Employee Attrition: Prediction and Explainability Report (XAI)

Alejandro Martínez de Guinea García
Master in Artificial Intelligence
Pontifical Comillas University (ICAI)

December 31, 2025

Abstract

This report presents the findings of the final project for the course Ethics and Explainability in Artificial Intelligence. It details the development of a Machine Learning model designed to predict employee turnover. Using **XGBoost** and Explainable AI (XAI) techniques such as SHAP and LIME, important factors influencing attrition were identified. Key recommendations for HR strategies to mitigate turnover are provided based on the insights derived from the model.

Contents

1 Introduction and Business Context

1.1 Problem Description

Employees are the most valuable asset within a company, but keeping them is hard due to various factors such as job dissatisfaction, lack of growth opportunities, and work-life imbalance.

Despite having plenty of data, traditional Machine Learning models often act like "black boxes". They might correctly flag an employee who is about to quit, but they won't explain *why*. This leaves managers in the dark, as they don't know the reasons of someone quiting, so they can't take effective action.

In this project, two main challenges are addressed:

1. **Catching it early**: We need a model smart enough to spot the subtle signs of a leaving employee.
2. **Opening the Black Box**: We don't just want a prediction; we want the explanation. We need to translate complex algorithms so HR can actually intervene.

1.2 Motivation

In the business world, employee turnover is a costly issue. It includes direct costs (like recruiting and training new hires) and indirect costs (like lost productivity and morale of remaining staff). Reducing turnover can lead to significant savings and a more stable workforce.

Most companies rely on "Exit Surveys" — asking people why they are leaving after they have already quit. That is too late. This project aims to shift the strategy from reactive to proactive. The goal is to detect at-risk employees before they decide to leave, identifying the root causes and enabling targeted interventions.

1.3 Stakeholders

Who is this tool for?

- **HR Managers**: They need to understand the *why* behind attrition. The model will provide them with actionable insights to design better retention strategies.
- **Team Leaders**: They can use the insights to improve team dynamics and address specific issues that may lead to dissatisfaction.
- **The Employees**: Ultimately, happier employees lead to lower turnover. By addressing the root causes of attrition, the company can create a better work environment.

2 Data Analysis and Preprocessing

2.1 Dataset Overview

The study uses the *IBM HR Analytics Employee Attrition* dataset, which contains 1,470 observations representing a diverse workforce. The features provide a comprehensive view of the employee profile, ranging from demographic details (e.g., Age) to job-specific metrics (e.g., Job Role, Overtime) and financial indicators.

The target variable for the predictive task is `Attrition` (Yes/No). An initial exploratory analysis revealed a significant class imbalance, with only $\sim 16\%$ of the samples belonging to the positive class (Attrition = Yes). This distributional skewness led to specific modeling strategies, as standard algorithms tend to bias predictions toward the majority class to maximize crude accuracy.

2.2 Data Cleaning and Noise Reduction

Prior to modeling, a cleaning process was implemented to remove non-predictive noise and prevent model bias:

- **Removal of Constant Values**: The features `StandardHours`, `Over18` and `EmployeeCount` were excluded from the dataset, since these variables exhibit zero variance across all observations and, therefore, possess no discriminative power.
- **Exclusion of Identifiers**: The variable `EmployeeNumber` was dropped. Administrative identifiers do not contain behavioral information, and their inclusion carries the risk of the model memorizing specific IDs rather than learning generalizable patterns.

2.3 Feature Encoding

To render categorical data compatible with the XGBoost algorithm, variables such as `Department` and `JobRole` were transformed. One-Hot Encoding was selected over Label Encoding. This approach prevents the algorithm from inferring spurious ordinal relationships (e.g., assuming Department 2 is mathematically "greater" than Department 1), thereby preserving the true categorical nature of the features.

2.4 Addressing Multicollinearity for Explainability

A critical challenge identified during the correlation analysis was the presence of Multicollinearity — redundancy between features. Strong linear relationships were detected in specific pairs:

- **JobLevel vs. MonthlyIncome**: Correlation > 0.90.
- **TotalWorkingYears vs. JobLevel**: Significant positive correlation.
- **Age vs. several features**: Moderate correlations with many different variables, such as `YearsAtCompany` or `MonthlyIncome`.

