Detection of Distressed Employees Using Isolation Forest and Explainable AI

@ashutosh

April 6, 2025

Abstract

This report presents a methodology for detecting distressed employees by leveraging unsupervised anomaly detection and model interpretability techniques. An Isolation Forest is used to flag employees whose behavioral and performance metrics deviate from the norm. To provide an interpretable explanation of the anomalies, SHAP (SHapley Additive exPlanations) values are computed, with particular focus on the absolute values of negative contributions. A sensitivity factor is incorporated to flag even slightly distressed employees by adjusting the detection threshold. The methodology is summarized in a consolidated database (`af_db`) which contains key indicators for each flagged employee. This report details the theoretical background of the methods, the processing pipeline, and a discussion on why the approach is effective.

1 Introduction

Employee well-being is critical for maintaining productivity and ensuring a positive workplace environment. Detecting early signs of distress can enable organizations to intervene proactively. In this work, we introduce a system that:

- Utilizes an **Isolation Forest** for anomaly detection, where more negative anomaly scores indicate higher distress.
- Employs **SHAP** values to provide interpretability by identifying which features most strongly contribute to the detection.
- Incorporates a **sensitivity factor** to adjust the model's responsiveness and capture even slightly distressed employees.
- Summarizes the key findings in a dataset (`af_db`) for further analysis.

2 Theoretical Background

2.1 Isolation Forest

Isolation Forest is an unsupervised anomaly detection algorithm that isolates observations by randomly selecting a feature and then randomly selecting a split value between the minimum and maximum values of the selected feature. Its key properties include:

a. **Random Partitioning:** The algorithm builds an ensemble of trees (isolation trees) where each tree is created by recursively partitioning the data using randomly chosen features and split values.

b. **Path Length:** The number of splits required to isolate a data point, called the *path length*, tends to be shorter for anomalies because they are few and different.

c. **Anomaly Score:** The anomaly score $s(x, n)$ for an observation x in a dataset of size n is given by:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}},$$

where $E(h(x))$ is the expected path length for x and $c(n)$ is the average path length of unsuccessful searches in a binary tree, approximated as:

$$c(n) \approx 2H(n-1) - \frac{2(n-1)}{n},$$

with $H(i)$ being the ith harmonic number. A lower (more negative) anomaly score indicates a higher likelihood of distress.

2.2 SHAP (SHapley Additive exPlanations)

SHAP values provide a way to interpret model predictions by quantifying the contribution of each feature. Based on concepts from cooperative game theory, SHAP assigns each feature an importance value such that:

- The sum of the SHAP values equals the difference between the model prediction and a baseline value.
- They provide local interpretability, explaining individual predictions.

In our context, we focus on the *negative* SHAP values because these indicate the features that push the anomaly score downward (i.e., contribute to flagging an employee as distressed). The absolute values of these negative contributions are used to rank the features by their impact.

2.3 Sensitivity Factor

A sensitivity factor is incorporated into our model to adjust the threshold for flagging anomalies. By increasing the sensitivity factor:

- The system becomes more responsive to slight deviations from normal behavior.
- More employees with only modest changes in their feature values can be flagged as distressed.

Thus, a higher sensitivity factor leads to capturing a larger set of potentially distressed employees, even if the deviations are not as extreme.

3 Methodology

3.1 Data Preprocessing and Feature Engineering

The dataset contains 36 columns of employee-related data. For anomaly detection, we select a subset of numerical and encoded features that capture:

- Communication metrics (e.g., average and median daily Teams messages and emails sent).
- Work patterns (e.g., average work hours, median work hours, and work hour variability).
- Attendance and leave factors (e.g., Annual, Casual, Sick, and Unpaid Leave Impact Factors).
- Temporal factors (e.g., Days since joining, Onboarding Factor).
- Performance-related metrics (e.g., Total Decayed Reward Points, Decayed Vibe, encoded feedback).

Categorical features (e.g., `Onboarding_Feedback`) are mapped to numerical scales, and boolean features are converted to binary values. Missing values are imputed using the median, and the features are scaled using `StandardScaler`.

3.2 Anomaly Detection using Isolation Forest

The preprocessed feature matrix is used to train an Isolation Forest model. The model assigns an anomaly score to each employee:

- **Anomaly Score:** More negative scores indicate a higher degree of anomaly (and hence higher distress).
- **Sensitivity Factor:** By adjusting this parameter, the model can flag even slight deviations. A higher sensitivity factor results in more employees being identified as anomalous.

3.3 Explainability using SHAP

To interpret the anomaly detection results:

- SHAP values are computed for the Isolation Forest using a TreeExplainer.
- We focus on the negative SHAP values, as these highlight the features that contribute to an employee being flagged as distressed.
- A horizontal bar chart is generated for each anomalous employee, ranking features by the absolute value of their negative SHAP contributions.

3.4 Sample Graphs and verification

The following types of graphs can be plotted using available flagged employee database `af_db`. The following pie chart represents the major problems faced by the employee (problematic features for the flagged employee).

