín Díaz et al., 2023).

Feature scaling was applied appropriately through standardisation — essential for PCA and KMeans — but feature selection lacked any justification. All variables were used without assessing predictive contribution, redundancy, or multicollinearity. Additionally, skewness was not assessed, and no transformations were applied to correct non-normal distributions.

Interaction effects were also overlooked. Relationships such as *satisfaction_level* x *last_evaluation* or *number_project* x *time_spent_company* could provide more insight than individual variables alone. In response, I created an *eng_idx* , a composite of satisfaction, evaluation, and workload, to reflect employee engagement more holistically.

Finally, the report did not include any formal train-test split. Except for the use of 10 fold cross-validation with , it is unclear how model performance was evaluated. This absence raises the risk of data leakage and undermines the credibility of reported results. Proper data splitting is fundamental to ensure models generalise beyond the training data (Shafie et al., 2024).

4. Modelling

Ghouzam's report relies on Decision Tree and Random Forest classifiers — both valid choices for binary classification tasks and well-suited to scenarios requiring model transparency. However, no rationale is given for their selection, nor are alternatives considered. Including a Logistic Regression model as a baseline would have provided a meaningful benchmark to evaluate whether these tree-based models genuinely improved performance.

More concerning is the lack of clarity around evaluation strategy. The report doesn't specify how the data was split — there's no mention of a training/test ratio or whether stratified sampling was used. Although 10-fold cross-validation is briefly mentioned for Random Forest, this is not applied consistently across models, leaving the possibility of data leakage and inflated performance metrics.

The Random Forest is reported to achieve 99% accuracy, which strongly suggests overfitting. Without context from precision, recall, or F1-score — especially in an imbalanced classification problem — this value is misleading. A high accuracy may simply reflect the model's tendency to predict the dominant class. This indicates a broader issue: the report uses only accuracy as an evaluation metric, which is insufficient in isolation. A confusion matrix or class-wise performance would have revealed how well the model captures attrition cases.

The Decision Tree visualisation is a positive inclusion — it clearly shows where splits occur and highlights important features like *satisfaction_level* and *time_spent_company*. However, no equivalent transparency is offered for the Random Forest. Since ensemble models are harder to interpret, the use of a model-agnostic explainer like LIME or SHAP would have been more appropriate here (Byeon, 2023).

In my improved analysis, I selected XGBoost, a widely used gradient boosting model that performs particularly well on structured HR datasets (Hambali et al., 2024). It offers a strong balance of accuracy and interpretability, especially when paired with SHAP to visualise global and local feature contributions. Importantly , I ensured that models were trained on a properly split dataset and evaluated using multiple metrics including F1-score, precision, and recall. Without these steps, it's difficult to trust any claims about model performance or business impact.

XGBoost Model

When developing my own model, I focused on addressing the main limitations of the previously mentioned study. This consisted of removing duplicates from the dataset to remove their impact from the model, introducing an 80-20 training/test split and regularisation techniques to reduce overfitting, address the class imbalance and evaluate the models on unseen data. Along with including more performance metrics; precision, recall, F1-score and area under the curve (AUC) to give a better evaluation of the model. Hyperparameter tuning is also used to improve the performance of the model.

I decided to develop an XGBoost model with LIME for explanation, as XGBoost, being a powerful gradient boosting method, extends decision tree-based models and often outperforms Random Forests in structured data scenarios (Ribeiro et al., 2016; Lundberg & Lee, 2017). I have incorporated LIME to clearly explain how the model came to its classifications to make it easier for HR individuals to interpret the results.