Justification for Feature Selection: While tree-based models can handle collinearity for prediction purposes, it poses a severe obstacle for Explainability (XAI). When two features carry identical information, interpretation algorithms (like SHAP) split the importance attribution between them, diluting the perceived impact of each. To ensure clear and actionable insights, redundant variables (e.g., `JobLevel`) were removed in favor of their more interpretable counterparts (e.g., `MonthlyIncome`). This strategy forces the model to consolidate importance on unique drivers, resulting in sharper explanations.

3 Methodology

This section outlines the evaluation framework, the benchmarking process, and the algorithmic selection criteria utilized to address the attrition prediction problem.

3.1 Evaluation Framework: F1-Score over Accuracy

Given the pronounced class imbalance ($\sim 16\%$ positive class), standard Accuracy was deemed an unsuitable metric. A model predicting the majority class (Stay) for every employee would achieve an accuracy of 84%, yet it would fail to identify any potential leavers, rendering it useless for retention strategies.

Consequently, the F1-Score was selected as the primary performance indicator. Specifically, the optimization focused on the **F1-Score for the Positive Class (Class 1: Attrition = Yes)**. This metric represents the harmonic mean of Precision and Recall, ensuring that the model is penalized both for missing leavers (False Negatives) and for raising too many false alarms (False Positives).

3.2 Baseline Benchmarking

To establish a minimum performance threshold, trivial baseline models were evaluated before training complex algorithms:

- **Constant-0 Model (Predicts "No" for everyone):** Resulted in an F1-Score of 0.0.
- **Constant-1 Model (Predicts "Yes" for everyone):** Resulted in an F1-Score of 0.28.

These results established that any viable machine learning model must significantly exceed an F1-Score of 0.28 to demonstrate predictive value beyond random guessing.

3.3 Algorithm Selection: Random Forest vs. XGBoost

Two primary ensemble algorithms were tested to surpass the baseline: Random Forest and XGBoost.

Initial Approach: Random Forest

A Random Forest model was initially implemented, achieving a robust performance with an F1-Score of 0.36, increasing significantly the baseline capability. However, despite its predictive power, the model presented significant challenges regarding Explainability (XAI). The resulting SHAP explanations were diffuse and lacked clarity.

Note: The Random Forest model is also available in the project repository in case further analysis is desired.

Final Selection: XGBoost

Subsequently, an XGBoost (Extreme Gradient Boosting) model was developed. The quantitative performance was significantly higher (F1-Score ≈ 0.45), almost doubling the baseline model. Also, the qualitative quality of the explanations improved significantly.

XGBoost builds shallow trees sequentially to correct previous errors. This boosting approach tends to be more selective with features, resulting in sharper, more distinct SHAP values.

4 Model Explainability and Interpretation (XAI)

This section details the application of Explainable AI (XAI) techniques to interpret the model's predictions. The analysis follows a structured framework: Global Interpretation (SHAP), Local Interpretation (LIME), and a Reliability/Sanity Check (Permutation Importance).

4.1 Phase 1: Analysis of the Full Model (Correlated Features)

Initially, the interpretation was performed on the model trained with the complete feature set, including highly correlated variables such as `Age`, `JobLevel`, and `TotalWorkingYears`.

4.1.1 Global Explanation (SHAP)

The SHAP Beeswarm plot (Figure 1) was generated to identify the macro-level drivers of attrition.

- **Dominant Factor**: `OverTime` appeared as the primary driver.
- **Demographics**: `Age` was identified as a top-tier feature. The SHAP values indicated a strong inverse relationship: younger employees (lower age values) showed significantly positive SHAP values, pushing the prediction towards "Attrition = Yes".
- **Financial Factors**: `MonthlyIncome` and `StockOptionLevel` were also prominent, with lower salaries and no stock options correlating with higher attrition risk.

