What do we know about professionals that leave a job?

Jasper Wilson

Wotter

4th August 2023

What do we know about professionals that leave a job?

Abstract

This paper explores user trends on the popular employee engagement application - Wotter - in the time periods prior to leaving a job. Wotter collects data on 24 metrics, known as the 'Core24', that attempts to give insight into how users feel about different aspects of their professional lives. Wotter achieves this through asking questions to employees, and uses a machine learning algorithm to identify particular trends in the Core24. In this paper, we analyze these metrics and attempt to find out what leads employees to leave their current jobs. We use

various machine learning classification and regression models to accomplish this. Firstly, we explain the methodology behind our analysis and then display our results. Finally, we hypothesize our 'Theory of Quitting', which attempts to explain why employees leave a job and what happens leading up to that point. Although it must be reiterated that this theory requires greater scrutiny and testing to be verified.

Methodology

Participants

We used a sample of user data from the Wotter database containing over 250,000 survey questions answered.

Assessments and Measures

Using Python and useful libraries such as NumPy, Pandas, Seaborn and SciKitLearn. We performed exploratory data analysis and employed various machine learning based modeling techniques such as Decision Trees, Principal Component Analysis (PCA) and Correlation Analysis.

Machine Learning Application

Here we outline some of the machine learning methods used in this research and provide detail into how they were specifically applied.

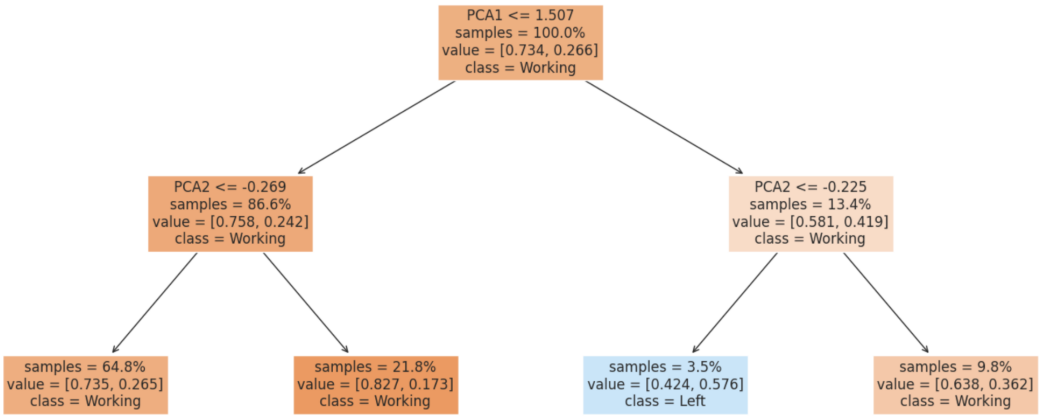
What do we know about professionals that leave a job?

Decision Tree Algorithm

Decision Trees[1] are a type of supervised machine learning algorithm that attempts to make predictions and/or categorize based on how previous questions about the data set are answered. These previous questions can be described as the decision nodes of the tree because the branch entered depends on how the question on the decision node is answered. This splitting of the data set continues until a leaf node is reached (a node that does not split the data set further and represents classification outcomes).

The decision tree used in this research was combined with Principal Component Analysis (PCA) (explained below).

Figure 1: Decision Tree Model for leaving employees where PCA1 = Principal Component 1 and PC2 = Principal Component 2.



This tree provides a means for classifying employees as still working or left the company. The ‘samples’ parameter represents the percentage of the data set that is being represented by that node. The ‘value’ parameter can be interpreted as probabilities of an employee still working or having left a company through conducting a random sample of that node. The ‘class’ parameter represents the classification result.

A common way of evaluating a model is through F1-score, Accuracy and a confusion matrix.[2] We used both of these to evaluate the decision tree model above, with the following results:

Figure 2: Classification Report and Confusion Matrix for Decision Tree/PCA model

	precision	recall	f1-score	support
False	0.75	0.98	0.85	7864
True	0.62	0.08	0.14	2838
accuracy			0.74	10702
macro avg	0.68	0.53	0.50	10702
weighted avg	0.71	0.74	0.66	10702

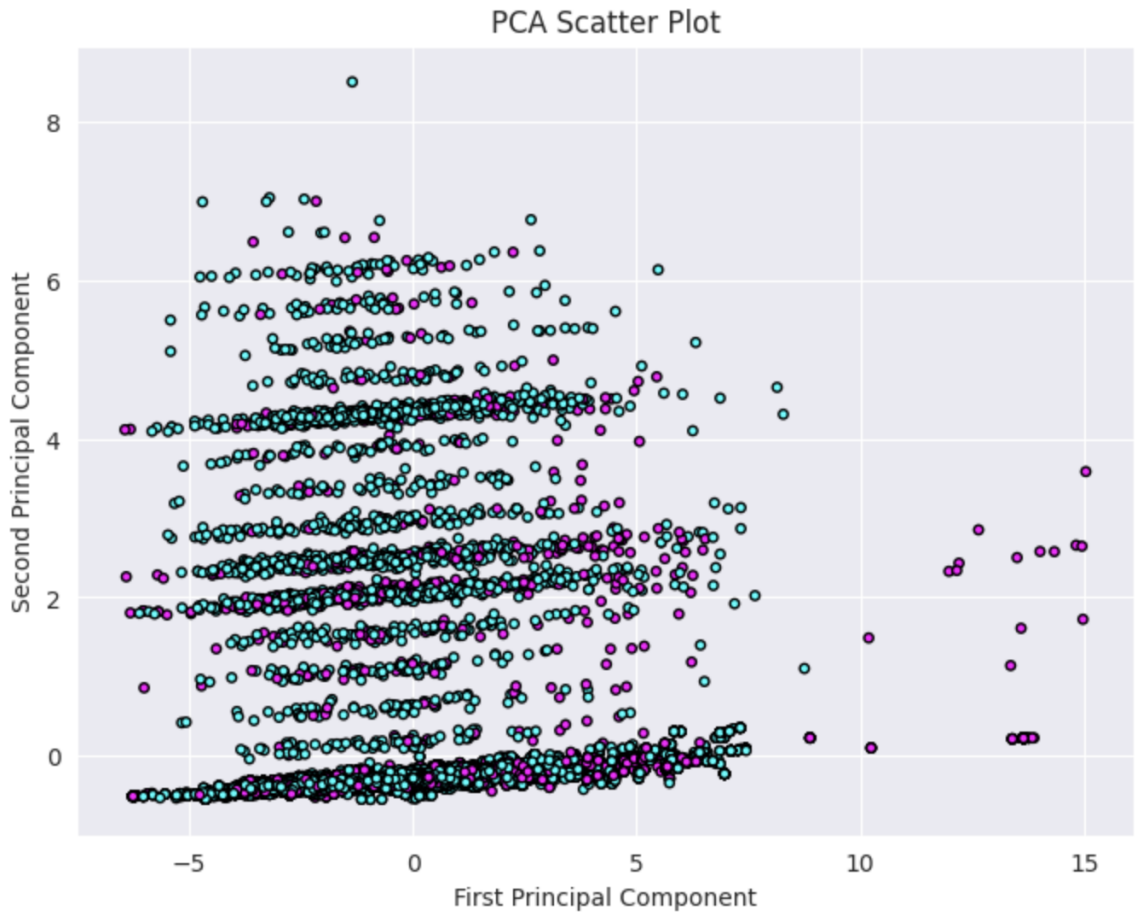
Predicted	False	True	All
Actual			
False	7723	141	7864
True	2607	231	2838
All	10330	372	10702

F1-scores are seen as the harmonic mean average of both Precision and Recall scores, it provides a metric for evaluating the accuracy of the model and is especially useful in situations where the data set is imbalanced (such as this dataset). The closer the F1-score is to 1.00, the better the model. In this case, the weighted average of the F1-score of 0.66 is indicative of 'mild' predictive power at best. This model must be refined before real-world deployment.

Principal Component Analysis (PCA)

Principal Component Analysis is a method of reducing the dimensionality of a data set by using eigenvectors and eigenvalues of data set covariance matrices to project the data onto dimensions that are orthogonal to each other[3]. The general idea is to find values for the Principal Components that maximize the variance of the data set to capture the most information about the data. For user data at Wotter, we used PCA to represent the whole data set in just 2 dimensions, shown in the scatter plot below.

Figure 3: PCA Analysis of Wotter User Data



This is a representation of a 31-dimensional data set in just 2 dimensions.

k-Nearest-Neighbours Data Imputing

k-Nearest-Neighbours (KNN) data imputing is a method of estimating missing values in a data set by finding the most similar data points to the one being imputed, then estimating the data values by taking the mean of those similar data points or using Euclidean distance[4]. This method has its limitations, such as introducing bias into the data set - however it can provide better estimates than other imputing strategies, such as using the Mean Average.

Results

We may be able to draw the following conclusions from our investigation:

Users could experience gradual wellbeing deterioration for up to 60 weeks before leaving a job.

Here we present results of our data analysis that support our hypotheses:

We split the different Core24 metrics into two distinct groups within leavers - 'Potential Triggers' and 'Deteriorators'. Potential Triggers are defined as metrics that fall rapidly below the employed mean average (≈ -0.5$\boldsymbol{\sigma}$ in a 5 week time period or less) before the Deteriorators begin to gradually decline over a time period of approximately 40-60 weeks, at which point the employee leaves the company. Potential Triggers tended to occur ≈20 weeks before the deterioration pattern was seen in the other metrics.

Graphical representations of the relevant Core24 metrics can be seen in Figures 4-21.