Table 1: Performance Metrics for XGBoost Model

Metric	Scores
Accuracy	0.982
Precision	0.992
Recall	0.902
F1-Score	0.945

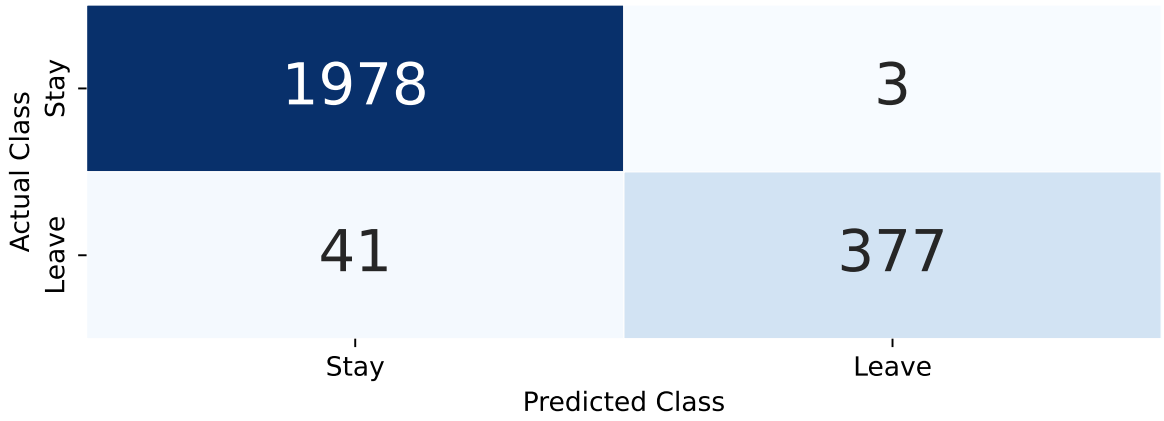

Figure 1: Confusion Matrix for XGBoost Model

The XGBoost model showed strong performance across all key evaluation metrics. On the test set, it achieved an accuracy of 98.2, which already indicates that the vast majority of predictions were correct. However, looking beyond accuracy, the precision score of 0.992 is very high, it means that when the XGBoost predicts a positive case, it is 99.2% likley to be correct. This is important in scenarios where false positives can waste resources.

Recall was also high at 0.902, which suggests the XGBoost is successfully identifying most of the actual positive cases. That said, there were still 41 false negatives (actual positives misclassified as negatives), as seen in the confusion matrix (Figure 1). While this number is relatively small, it could still be a concern depending on the context—especially if missing positive cases leads to missed opportunities or losses.

Figure 2, the ROC curve, backs up the XGBoost's solid classification ability. The AUC score of 0.987 shows that the XGBoost does a great job distinguishing between the two classes,

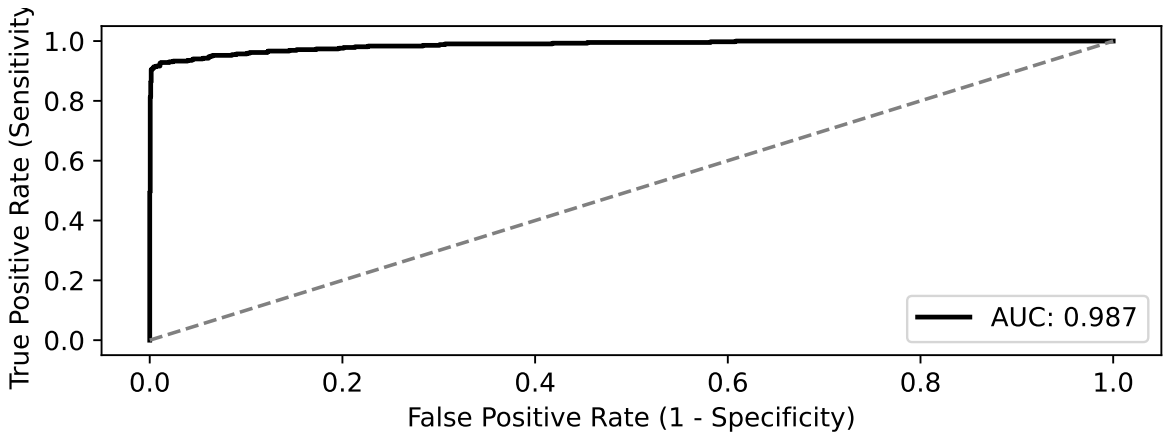

Figure 2: ROC Curve for XGBoost Model

even under different thresholds. This, combined with the F1 score of 0.945, suggests that the XGBoost handles class balance well and is not overly biased toward the majority class. Despite the high performance metrics which traditionally could indicate potential overfitting, due to the implementation of L1 + L2 regularisation, a train/test split, and 10 cross-fold validation, these metrics can be trusted allowing this model to be deployed within a HR environment for employee churn decisions.