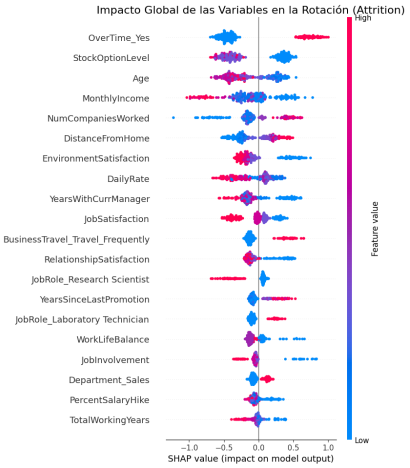

Figure 1: SHAP Summary Plot (Full Model).

4.1.2 Local Explanation (LIME)

An individual employee case was analyzed using LIME to understand specific predictions. This was the case of the employee that the model predicted with the highest risk of attrition ($\sim 97\%$). The LIME explanation revealed:

- **Key Drivers**: The primary contributors to the high attrition risk were positive `OverTime`, low `MonthlyIncome` and no `StockOptionLevel`.
- **Counterintuitive Findings**: Interestingly, despite having most factors indicating high risk, the employee had the highest score for `JobSatisfaction`.

With these findings, it is clear that HR should take action on employees like this one, who are engaged with their work but are at high risk of leaving due to workload and compensation issues.

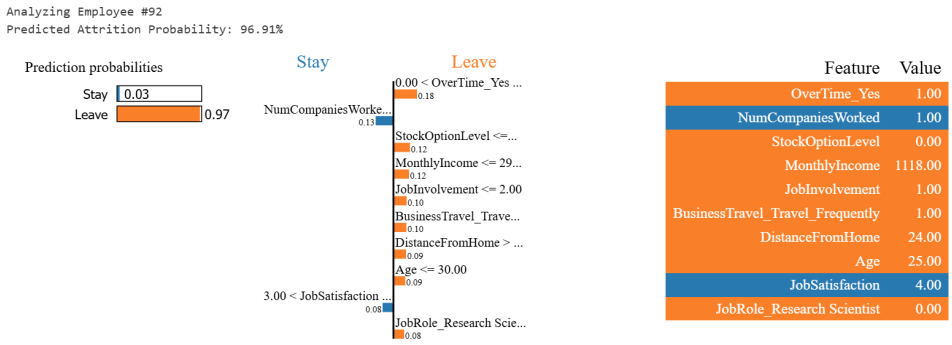

Figure 2: LIME Explanation for High-Risk Employee.

4.1.3 Sanity Check (Permutation Importance)

To validate the SHAP findings, a Permutation Importance test was conducted. This method involves randomly shuffling each feature and measuring the resulting drop in model performance (F1-Score for the positive class).

First, a permutation importance was calculated for the main feature: `OverTime`.

By randomly shuffling the values of the `OverTime` column (thereby breaking the link between the feature and the target), the model's F1-Score for the positive class was re-evaluated. The observed performance drop was substantial (from $\sim$ 0.45 to $\sim$ 0.3, meaning a decrease of about $\sim$ 30%), confirming that `OverTime` is indeed a critical driver of attrition, as indicated by SHAP in Figure 1.

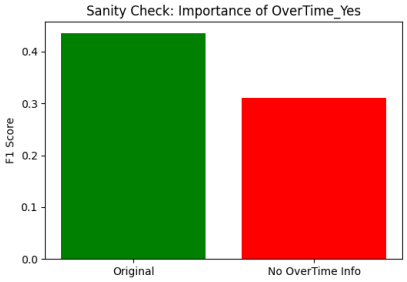

Figure 3: Permutation Importance on `OverTime`.

Next, the `permutation_importance` function from `scikit-learn` was used to compute the importance scores for all features. The result was the following:

The Permutation Importance plot (Figure 4) revealed that while `OverTime` remained the most critical feature, `Age` completely disappeared from the top important features. This was unexpected, as SHAP had indicated that `Age` was a vital predictor. This discrepancy raised concerns about possible presence of multicollinearity affecting the interpretability of the model. This is because if `Age` is correlated with other features, the model might be relying on those proxies instead, surviving the permutation of `Age` without a significant drop in performance.