Figure 4: Values

Figure 5: Self Worth

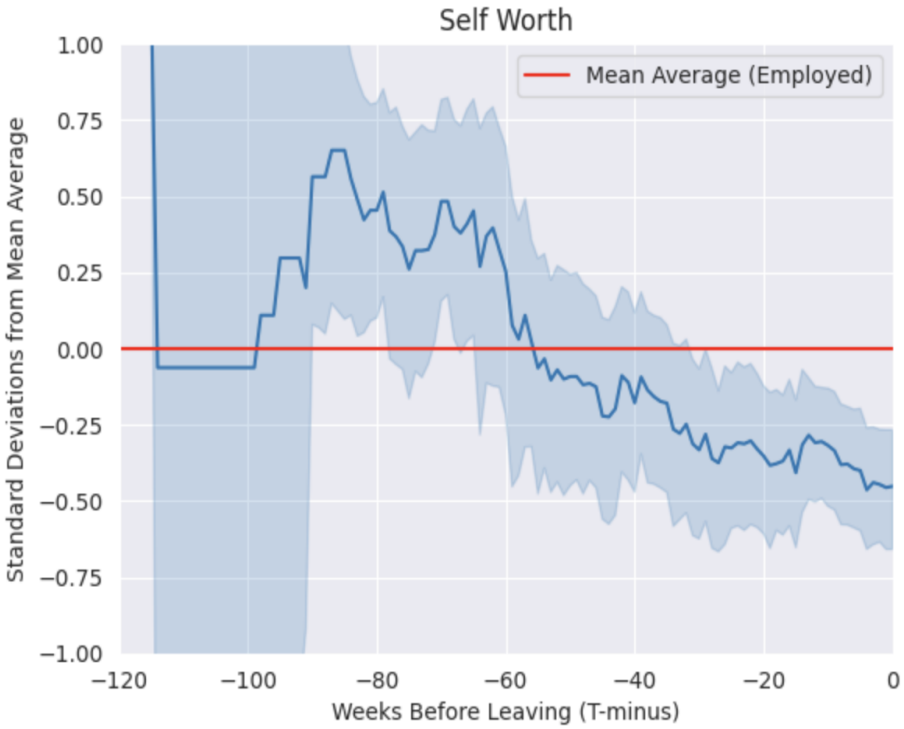


Figure 6: Future Progression

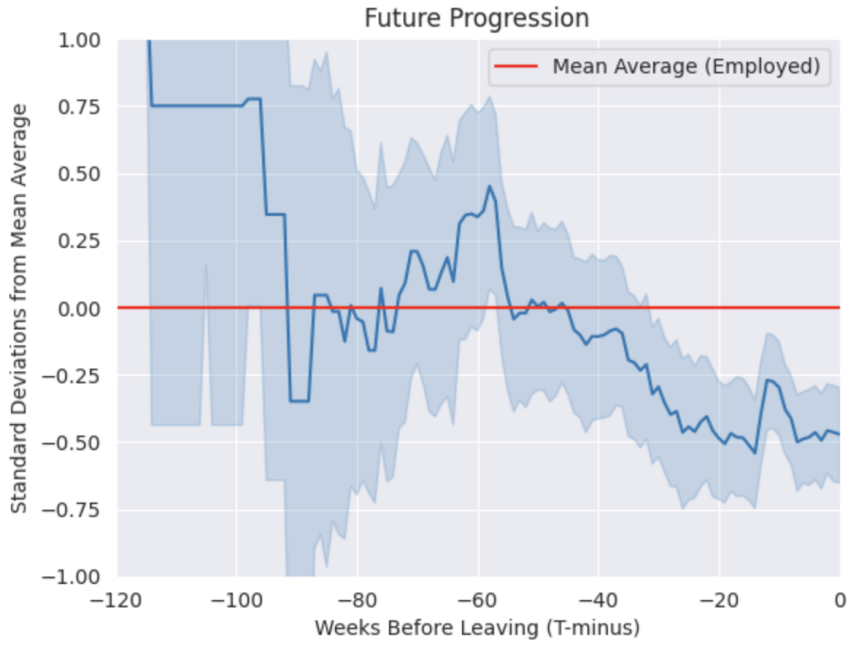


Figure 7: Pride

Figure 8: Transparency

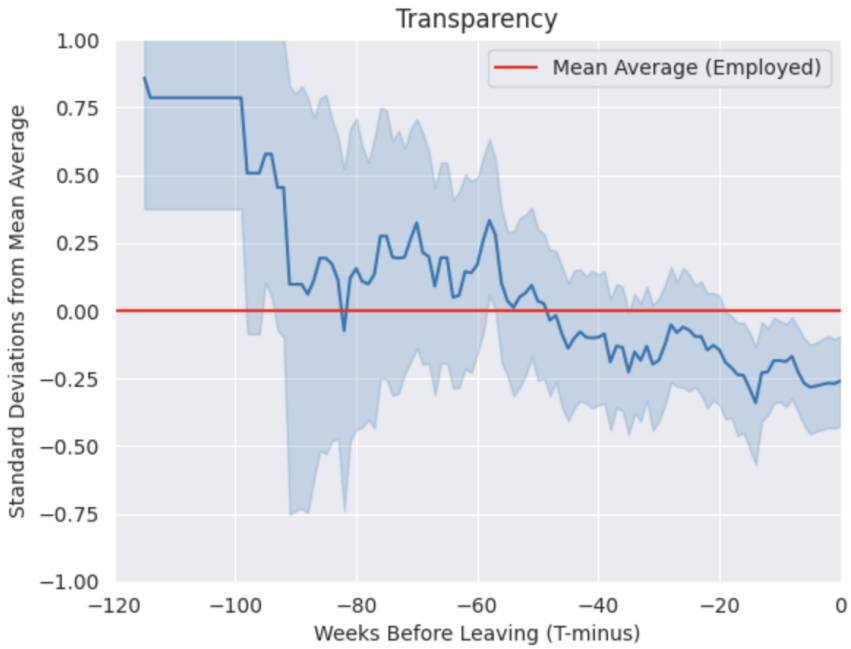


Figure 9: Employee Voice

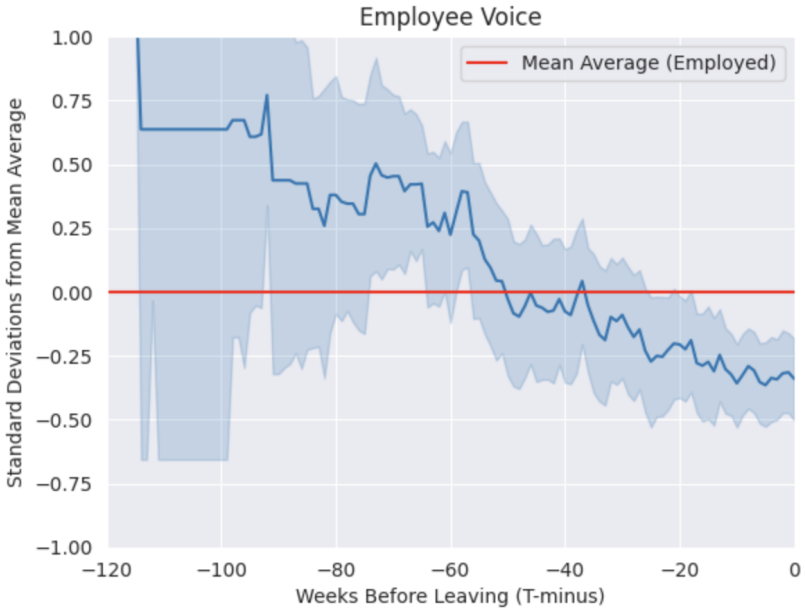


Figure 10: Grievances

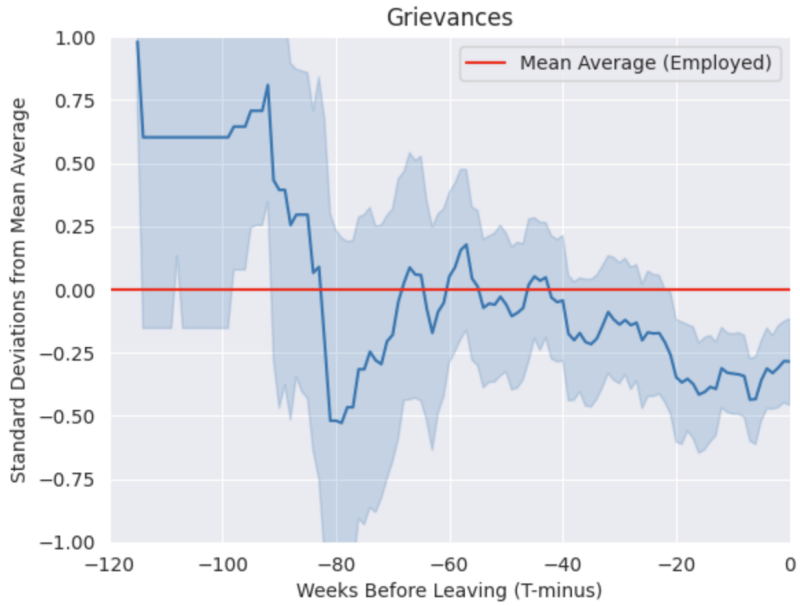


Figure 11: Employee Trust

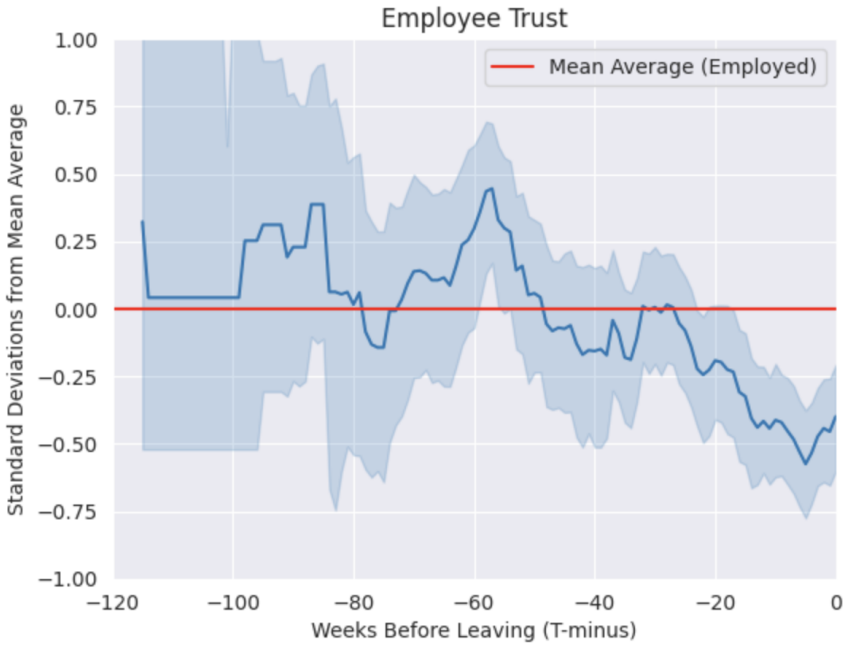


Figure 12: Ability to Take Leave

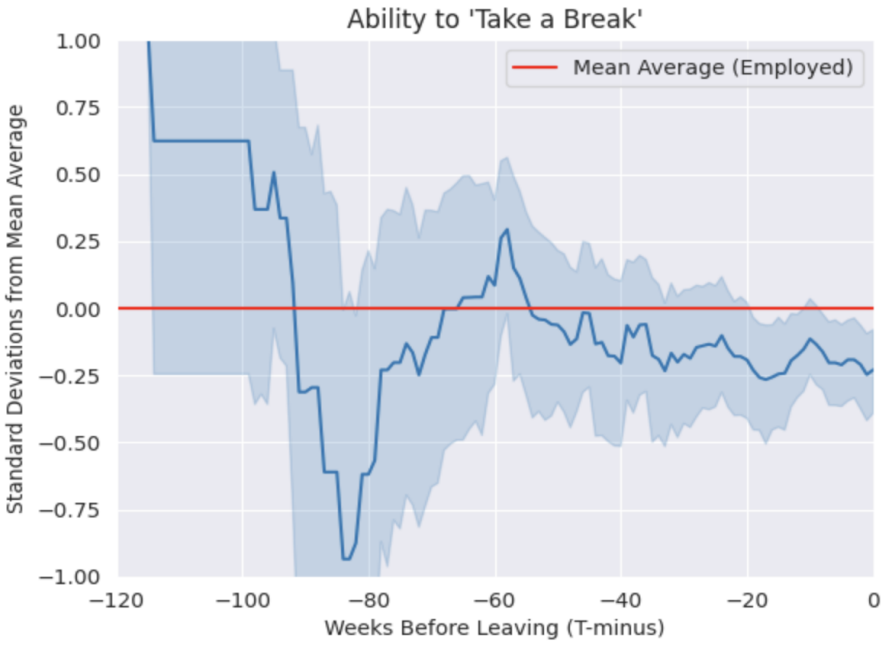


Figure 13: Manager Respect

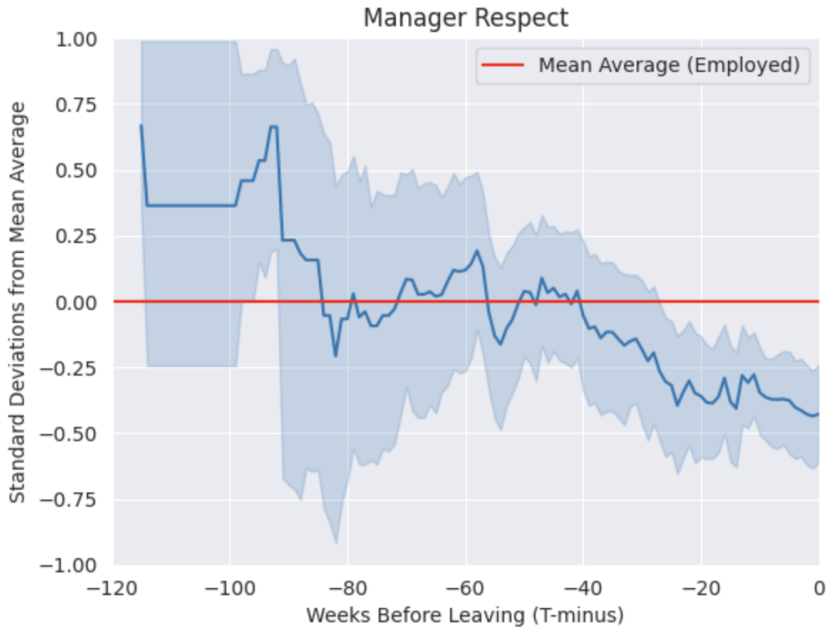


Figure 14: Expectations

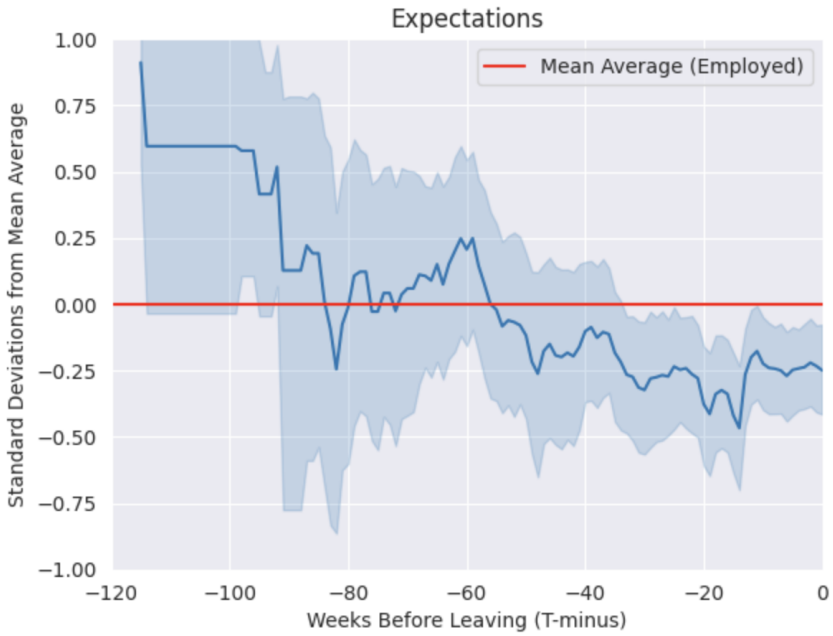


Figure 15: Competence

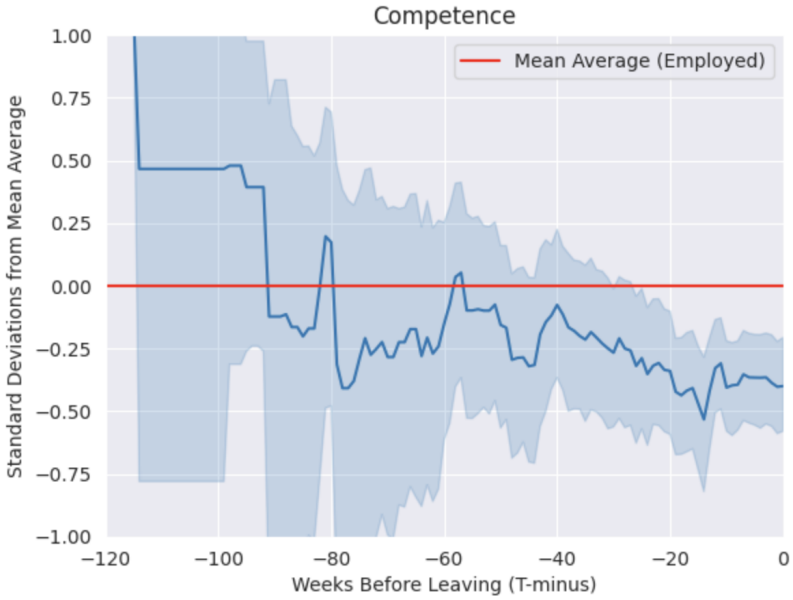


Figure 16: Support