5. Evaluation

Currently, the project is not suitable for deployment. While the initial analysis shows promise in identifying key drivers of employee churn, several fundemental gaps prevent it from being implemented in a real-world HR setting. Most notably, the model has not been validated against operational or unseen data. Without testing on external scenarios or real-world systems, there is a significant risk that the conclusions will fail to generalise, especially in fast-changing or department-specific organisational contexts.

The dataset itself, although informative, has a narrow scope. It excluded important features such as external labour market indicators, voluntary resignation reasons, and employee sentiment. These elements are often key to understanding workforce dynamics and employee decision-making. Without qualitative data, like exit interviews or engagement surveys, the model lacks the depth to provide actionable insight beyond correlation. This omission reduces the model's value for long-term strategic planning.

From an analytical standpoint, the project correctly identified key predictive features, but the overall modelling approach lacked rigour. No precision, recall, or F1-score metrics were reported, which makes it difficult to judge how well the model handles the minority class (employees who leave). There is also no evidence of hyperparameter tuning or regularisation,

essential for avoiding overfitting and ensuring consistent performance. Fairness was not considered either; for example, no subgroup analysis was conducted to assess whether the model performs equally well across departments, seniority levels, or salary bands. Given the ethical implications of using predictive models in HR, the absence of bias testing is a significant concern (Byeon, 2023).

Visual outputs were present but incomplete. While they helped communicate feature relevance and model logic, they lacked comparative or temporal elements such as time-series trends or benchmarking against industry norms. These would be vital for contextualizing insights and tracking change over time. Additionally, model interpretability was limited. Although feature importance was displayed, no tools such as LIME or SHAP were used to explain individual predictions. Without interpretable outputs, stakeholders may struggle to trust or act on the model's decisions, making adoption unlikely.

For example within my XGBoost, Figure 3 shows how LIME can enhance transparency by breaking down the individual contribution of features in a specific observation. Within this observation, the XGBoost predicted a probability of 97% that the employee would leave, driven largely by a high time spent at the company (>4 years), no promotions in the last 5 years, and satisfaction levels being below 0.66. Additional factors that contributed to a leave prediction were having to work more than 199 hours a month on average and a number of projects above 4. Interestingly, the new feature created *eng_idx* which represents an engagement index combining *satisfaction_level*, *last_evaluation* and *average_monthly_hours* has low importance, validating the previous study not doing any feature selection and using the standard features. Another area that the XGBoost has an unexpected result is predicting people to leave if they have a last evaluation of above 0.87, this could be attributed to competitors hiring the high performing individuals or them looking for opportunities with better compensation packages. However, LIME only explains one observation and this is where feature importance is needed, correctly included in Ghouzam's project although it could've included F-score to quantify the importance instead of only showing relative importance.