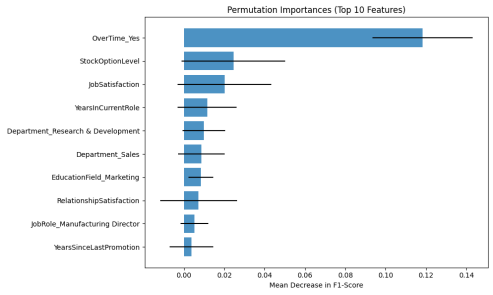

Figure 4: Permutation Importance (Full Model).

4.2 Phase 2: Analysis of the Reduced Model (Selected Features)

To resolve the inconsistency and obtain a trustworthy explanation, a second model was trained after removing the redundant correlated variables (`Age`, `JobLevel`, `TotalWorkingYears`). The XAI framework was re-applied.

4.2.1 Global Explanation (SHAP)

With redundancies removed, the SHAP Summary Plot (Figure 5) provided a similar result, but with clearer attributions:

- **OverTime**: Confirmed as the absolute primary cause of attrition.
- **MonthlyIncome**: Without `JobLevel` and `Age` diluting its impact, Salary emerged clearly as the second most important factor. Also, points are less concentrated in the center, where the SHAP values are close to zero. This indicates that the model is more decisive when using `MonthlyIncome` alone.
- **StockOptionLevel**: Remained as a significant predictor, however with less impact.

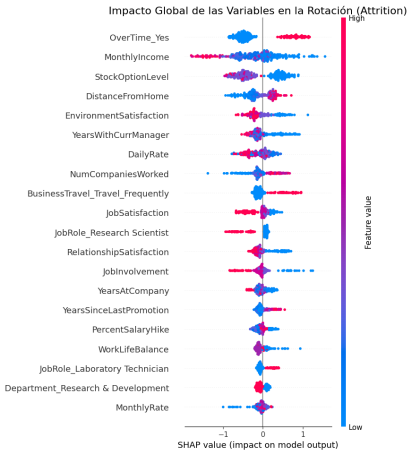

Figure 5: SHAP Summary Plot (Reduced Model).

As we can see, `Age` is no longer present, as it is now represented through its proxies.

4.2.2 Local Explanation (LIME)

The same high-risk employee case was re-evaluated using LIME on the Reduced Model. In this case, the model defined a probability of attrition of $\sim 98\%$. The LIME explanation revealed:

- **Key Drivers**: The main contributor to the high attrition risk surprisingly was `MonthlyIncome` instead of `OverTime`. Since `MonthlyIncome` was correlated with `Age` and `JobLevel`, in the Full Model its importance was diluted. The high impact makes sense, since the employee's salary is one of the lowest in the company, and there is no way to distinguish between overtimes as it is a binary variable.
- **Counterintuitive Findings**: Again, despite having most factors indicating high risk, the employee had the highest score for `JobSatisfaction`, and a very high score for `EnvironmentSatisfaction`.

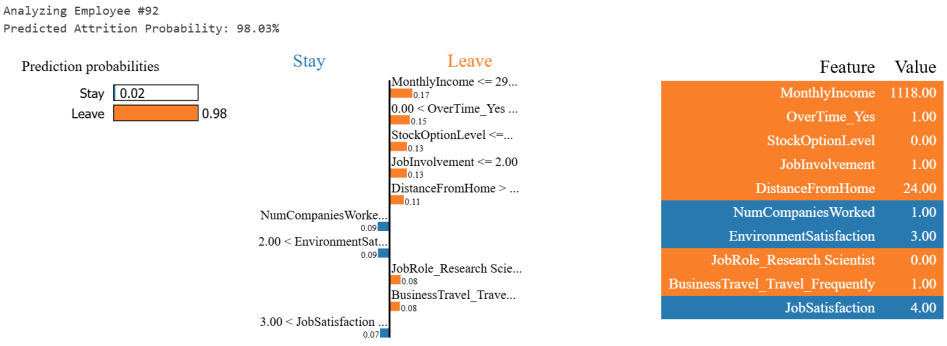

Figure 6: LIME Explanation for High-Risk Employee (Reduced Model).