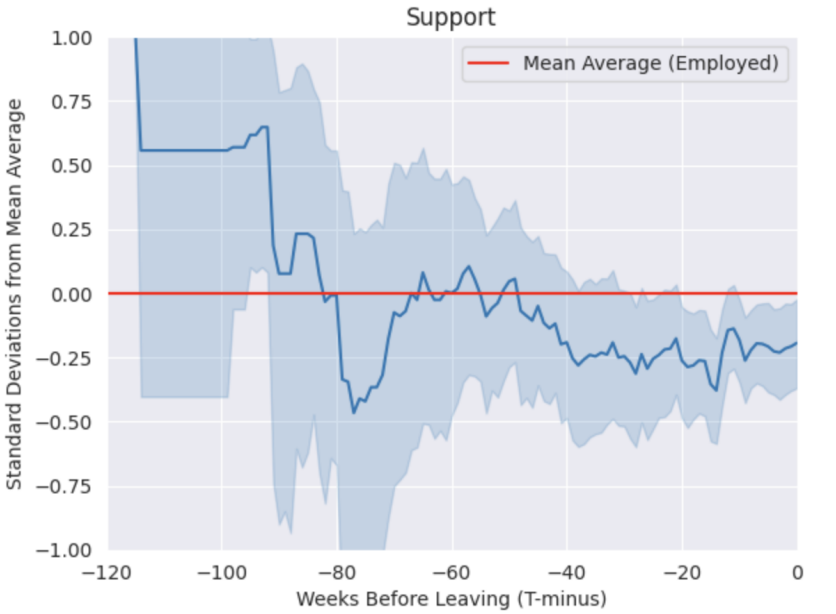


Figure 17: Inter-Team

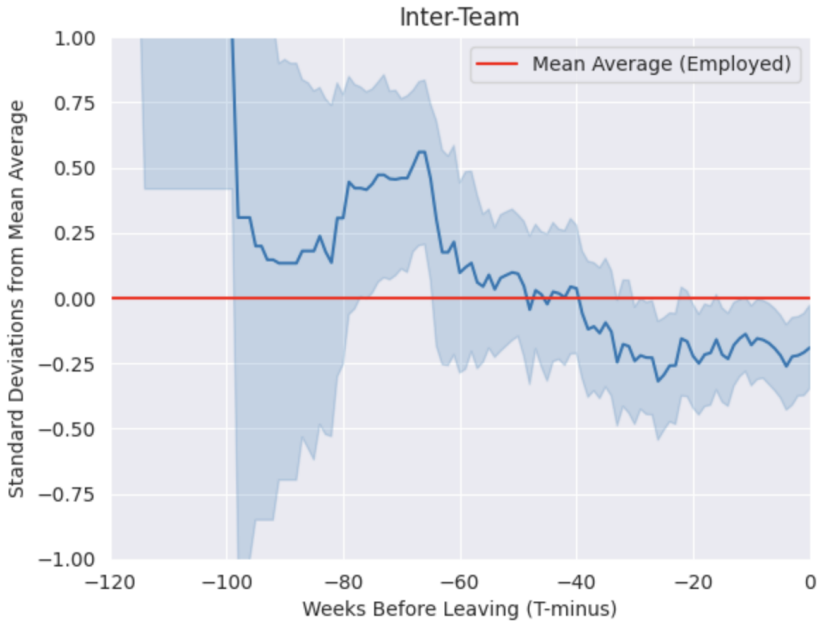


Figure 18: Integrity

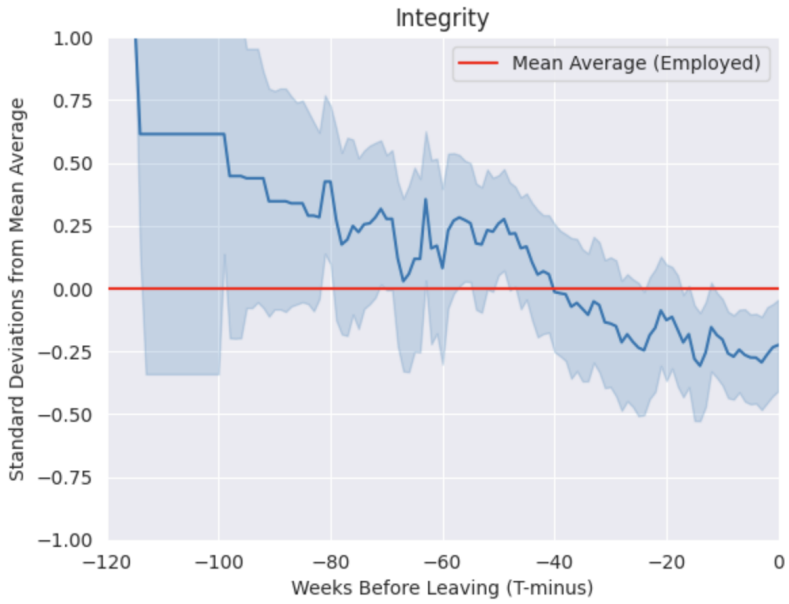


Figure 19: Social

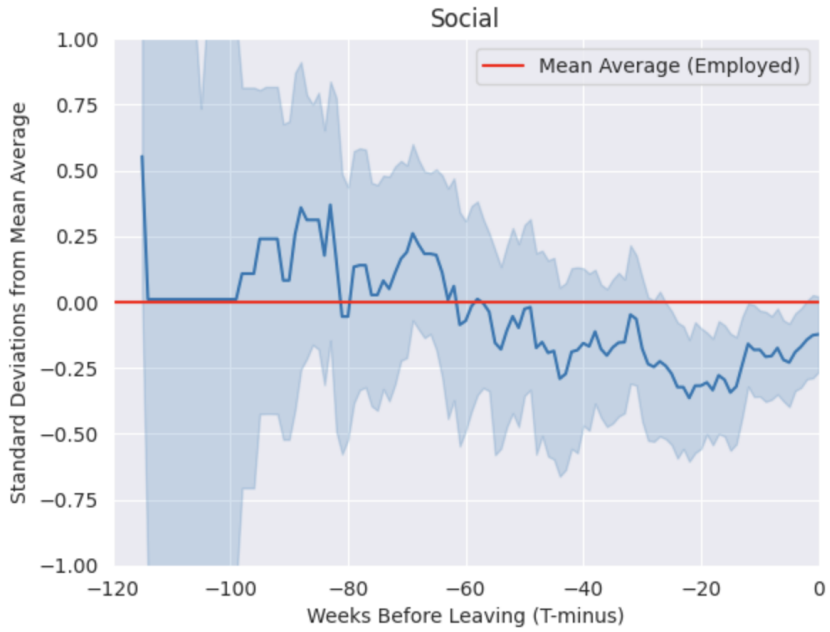


Figure 20: Financial Wellbeing

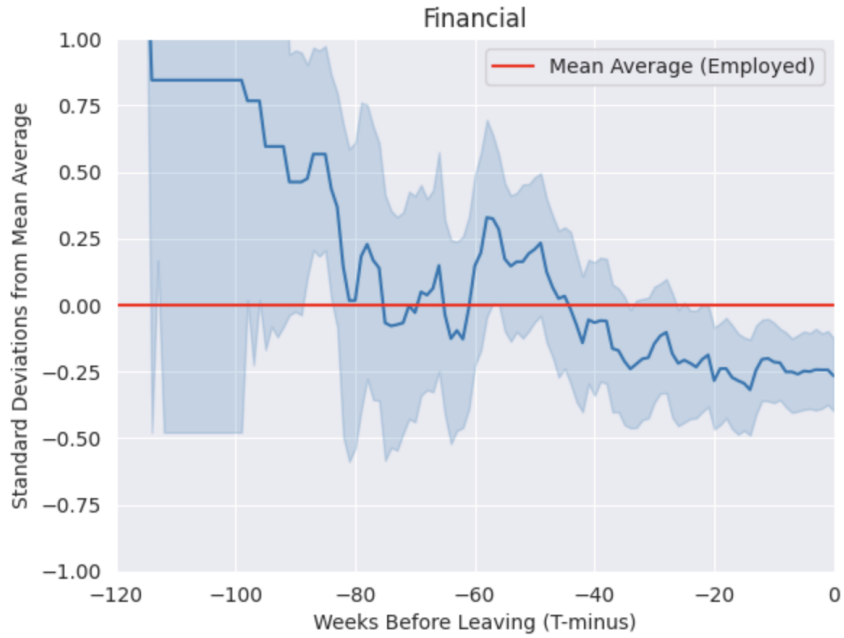
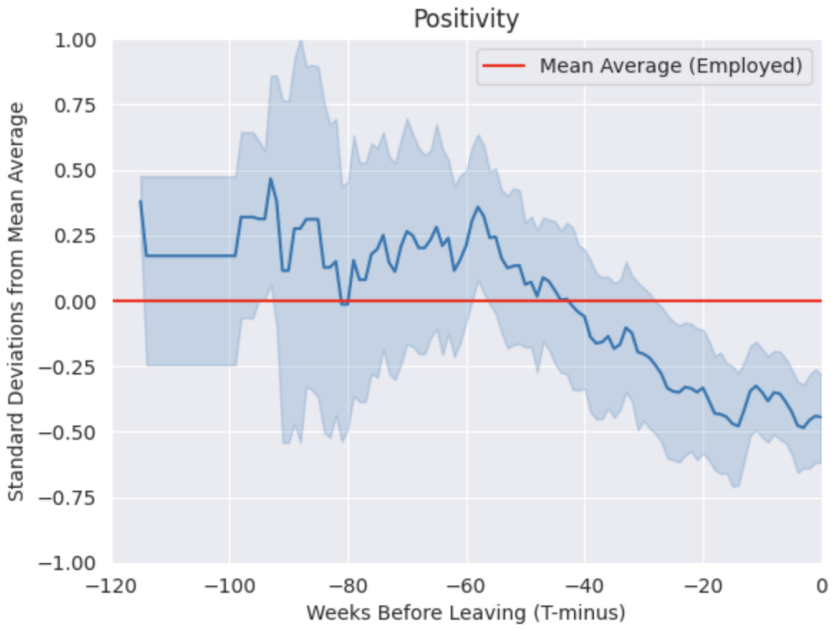


Figure 21: Positivity



What do we know about professionals that leave a job?

There may be gender disparities in professional experiences leading up to leaving a company.

Men seemed to score higher on Core24 metrics initially before converging with those from other genders below the mean employed average for the given metric. Examples of this can be seen in the figures below:

Figure 22: Expectations by Gender

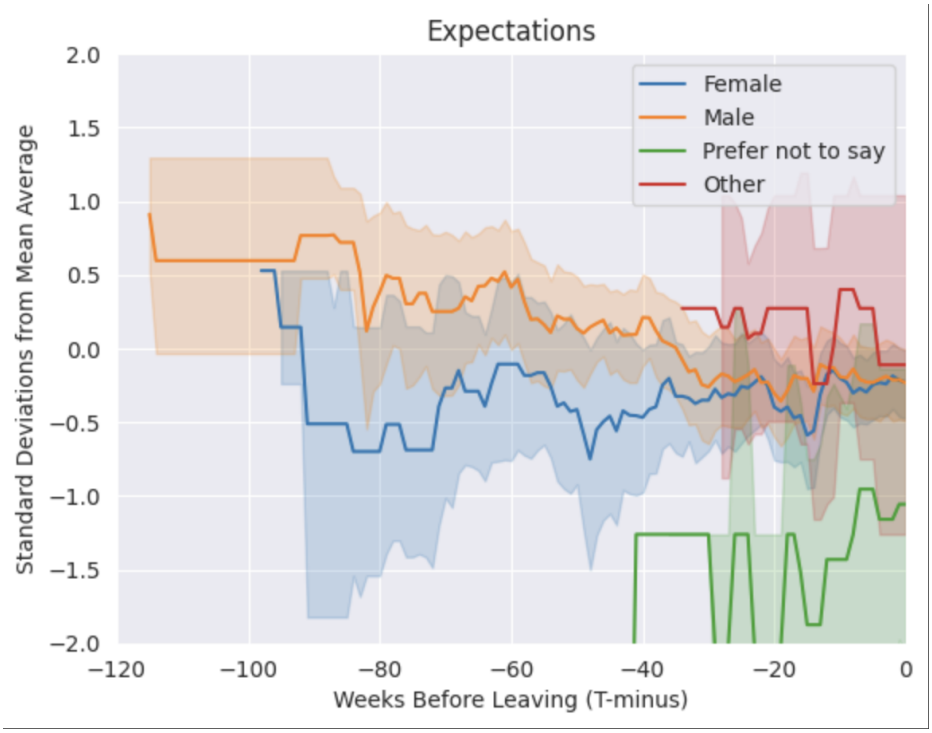


Figure 23: Values by Gender

Figure 24: Future Progression by Gender

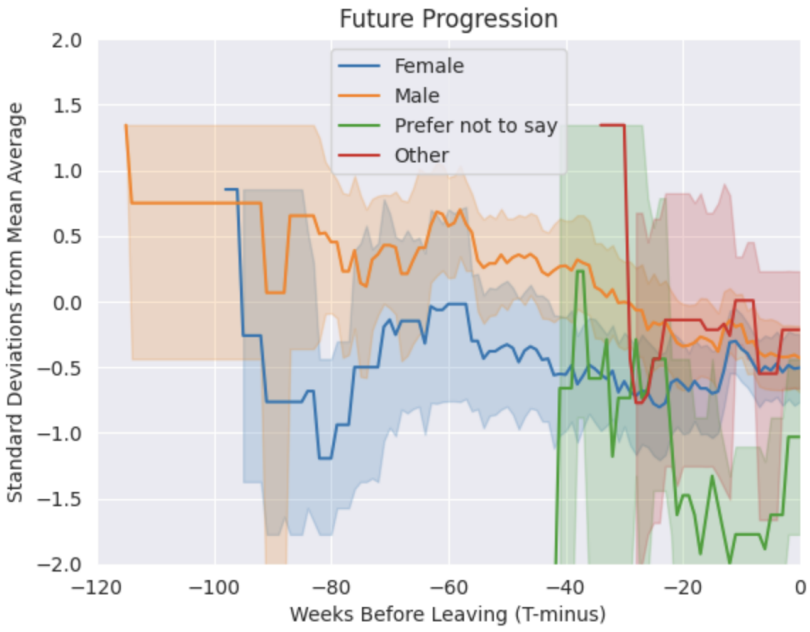


Figure 25: Pride by Gender

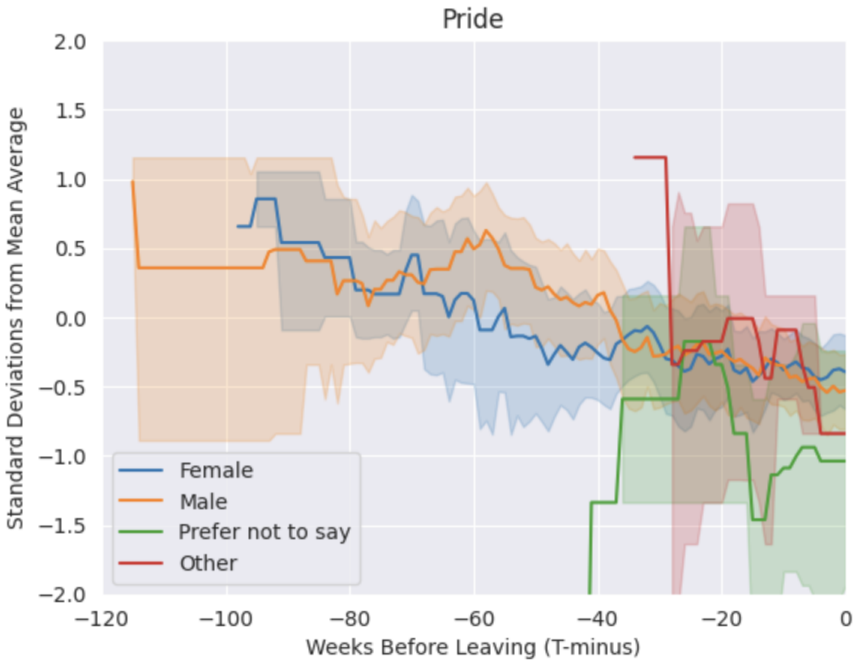


Figure 26: Transparency by Gender

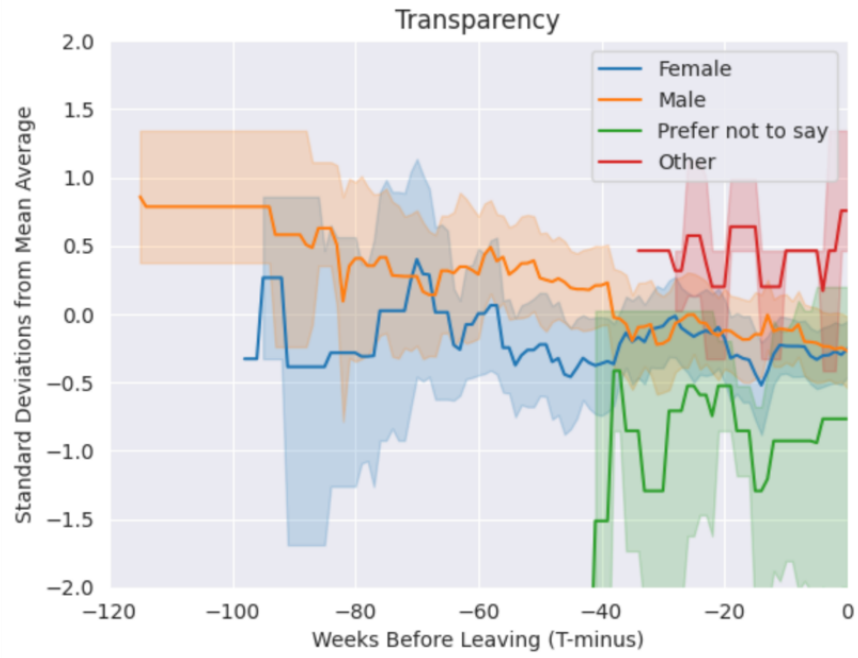


Figure 27: Pride by Gender

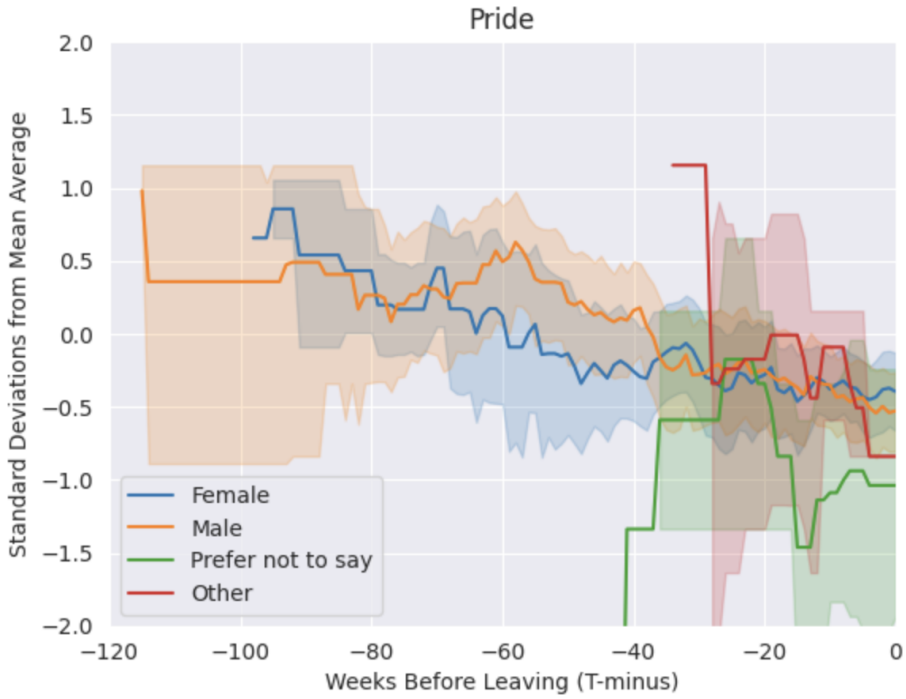


Figure 28: Employee Voice by Gender

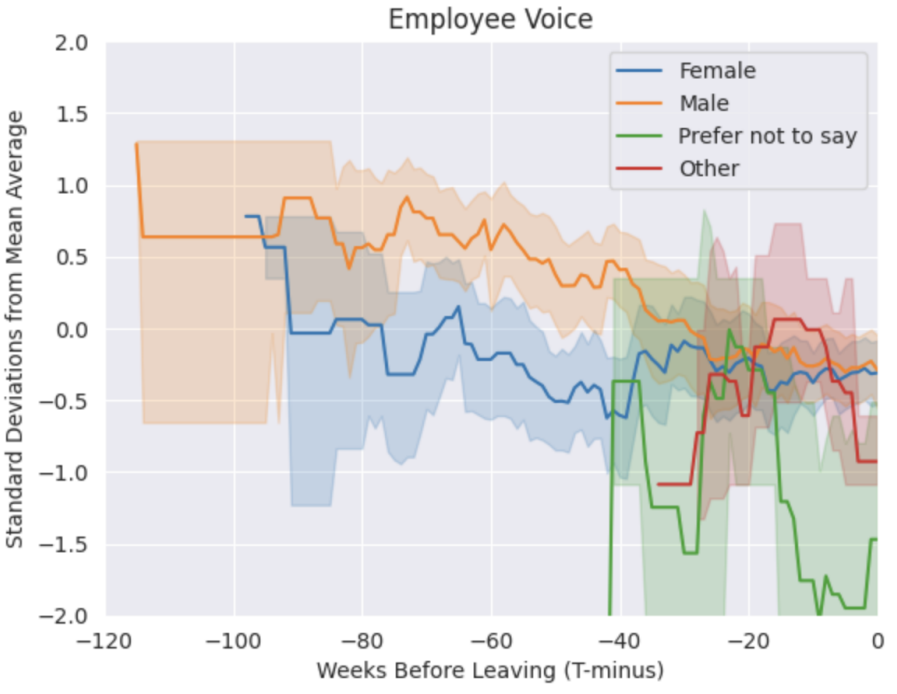


Figure 29: Grievances by Gender

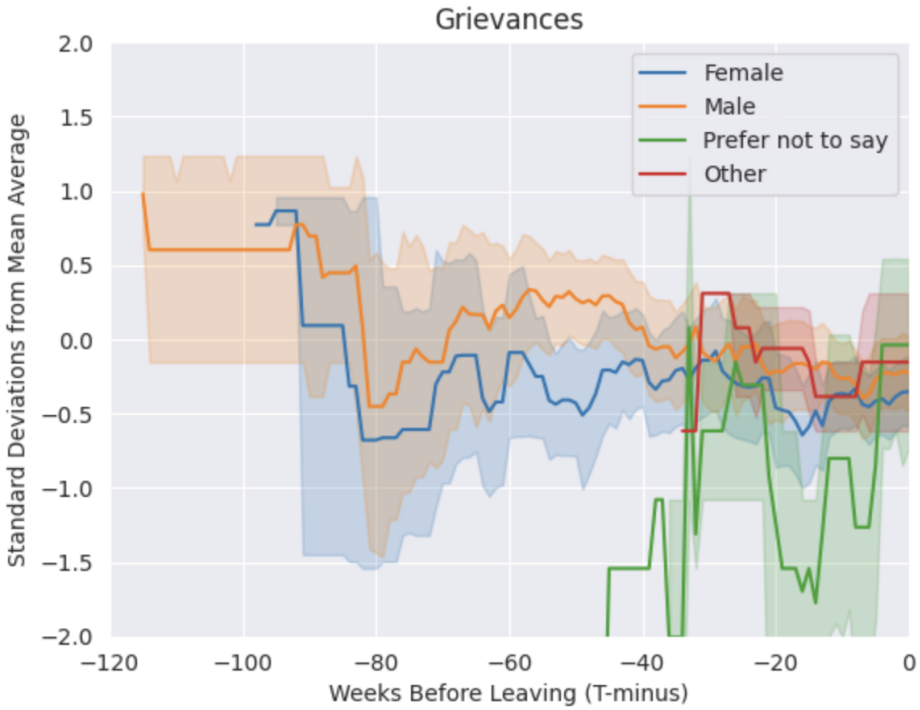


Figure 30: Employee Trust by Gender

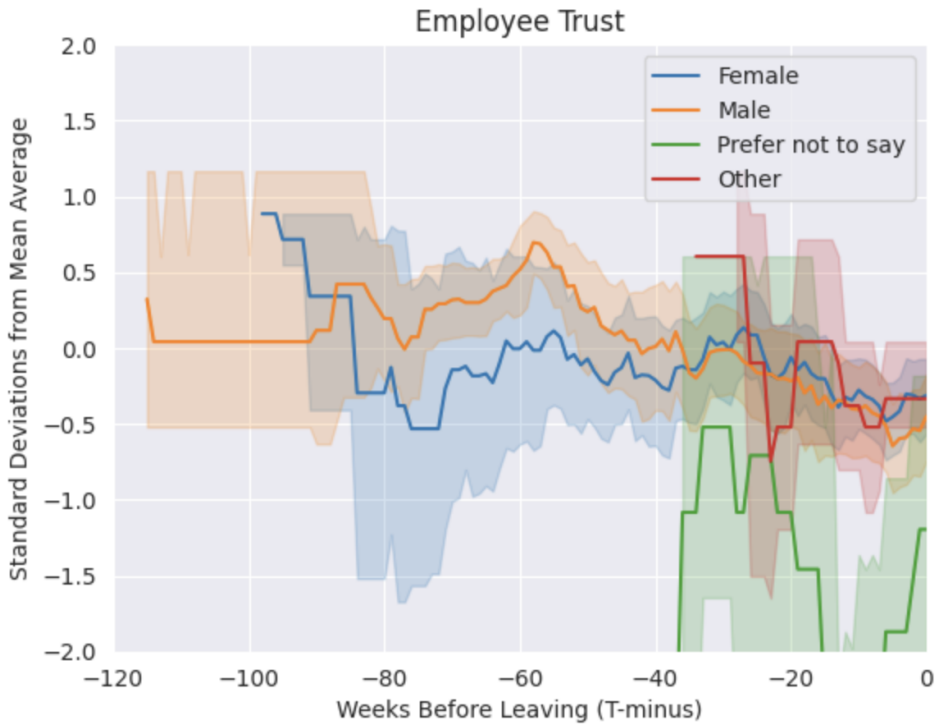


Figure 31: Support by Gender

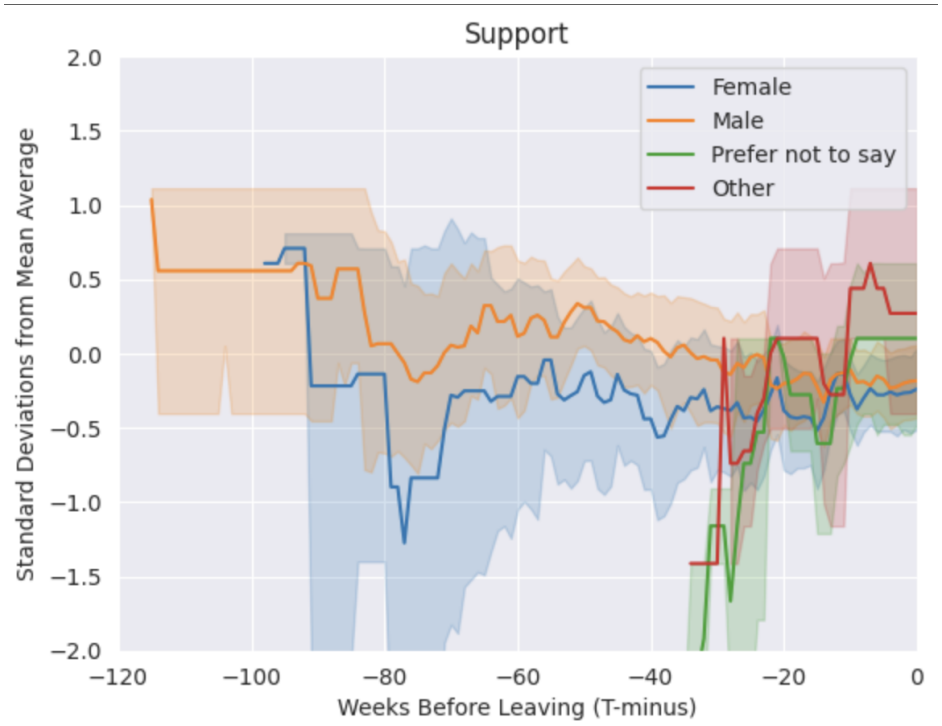


Figure 32: Manager Respect by Gender

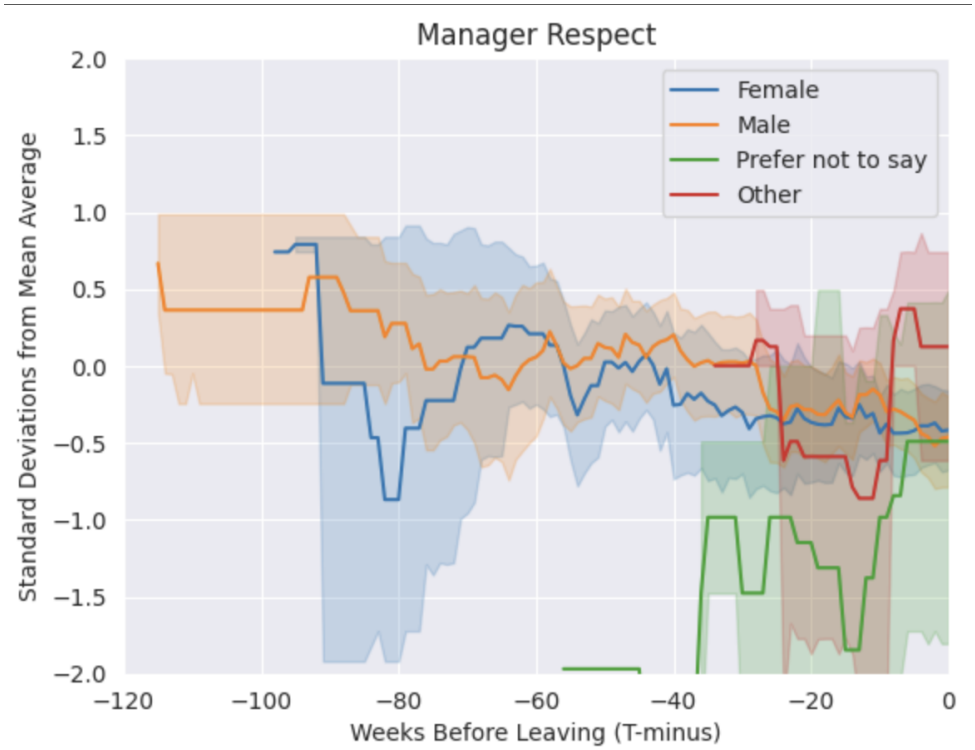


Figure 33: Competence by Gender

Figure 34: Integrity by Gender

Figure 35: Team Respect by Gender

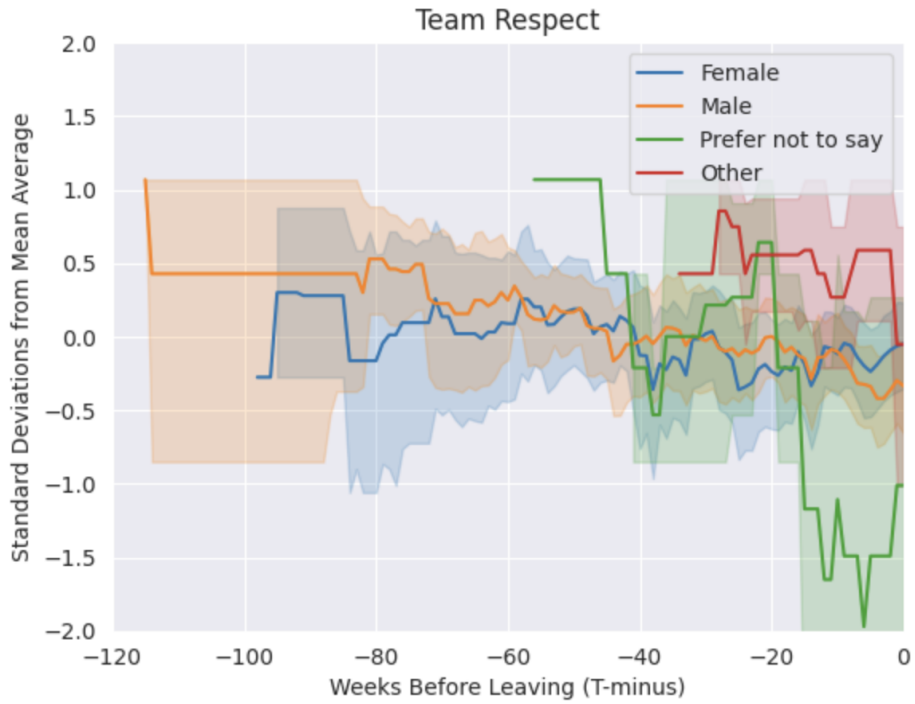


Figure 36: Inter-team Relations by Gender

Figure 37: Social Metric by Gender

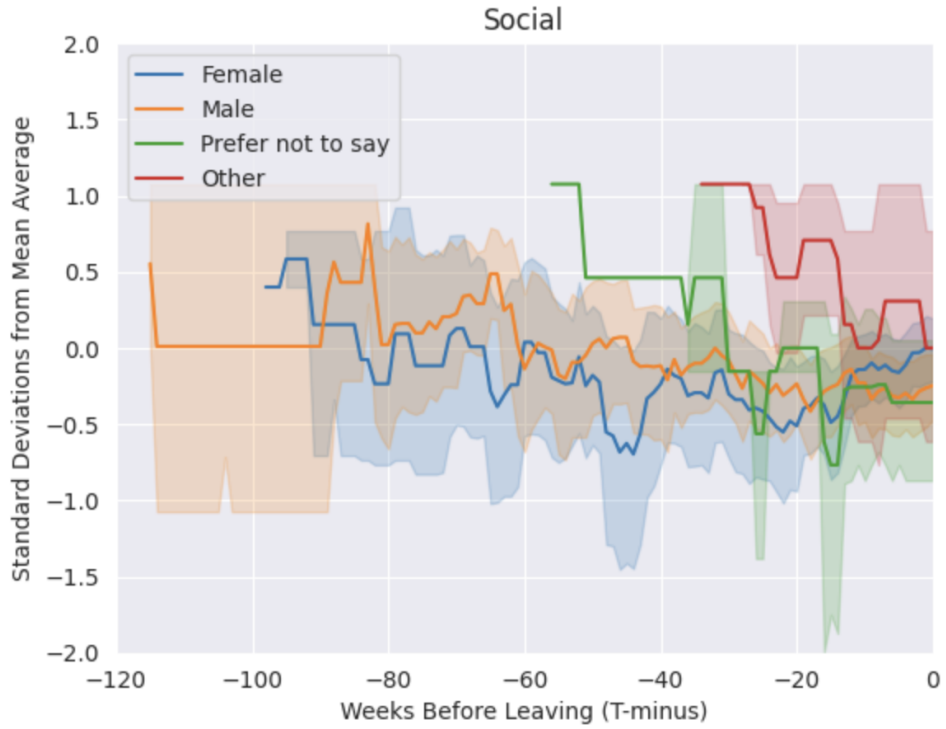


Figure 38: Ability to take a break/take leave by Gender

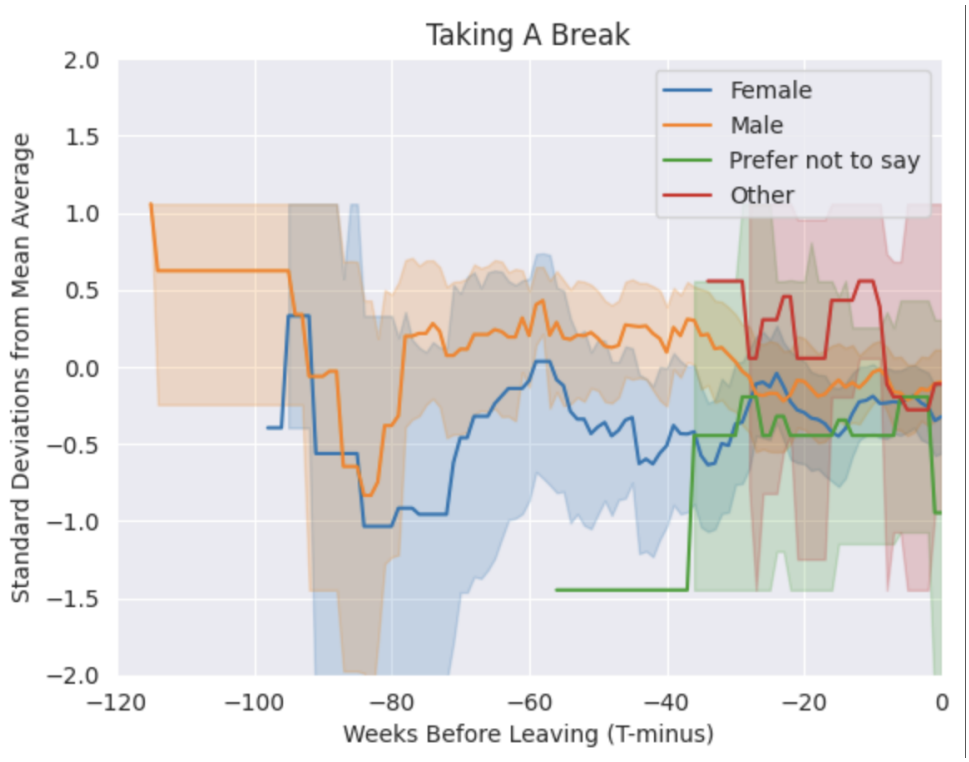


Figure 39: Work/Life Balance by Gender

Figure 40: Recognition by Gender