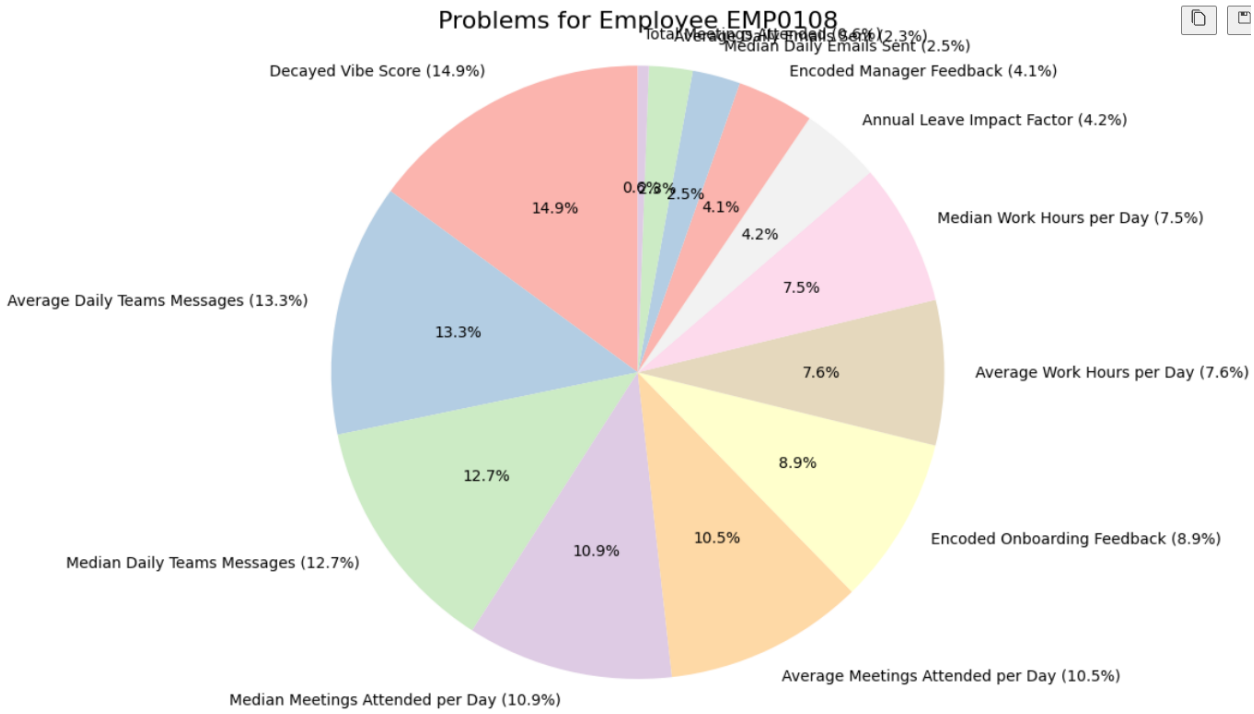

Figure 1: Test Results

This model output is cross-verified by using ChatGPT Reasoning mode applied on the smaller database of employees.

ChatGPTs prompt response: `Yes, your model's output aligns with the reasoning. Both my inference and the chart indicate that EMP0108's distress is primarily driven by factors such as a low decayed vibe score and challenges in the onboarding process, along with managerial feedback and communication workload issues. This suggests your anomaly detection approach is capturing the key stress indicators effectively.`

3.5 Summary Database (`af_db`)

A summary dataset, `af_db`, is created to consolidate key insights for each flagged employee. Its columns include:

Employee_ID: The unique identifier for the employee.

Problems: A JSON-formatted list of the top 5 features (with their absolute negative SHAP contributions) that most strongly indicate distress.

Other Problems: A JSON-formatted list of the remaining features with negative SHAP contributions.

Anamaly_Score: The anomaly score from the Isolation Forest (more negative indicates higher distress).

Average Work Hours: The average daily work hours, providing context on employee engagement.

Reward Factor: The total decayed reward points, reflecting recognition and motivation.

Performance Rating: The performance rating of the employee.

Vibe Factor: The decayed vibe score indicating overall employee sentiment.

4 Results and Discussion

The Isolation Forest model successfully isolates employees with anomalous behavior patterns. Key observations include:

- **Anomaly Scores:** Employees with very negative anomaly scores are flagged as highly distressed.
- **Negative SHAP Contributions:** The focus on negative SHAP values reveals which features (e.g., high *Annual Leave Impact Factor* or low *Average Daily Emails Sent*) drive the model's decision.
- **Sensitivity Adjustment:** The sensitivity factor allows the system to detect even slight deviations, enhancing its utility as an early-warning tool.
- `af_db` **Dataset:** This summary table provides a consolidated view of the most impactful features for each distressed employee, alongside contextual information like work hours and reward points.

This layered approach ensures not only that distressed employees are detected but also that the reasons behind the detection are transparent and actionable.

5 Conclusion

We have presented a robust and interpretable methodology to detect distressed employees using an unsupervised Isolation Forest model, complemented by SHAP-based explanations. The approach leverages a sensitivity factor to adjust the detection threshold, ensuring that even slightly distressed employees can be identified. The resulting summary database (`af_db`) consolidates key indicators and provides actionable insights for early intervention. The theoretical grounding in Isolation Forest and SHAP, coupled with the practical results, demonstrates that this method is both effective and interpretable.