4.2.3 Sanity Check (Permutation Importance)

Finally, the Permutation Importance test was repeated on the Reduced Model. As well as before, the importance of `OverTime` was first validated by shuffling its values and measuring the drop in F1-Score for the positive class.

The results confirmed that `OverTime` is still an essential factor for the model, with a significant performance drop (from ~ 0.47 to ~ 0.28, meaning a decrease of about $\sim 40\%$) when its values are shuffled. This reaffirms its dominant role in predicting attrition.
In fact, the effect is even larger than the one with the Full Model, which makes sense due to the diluted effect it had with the presence of correlated variables.

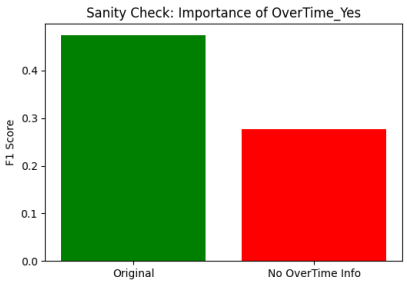

Figure 7: Permutation Importance on `OverTime` (Reduced Model).

Then, the full Permutation Importance plot was done. This plot confirmed that `OverTime` remained the most critical feature, followed by financial factors such as `MonthlyIncome` and `StockOptionLevel`. This aligns perfectly with the SHAP results, confirming the reliability of the explanations after addressing multicollinearity.

Surprisingly, `JobSatisfaction` appears as the fourth most important feature, despite not being as prominent in the SHAP plot. This suggests that while `JobSatisfaction` may not have a strong average effect across all employees, it could be crucial for specific subgroups, warranting further investigation. Also, it is a feature that comes up inmediately to mind when thinking about retention strategies.

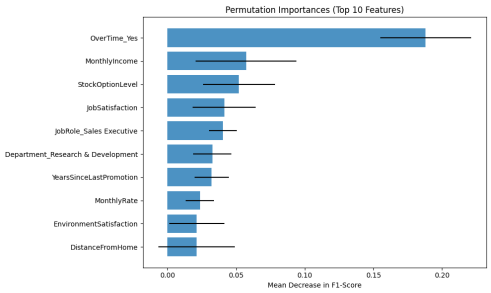

Figure 8: Permutation Importance (Reduced Model).

Now, the results are consistent between SHAP and Permutation Importance, confirming the validity of the interpretations.

5 Conclusions and Recommendations

Summary of Findings

This project successfully validated that Explainable AI (XGBoost + SHAP) can transform raw data into strategic HR assets. The analysis identified **OverTime** as the dominant driver of attrition, indicating that burnout (rather than dissatisfaction) is the primary cause of turnover. Conversely, financial factors like `MonthlyIncome` and `StockOptionLevel` act as effective retention mechanisms.

Methodological Insight: The Value of Correlation

A key conclusion involves the dual role of correlated features. While removing redundant variables (like `Age`) was essential to obtain alignment between SHAP and Permutation Importance results, retaining them in the *Full Model* proved valuable for contextual analysis. Variables such as `Age`, despite being redundant, provide insights into demographic patterns of attrition. For instance, younger employees are more prone to leave, likely due to lower salaries, less experience or higher career mobility, which are captured through correlated proxies.

Therefore, a balanced approach is recommended: use the *Reduced Model* for reliable feature importance and the *Full Model* for demographic insights.

Strategic Recommendations

Based on the data, two immediate actions are proposed:

1. **Mitigate Burnout**: Since `OverTime` is the strongest predictor of churn, reviewing workload distribution offers the highest ROI for retention.
2. **Targeted Increases**: For employees identified as high-risk (e.g., low `MonthlyIncome` and no `StockOptionLevel`), targeted salary adjustments and stock options can effectively reduce attrition risk.

Future Work

Future research could focus on the following areas:

- **Temporal Dynamics**: Incorporate time-series data to capture how attrition risk evolves.
- **Wider Data Sources**: Integrate qualitative data (e.g., employee surveys) for a holistic view. This could look into whether attrition was a personal decision or influenced by external factors, such as market conditions or job offers.