Figure 41: Company Care by Gender

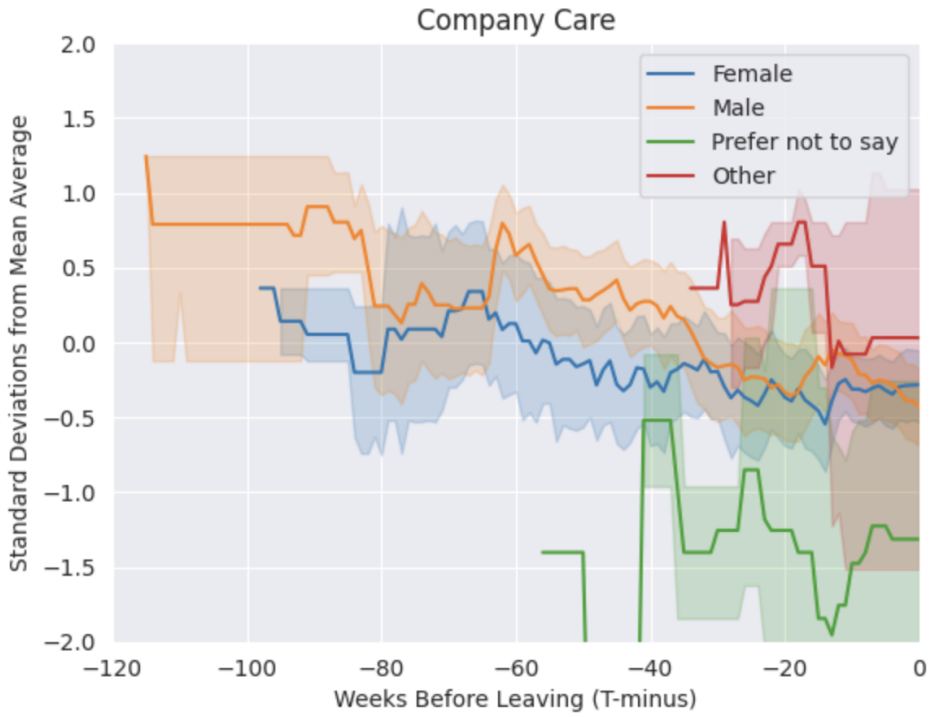


Figure 42: Composure by Gender

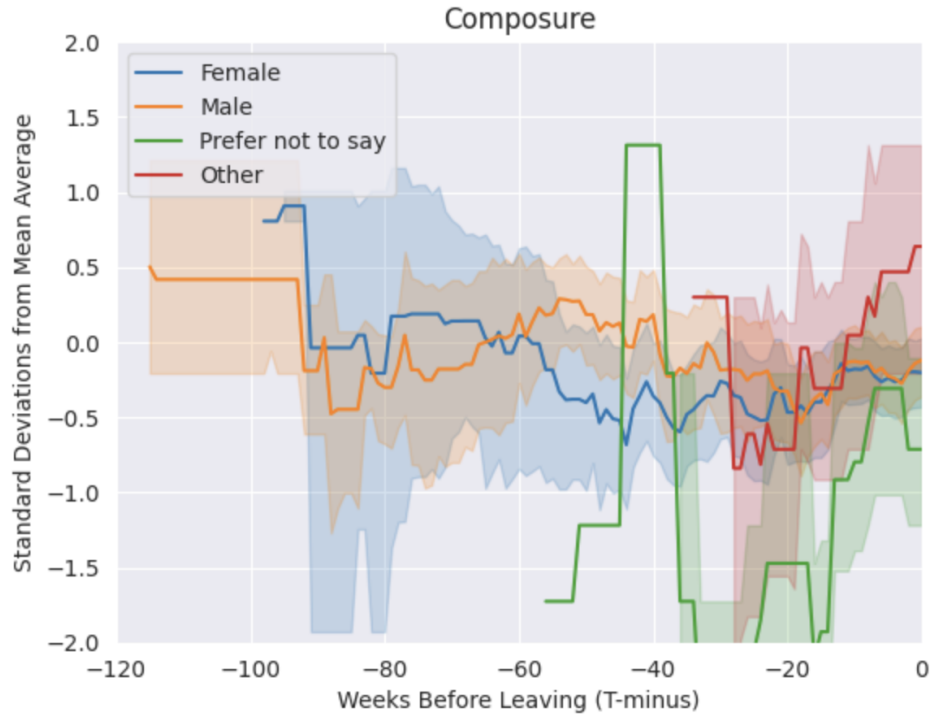
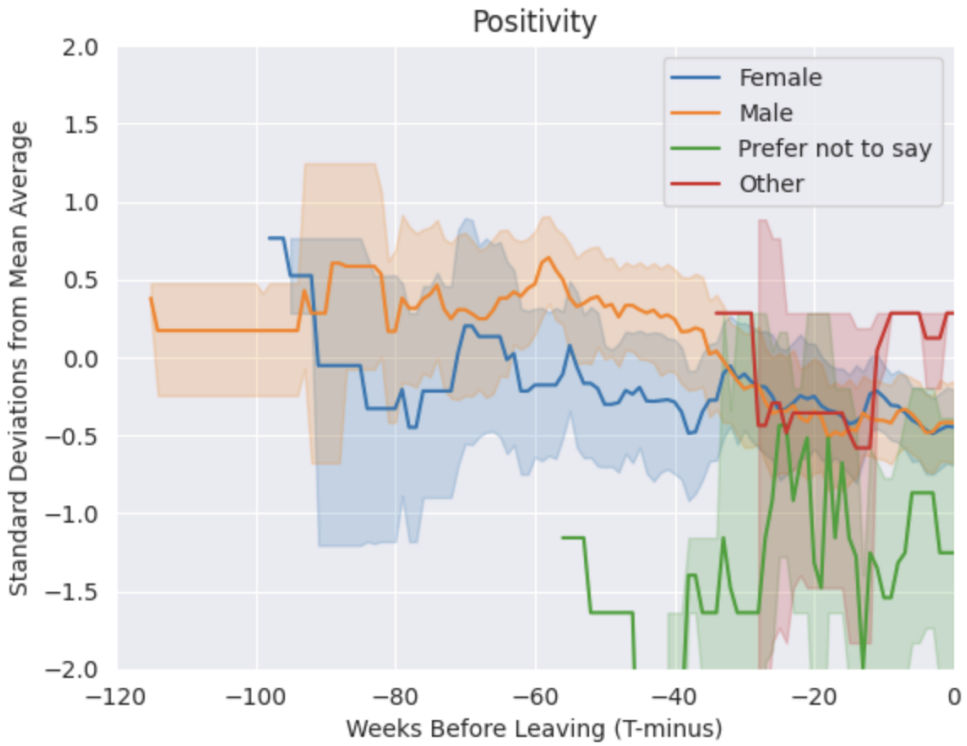


Figure 43: Positivity by Gender



This could be indicative of experiences of men in professional environments being significantly better (or at least, feeling significantly better) initially whilst those from other genders tend to have negative experiences in these environments from the outset of their employment. Also, those who prefer not to say their gender or identify as 'Other' tend to skew the data to the downside. People from these groups may need more support in the workplace.

Note that these disparities are most pronounced in Employee Voice, Employee Trust, Support and 'ability to take leave' metrics.

Those from Generation Z may be less likely to exhibit signs of unhappiness prior to leaving a job than their Millennial and/or Generation X counterparts.

Employees who left a job as part of Generation Z scored consistently higher on many Core24 metrics, commonly scoring above the mean average of those still employed in the database. Examples of this can be seen graphically figures 44-67 below.