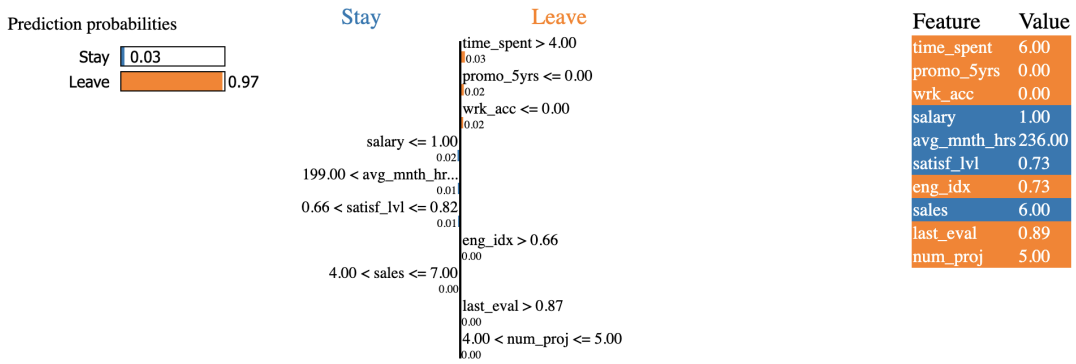

Figure 3: LIME Explanation

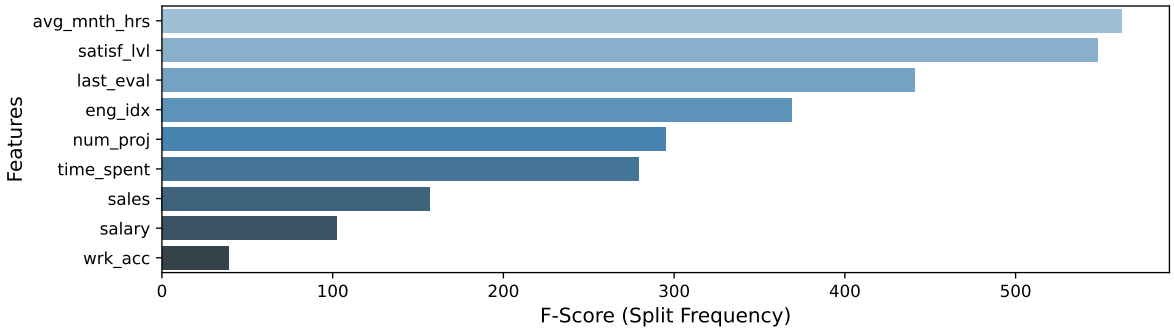

Figure 4: Feature Importance for XGBoost Model

Figure 4 shows global feature importance based on XGBoost's F-score, which counts how often each feature is used to split the data across all trees in the XGBoost. High F-scores for *satisfaction_level*, *average_montly_hours*, and *last_evalulation* indicate these features are frequently used in decision rules, suggesting strong individual predictive power. The engineered feature *eng_idx* also ranks highly, confirming its added value as a standalone input. It's important to note that F-score reflects only how often a feature is used; not the size of its effect or any interactions with other features. Therefore, while useful for identifying influential predictors, this metric does not account for how features might interact in the model.

6. Deployment

Based on the analysis, Ghouzam's model is not yet ready for deployment, but several clear and feasible steps could bring it closer to operational use. The most immediate priority is planning integration with existing HR systems. At present, there is no deployment framework in place — no API, no dashboard, and no defined output format for HR teams. For the model to have practical impact, it must deliver insights in an accessible, actionable way. A basic interactive dashboard, built using tools like Streamlit or Power BI, could serve as a front-end to surface individual attrition risks and feature-level explanations.

A second major gap is the absence of a monitoring and retraining strategy. HR data changes over time due to seasonal trends, policy changes, and broader economic shifts. Without a mechanism to track model drift and performance decay, the model risks becoming obsolete. Establishing a structured retraining pipeline, supported by a clear ETL (Extract, Transform, Load) process, would enable continuous updates as new employee data becomes available. Key evaluation metrics — including F1-score, recall, and precision — should be tracked on a live dashboard to monitor real-time performance.

To support ongoing use, a shift toward a prescriptive analytics model is recommended. This would enable the system not just to flag high attrition risk, but also to suggest interventions

such as bonus adjustments, role transfers, or engagement strategies. However, this requires a high degree of interpretability. Currently, the model lacks explanation tools like LIME or SHAP, making it difficult for HR professionals to understand or act on individual predictions. Without interpretability, stakeholders may distrust the outputs, a key barrier to adoption in any decision-support system.

Finally, ethical and regulatory considerations must be addressed. The model does not appear to incorporate any fairness assessments or bias audits, and there is no discussion of data privacy or GDPR compliance. These omissions are especially important when using personal employee data in predictive systems. A deployment plan must include bias mitigation strategies, subgroup performance checks, and a clear governance framework for responsible AI use in HR settings.

Link to Github Repository = BEE2031 Introduction to Business Analytics