Figure 44: Expectations by Generation

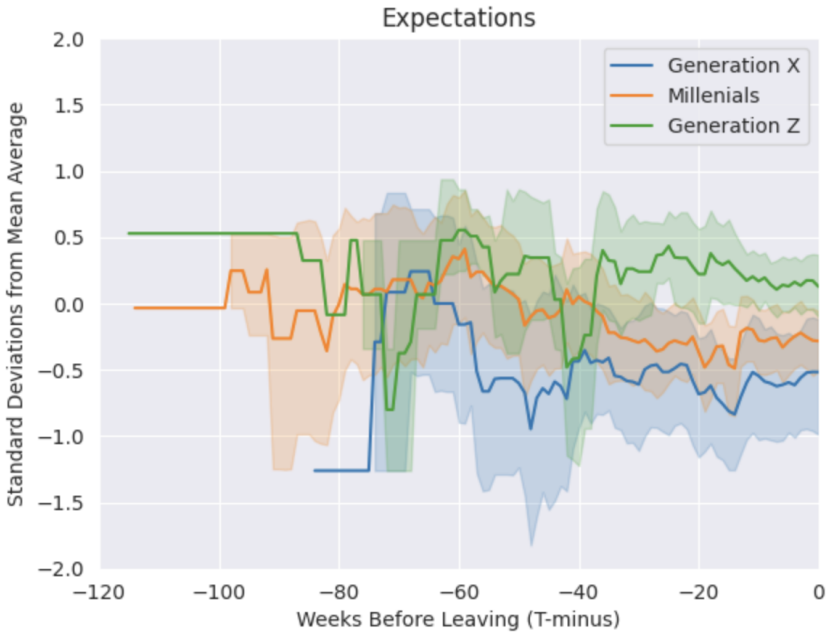


Figure 45: Self Worth by Generation

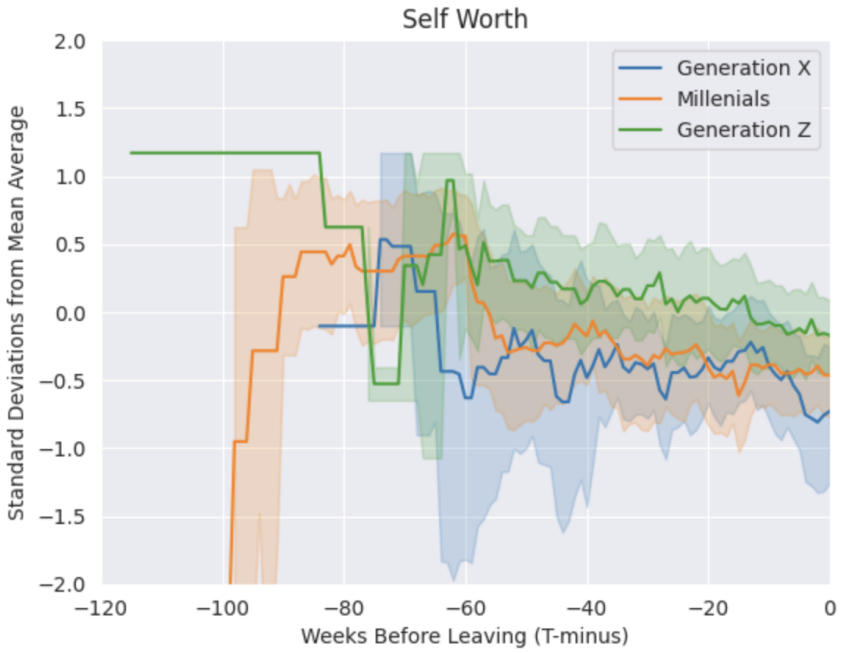


Figure 46: Financial Wellbeing by Generation

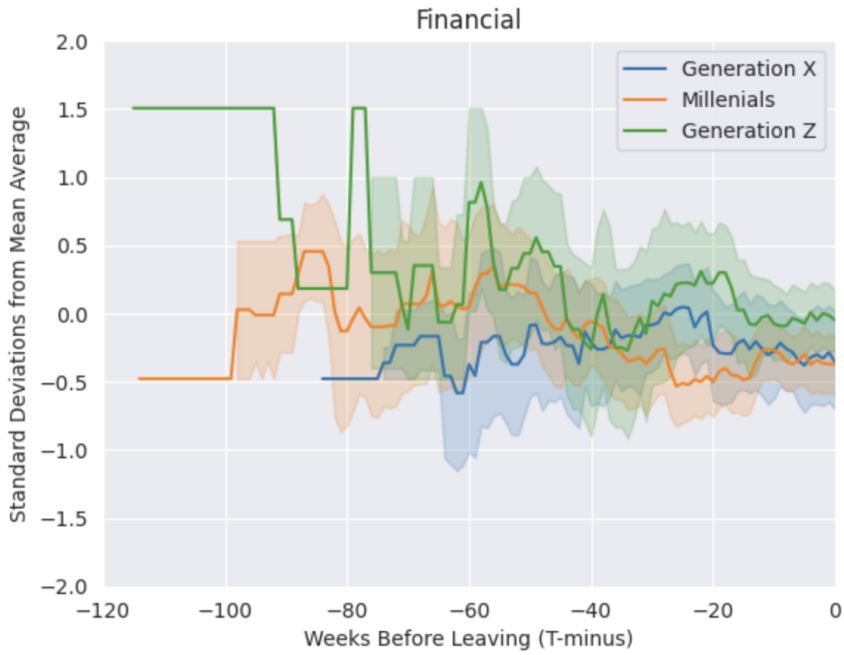


Figure 47: Values by Generation

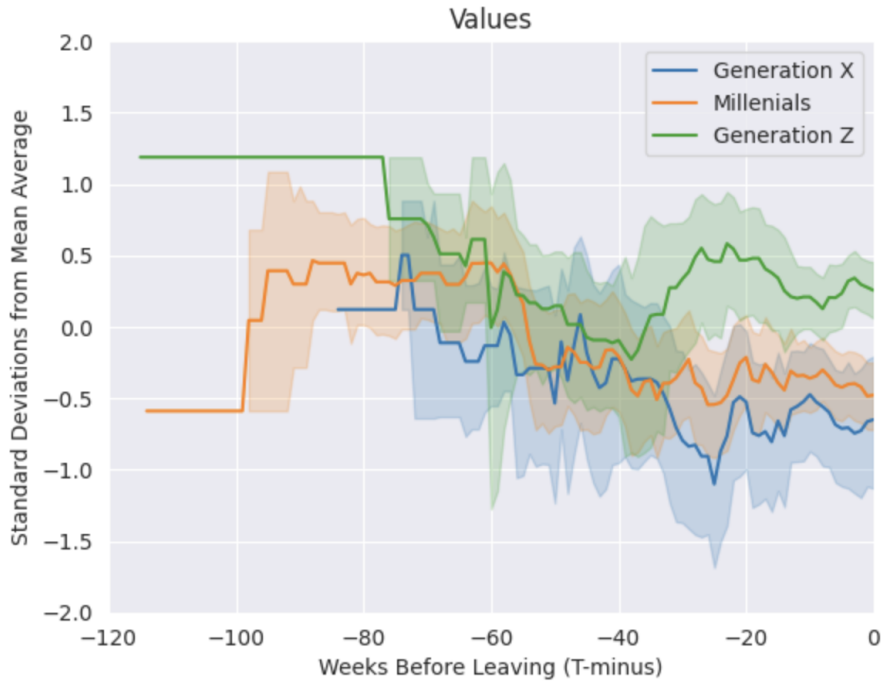


Figure 48: Future Progression by Generation

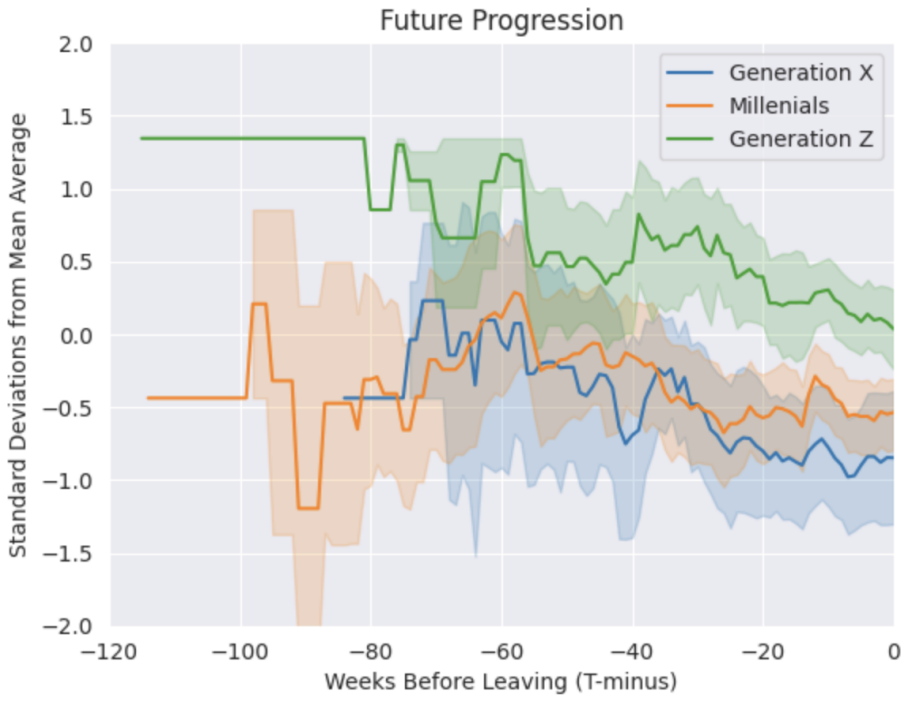


Figure 49: Pride by Generation

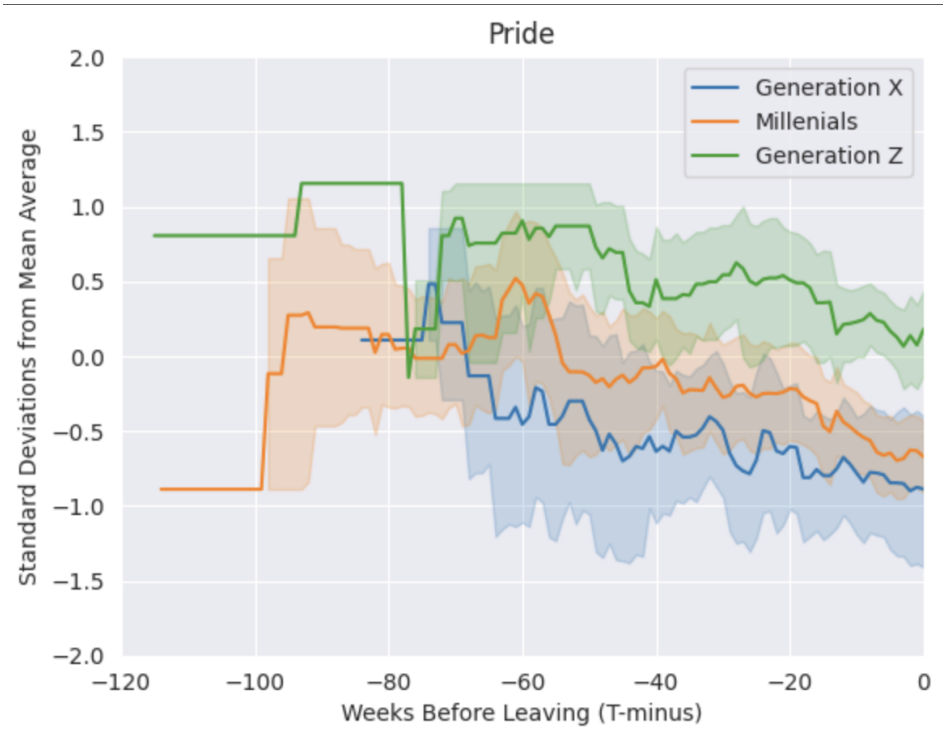


Figure 50: Transparency by Generation

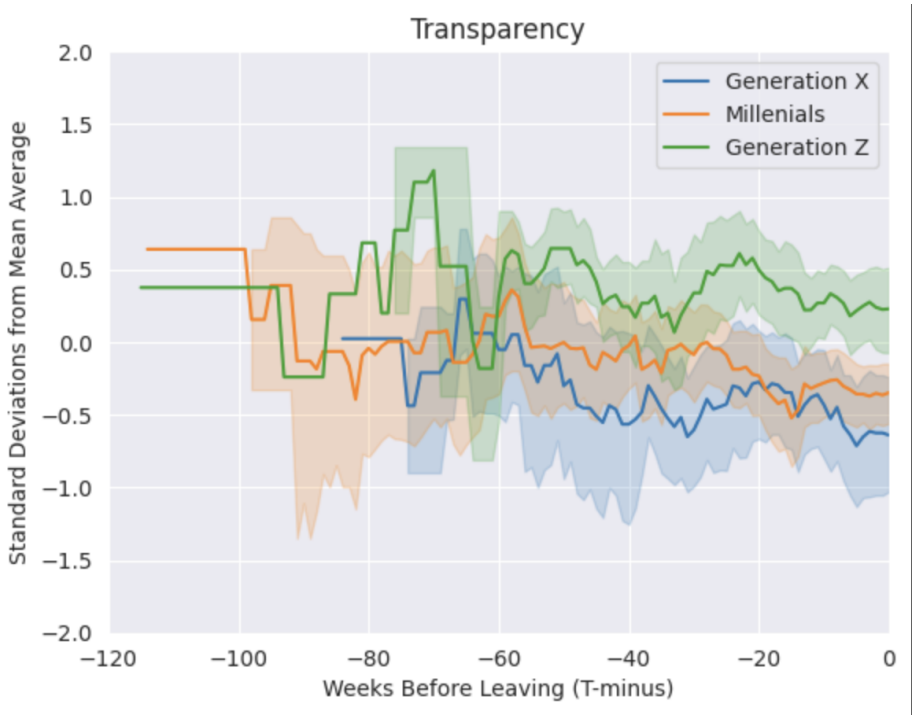


Figure 51: Employee Voice by Generation

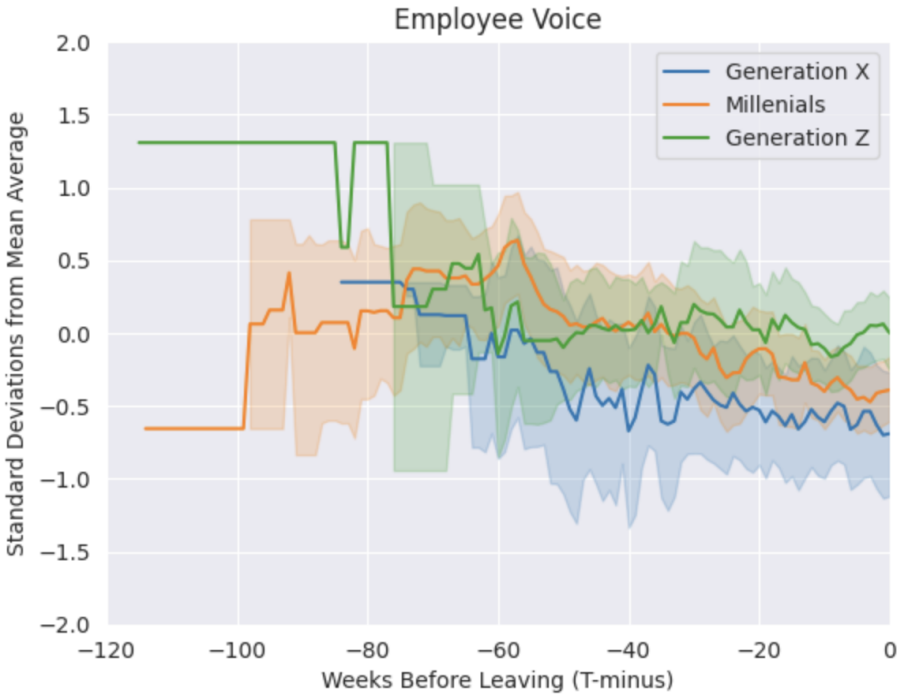


Figure 52: Grievances by Generation

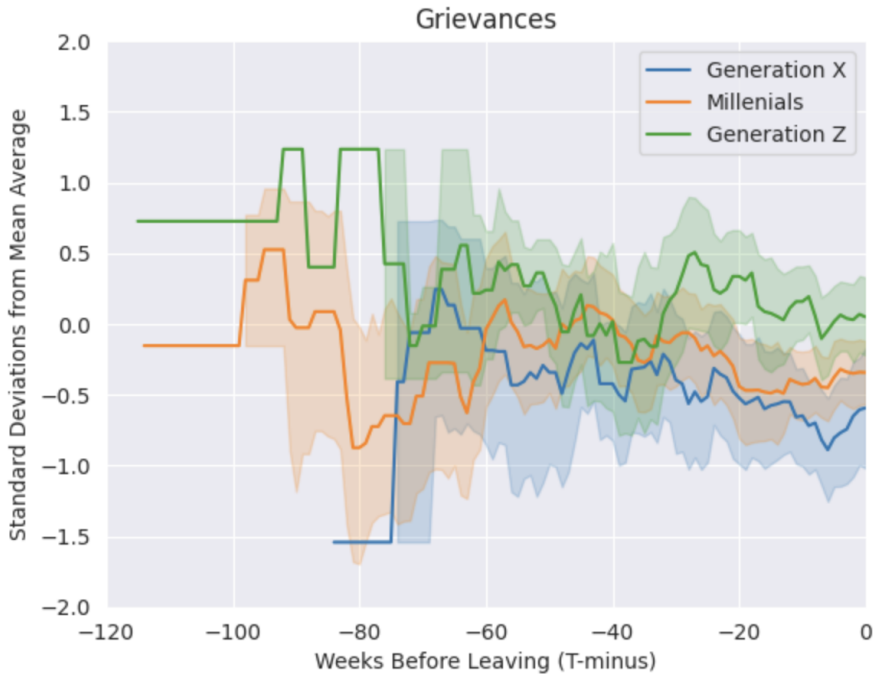


Figure 53: Employee Trust by Generation

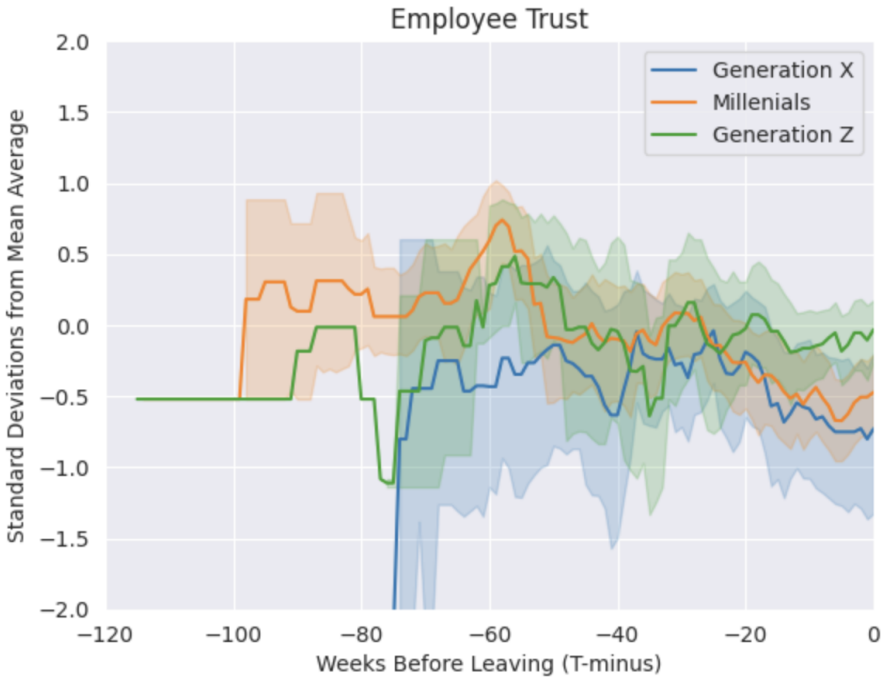


Figure 54: Support by Generation

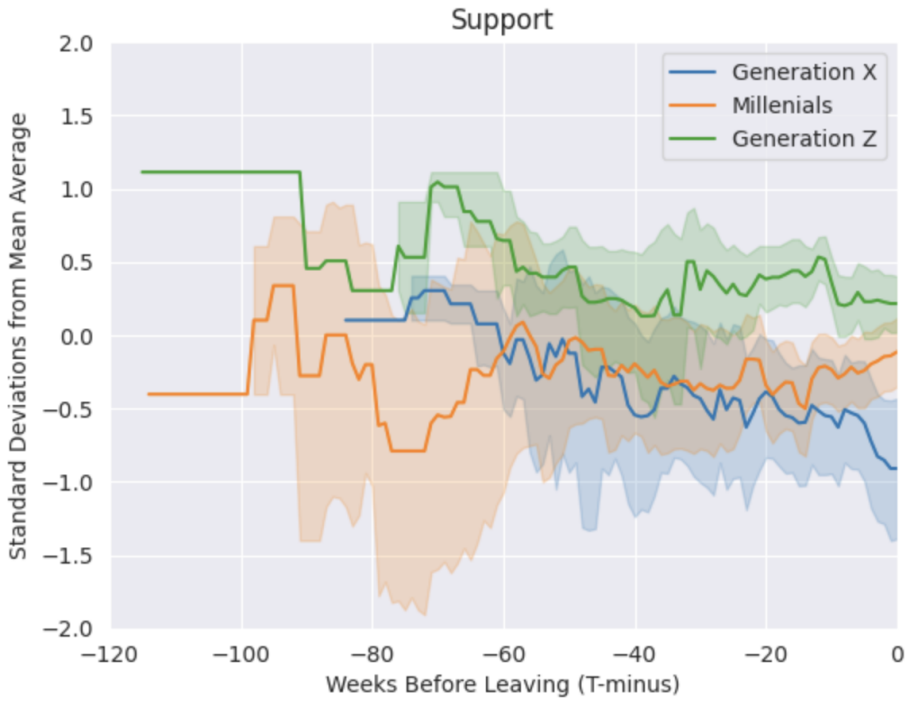


Figure 55: Manager Respect by Generation

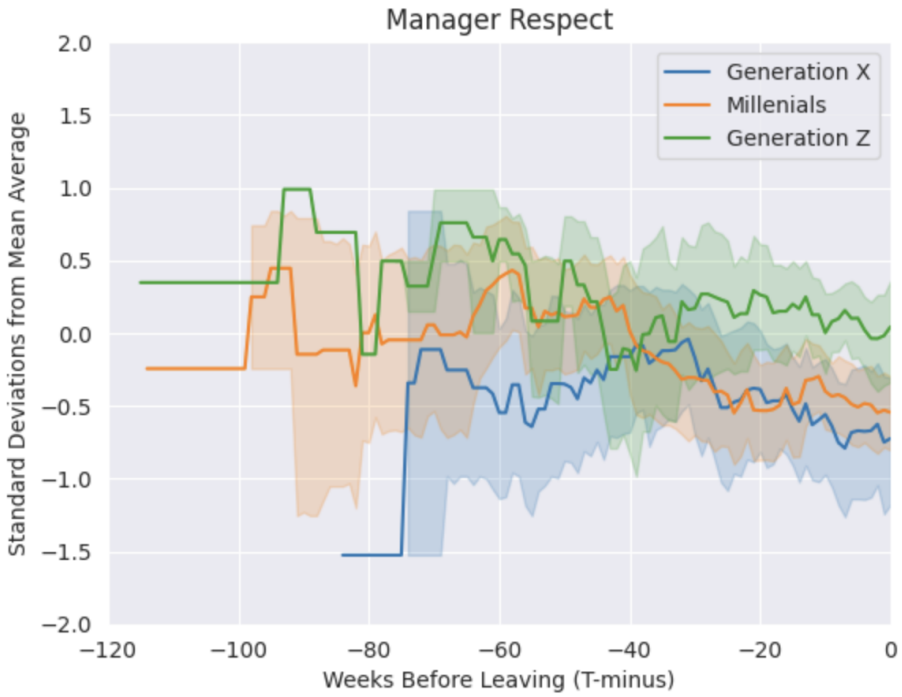


Figure 56: Competence by Generation

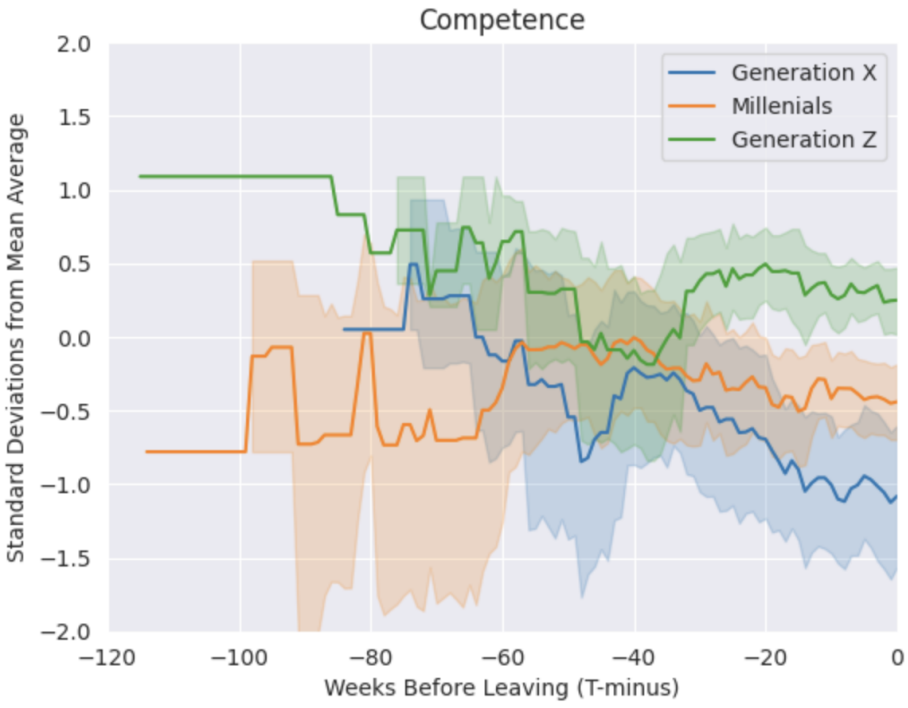


Figure 57: Integrity by Generation

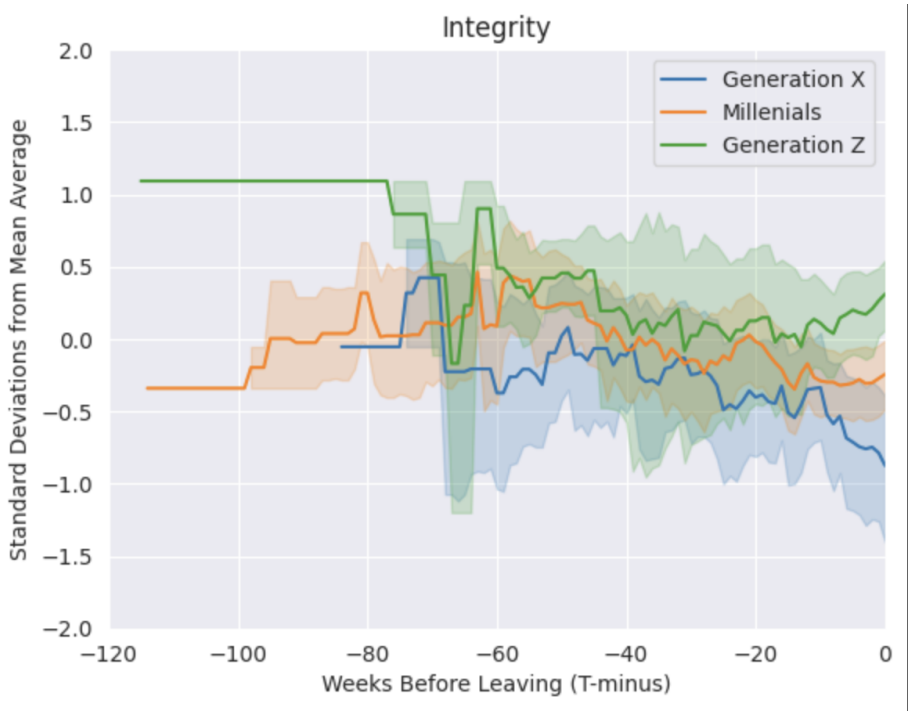


Figure 58: Team Respect by Generation

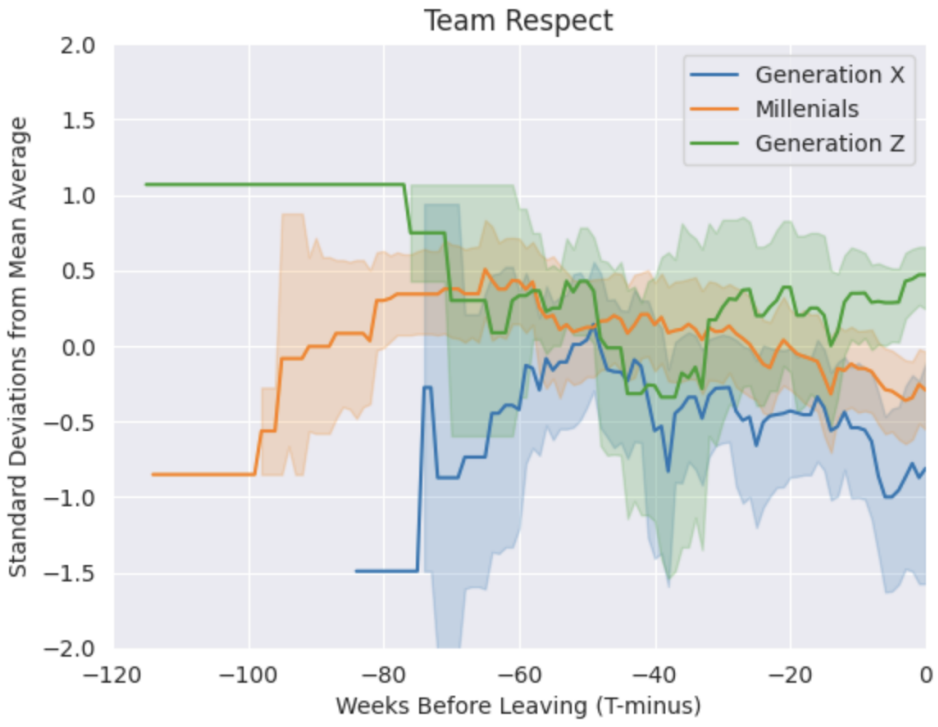


Figure 59: Inter-Team Relations by Generation

Figure 60: Social Wellbeing by Generation

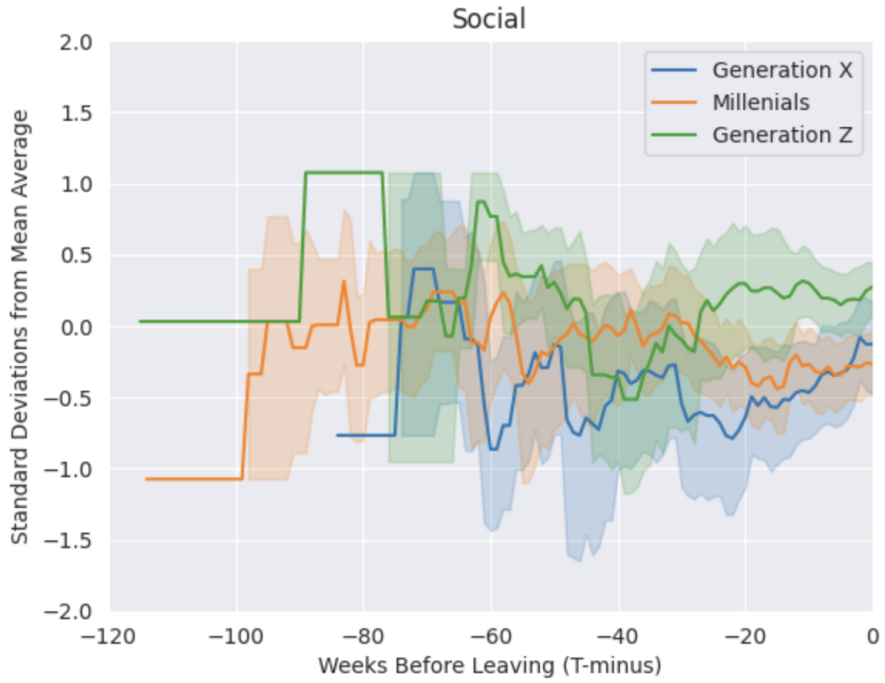


Figure 61: Health by Generation

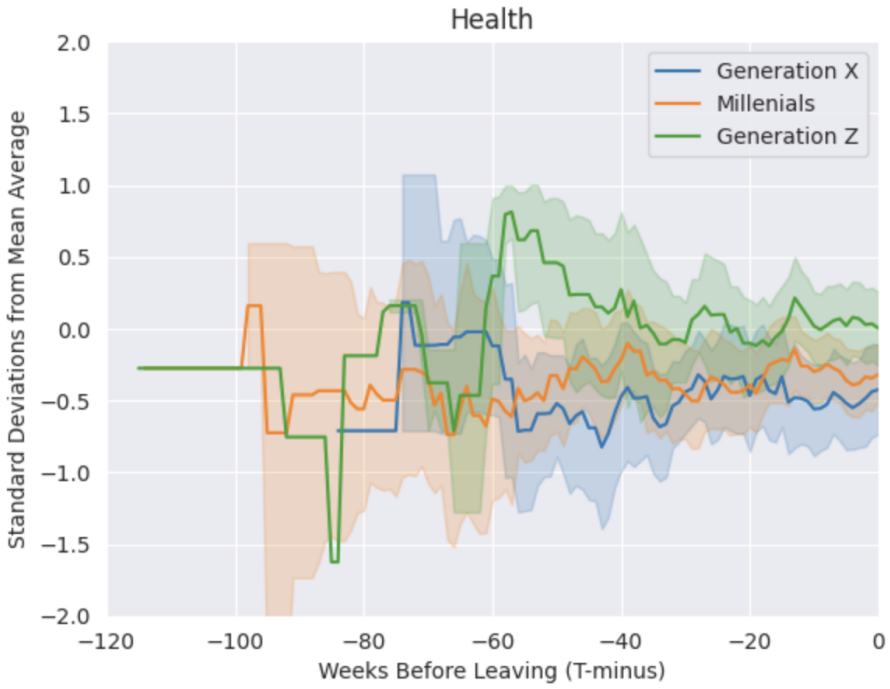


Figure 62: Ability to take a break/take leave by generation

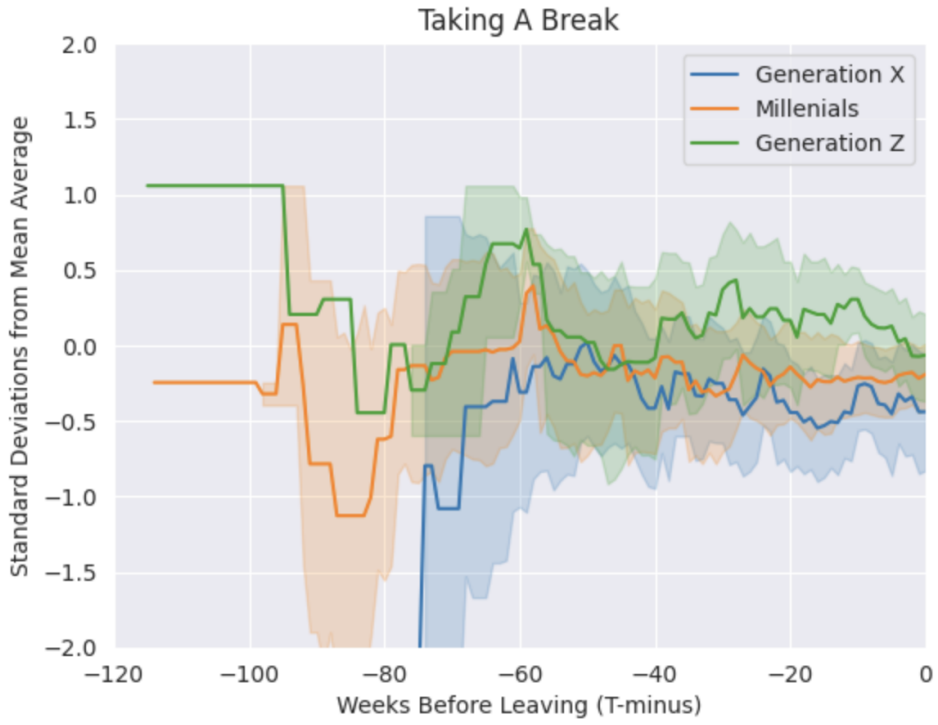


Figure 63: Balance by Generation

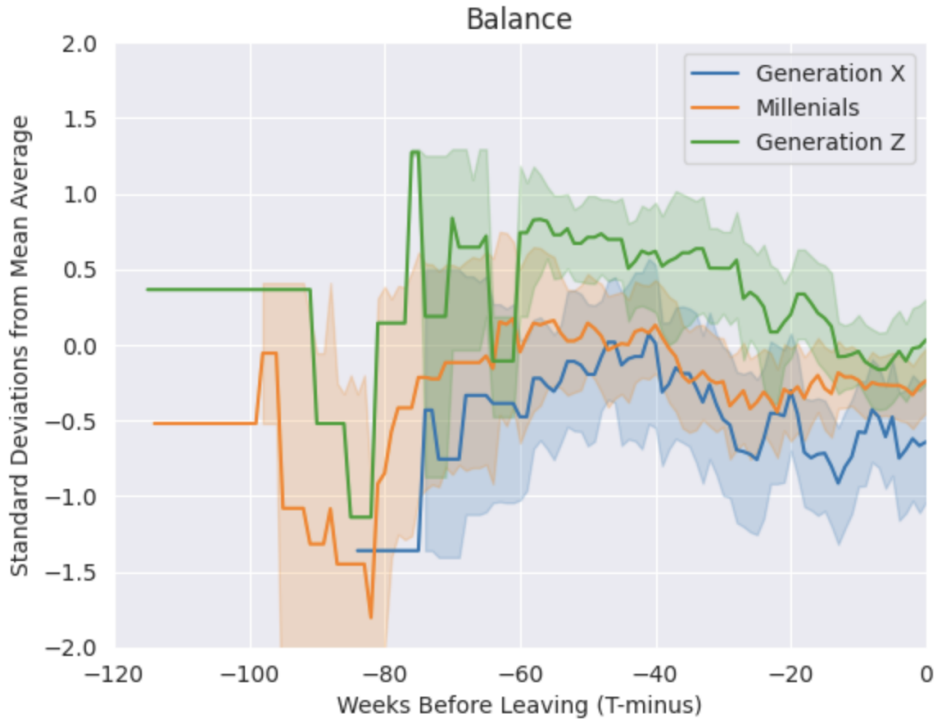


Figure 64: Recognition by Generation

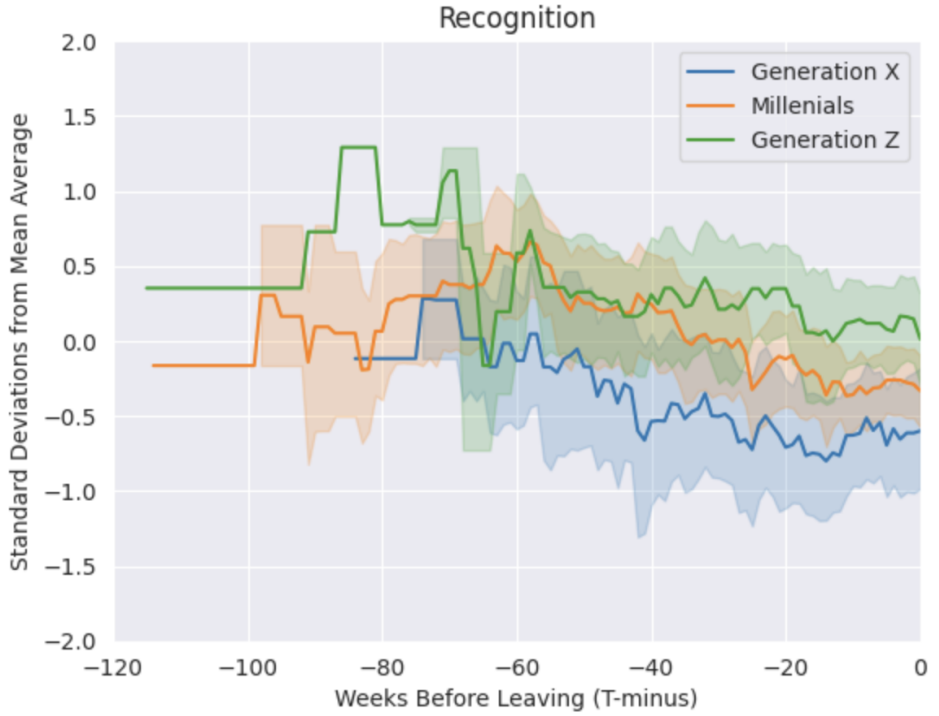


Figure 65: Company Care by Generation

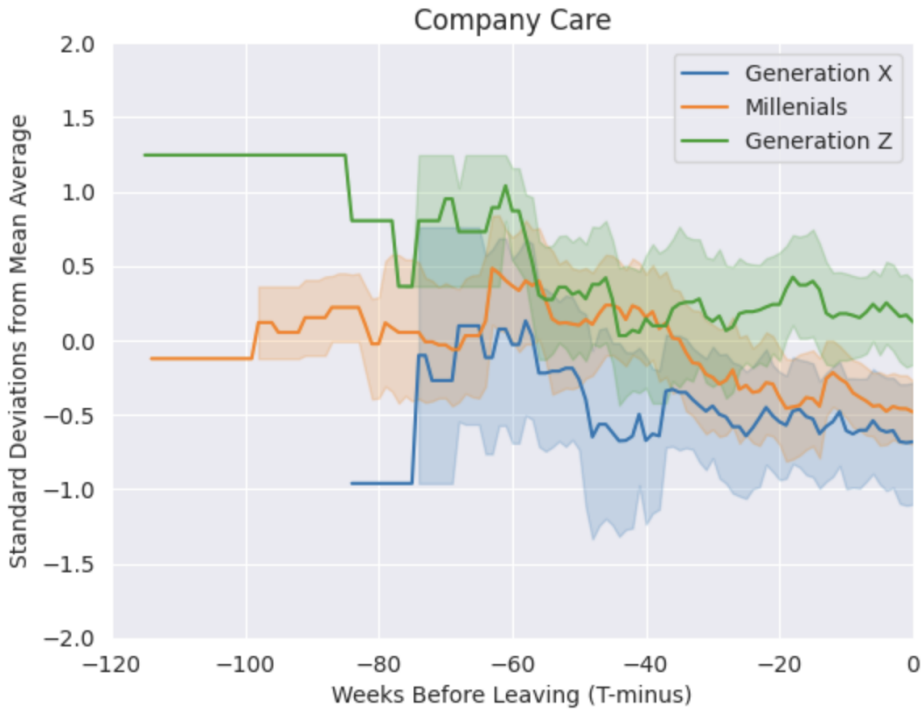


Figure 66: Composure by Generation

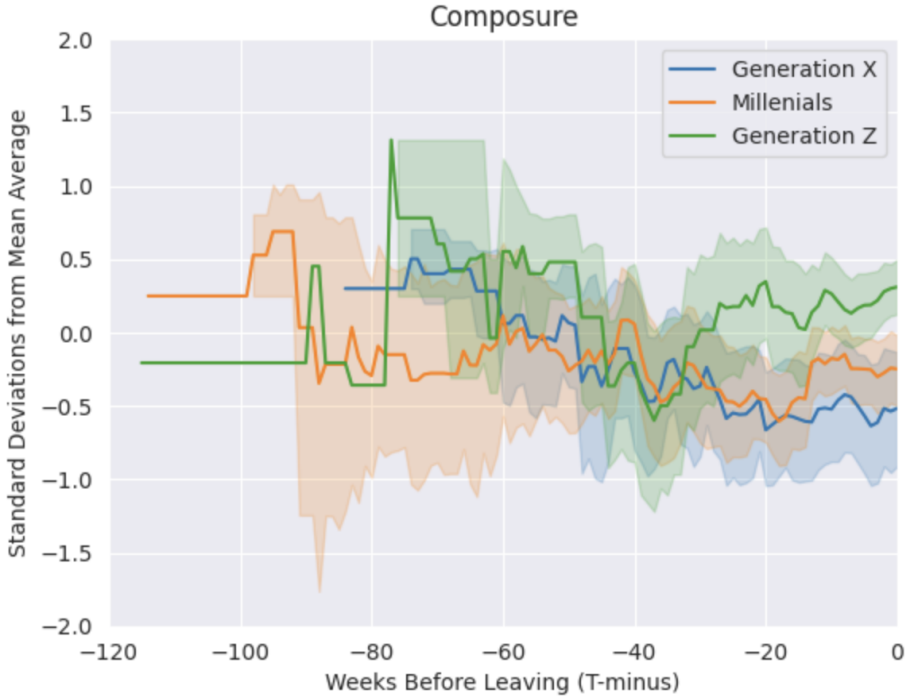
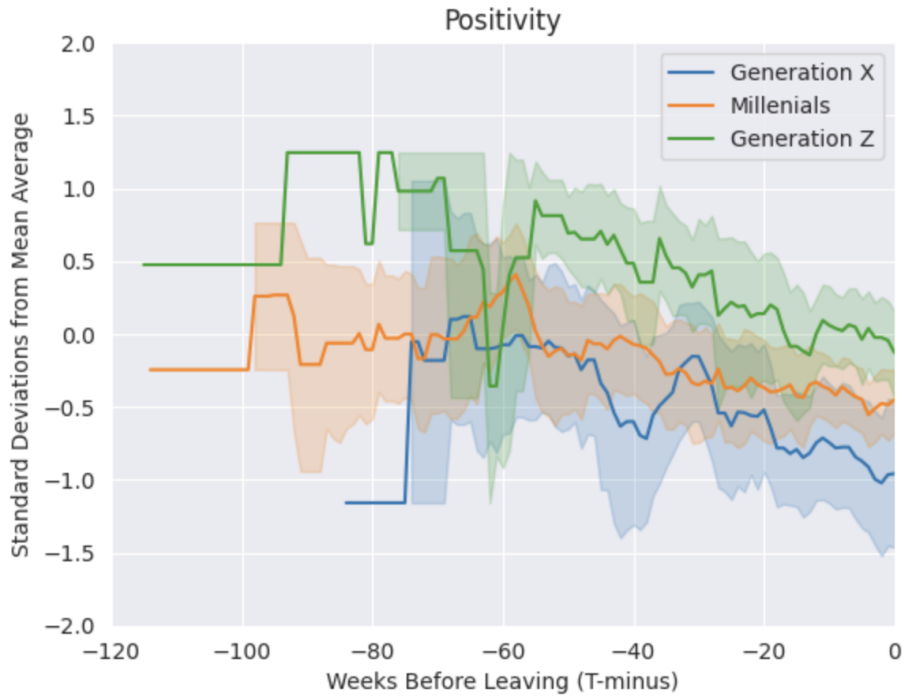


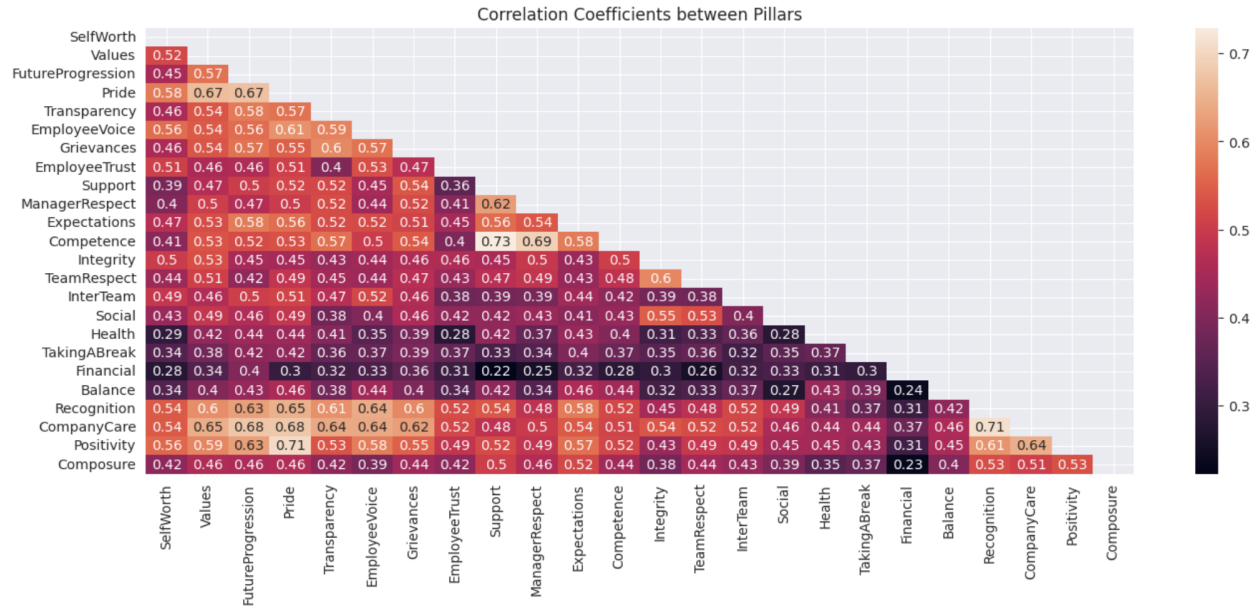
Figure 67: Positivity by Generation



Moreover, research from Gallup (2023)[5] shows that those under 40 years of age are more likely to actively look for other jobs, even if they are perceived to be happy in their current job - which may support this hypothesis (38% of under-40s compared to 31% of over-40s in Europe 2022).

Correlations between certain pairs of Core24 metrics may provide insight into character traits of professionals that employees associate with each other.

Figure 68: Correlation between Core24 metrics



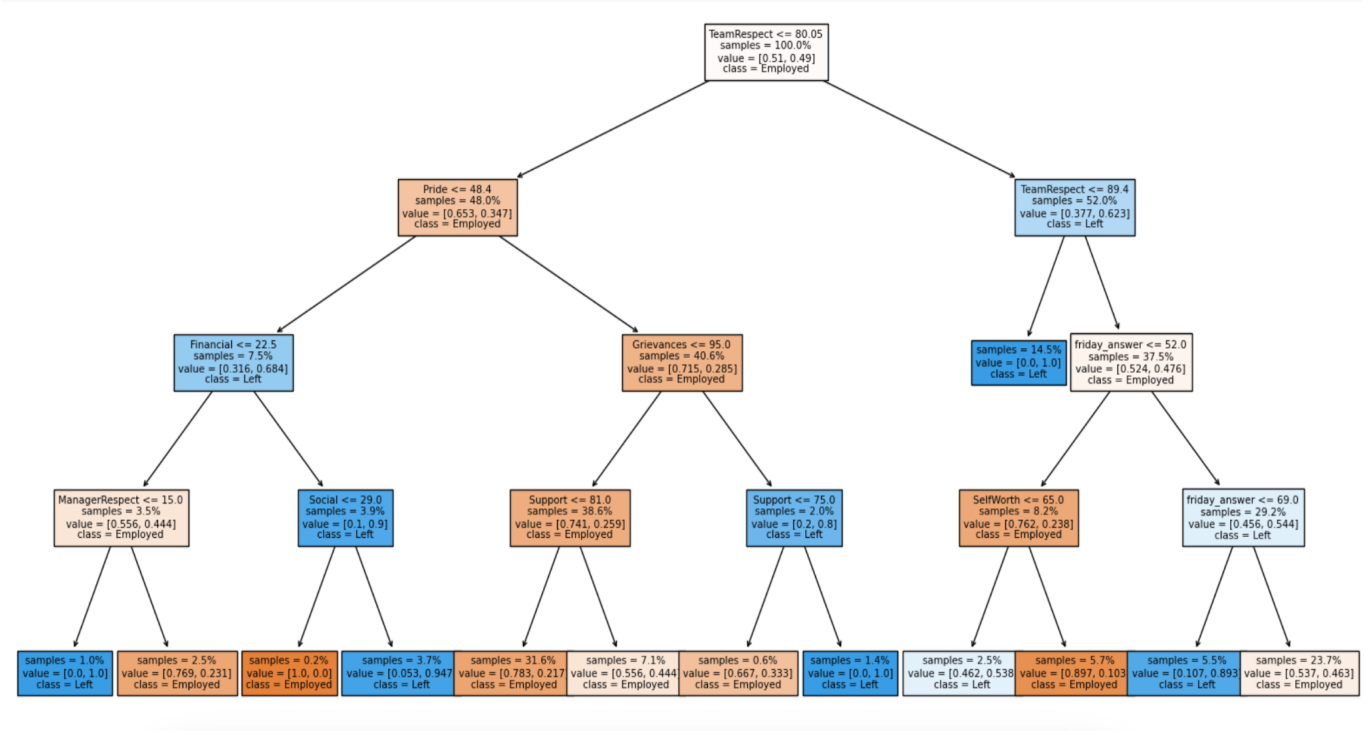
For example, the highest correlation coefficient was between Support and Competence (+0.73). This could indicate that employees feel that the competence of their respective managers is highly associated with the amount of support they receive from them. The joint-second highest correlation coefficient was between Pride and Positivity (+0.71), which is indicative that employees associate how proud they are to work at a given company with how much they look forward to coming to work. Similar relationships were found between how much employees feel cared for by their employer (Company Care) and recognition of their work (Recognition); and manager competence (Competence) and respect felt for that manager (+0.71 and +0.69 respectively). Interestingly, and possibly due to the UK cost of living crisis, Financial Wellbeing correlated least with other Core24 metrics.

What do we know about professionals that leave a job?

Team Respect and Pride may prove to be initial discriminating factors in whether an employee will quit a job or not.

Initial Decision Tree modeling indicated Team Respect and Pride to be the 'initial splitters' - that is, the initial metrics used to split the data set for classification. This may be indicative of these metrics being important in the initial stages of classifying whether or not an employee may be about to leave.

Figure 69: Decision Tree Model (Non-PCA)



This non-PCA decision tree model can show groups of Core24 pillars that are related as you move down the tree. Some interesting and rather 'intuitive' metrics that could possibly be grouped through this model are Team Respect, Pride, Grievances and Support Levels.

What do we know about professionals that leave a job?

Potential Triggers may hold the key to predicting whether or not an employee will quit a job from a very early stage.

The Core24 metrics listed as 'potential triggers': A grievance, inability to take a break/take leave, sudden lack of support, drop in expectations or a drop in perceptions of competence - seem to be the earliest indicators of potential problems for employees that lead up to them quitting and becoming disengaged in their work. Note that these events may also provide opportunity to support these workers from this very early stage and reduce staff turnover.

Discussion

Analyzing the relevant data led to a 'Theory of Quitting' which encompasses what current data from Wotter is telling us about what may happen in the time periods leading up to an employee leaving a company. It reads as follows:

1. **An initial problem arises** (i.e. a 'Potential Trigger') for a given employee in a given company, resulting in dramatic falls in certain Core24 metrics (such as Grievances). This happens approximately 80 weeks before the employee eventually leaves.
2. **Management formulate and propose a short-term solution** for the employee with the problem. In terms of the Core24 metrics, trends reverse and users exhibit gradual uptrends in their data as they answer questions for around 20-30 weeks.

What do we know about professionals that leave a job?

3. **The employee realizes that the solution implemented from management is ineffective and/or inadequate.** This stage has a tendency to begin around 50-70 weeks before the employee eventually leaves the company. At this stage, trends in the 'Deteriorators' reverse once more to begin a second downtrend.
4. **Core24 metrics gradually deteriorate** over the course of around 50 weeks. This trend is most prominent in the 'Deteriorator' metrics (see above).
5. **Employee quits.**

As stated in the Abstract, this theory must be tested more rigorously as the sample size for employees who have left a job is currently very small (319 samples). In addition, some data was imputed through a k-Nearest-Neighbors method which estimated values for metrics registered as NaN (not a number) values based on the 5 most similar data points in the dataset. Using such a method introduces bias into the dataset.

References

1. Chauhan, N. S. (2022, February 9). *Decision tree algorithm, explained*. KDnuggets. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html
2. Zach (2022, May 9). *How to Interpret the Classification Report in sklearn (With Example). Statology. https://www.statology.org/sklearn-classification-report/*
3. Cheng, C. (2022, March 22). *Principal component analysis (PCA) explained visually with zero math*. Medium.

https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d

4. Chowdhury, K. R. (2020, July 20). *KNNImputer: A robust way to impute missing values (using scikit-learn)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/07/knnimputer-a-robust-way-to-impute-missing-values-using-scikit-learn/
5. Gallup. (2023). *State of the global workplace report*. Gallup.com. https://www.gallup.com/workplace/349484/state-of-the-global-workplace.aspx?campaignid=18945816141&adgroupid=143633586437&adid=635680356863&gclid=Cj0KCQjwoK2mBhDzARIsADGbjeo7_bgPs4utvvxcSBt_HGYGPnsrzA66_sjX4s2U_JVZh0iPlvgGWf0aAjuMEALw_wcB