Wharton People Analytics Case Competition 2022

By: Kevin Xu and William Harkless

Abstract

Our research regarding the impact of the Great Resignation on the University of Pennsylvania's workforce and community suggests the manifestation of the Great Resignation on Penn's campus. Analysis of Penn employee data shows trends of higher resignations starting at the beginning of 2020 until present year, 2022. Employees are following this resignation trend regardless of identity and demographic. Although, we will see that some groups follow the trend more strongly than others. The highest predictors of resignation amongst employees are age, salary, and highest degree attained. We recommend that Penn's HR department prioritizes establishing a clear evaluation and advancement pipeline for all of its employees, particularly those who are new and among the lower paid workers, allowing employees to benefit from their efforts and experience, while the university can retain its employees better without the need to rehire and retrain.

Methodologies and Techniques

Cleaning and Wrangling

We took the four datasets provided by Penn's HR Division and cleaned the data for analysis. Our wrangling process centered around the unique employee IDs. To understand the trends surrounding attrition holistically, we thought it would be best to get a better understanding of each individual's characteristics. Ideally the cleaned data would contain a sufficient number of employees to generate statistically sound insights, and it would contain complete and organized data for each sample.

Each dataset contained many columns with a mix of quantitative and categorical data. To structure the data for future visualizations and predictive models, each category was one-hot encoded for each employee. This allows our machine learning algorithms and graphics to process categorical data more efficiently. After hot encoding each categorical column in each data frame, we cleverly constructed new columns that might be more informative for humans and models. Next, after carefully structuring and cleaning the data, we created a data frame that captures all of the characteristics of each employee by joining the modified tables. Lastly, workers that were involuntarily terminated were filtered out from the table. We only want to analyze the impacts of The Great Resignation, so workers who voluntarily work or leave are our target population.

Our cleaned data now consists of approximately 8,400 unique employees that have voluntarily stayed or departed from Penn since 2019 to present day. This data frame is what we will use to generate insights in the form of graphics and predictive models.

Visualizations

Generating intuitive and informative graphics using the data was fairly straightforward. Using the specially formatted time data in our table, we graphed the rolling average percentage of employees that resigned in order to view broader trends in resignations while smoothing out fluctuations found on a monthly basis (see Figure 1). This graphic clearly demonstrates the trend of resignations across recent years. Next, we decided to group employees based on sex and measure the rates of resignation over the recent years (see Figure 2). We used our formatted time data along with grouping functions to plot this figure. We followed the same procedure to produce a similar visualization measuring workforce termination rates based on race (see Figure 3). Lastly, we thought grouping employees by job type might produce useful and interesting takeaways. We produced two figures based on this notion. The first is a bar plot that shows the resignation rates among different job types (see Figure 4). The second is time series data measuring the termination rates for some of the jobs with the highest and lowest resignation rates (see Figure 5). We do the same for resignation rates based on age ventile (Figure 6 and Figure 7), as well as salary decile (Figure 8 and Figure 9). These are key graphics for our machine learning models as well, but we will discuss those more below.

Models

Keeping in mind our goal of analyzing The Great Resignation, we thought it would be very helpful and informative to reliably predict whether or not an employee will resign based on their characteristics. However, we would like to formally state that algorithms and models should not be used thoughtlessly when making hiring decisions. Selecting employees based purely on identity and surface characteristics would not be just or fair. Using algorithms to make important decisions can lead to self fulfilling prophecies and further discrimination and inequality. Our goal is to learn objectively from the provided data using machine learning techniques.

Returning to our original goal of understanding resignation at Penn, we decided to build many classification models. Our models classify employees as either likely to resign or unlikely, a classic binary classification model.

First, we specified the inputs and output to our models. The output is one or zero. One represents that an employee is likely to resign, and zero represents the opposite. Our inputs are all other available characteristics about the employee. Next, we split our data into training data and testing data. Seventy-five percent of the data is for training the models, and the rest is for testing.

Our baseline model was an easy to implement logistic regression model, which is mediocre at binary classification problems. It performed mediocrely as expected. The best performing logistic regression model used L2 regularization for selecting the most important features (inputs). The training accuracy and testing accuracy were 64.9% and 64% respectively.

The best performing models were the tree models. We decided to use Grid Search and Cross Validation to tune each model's parameters. Grid Search tries each combination of specified parameters for the model, and Cross Validation measures the success of each combination parameters on a random piece of the training data. By the end of the algorithm the model will use the parameters that produce the highest accuracy. The random forests classifier was the best performing model overall, and the gradient boosting classifier was the runner-up. The training accuracy and testing accuracy for the random forest classifier was 96% and 83.2% respectively. The accuracies for the gradient boosting classifier were 95.7% and 81.7% respectively.

Lastly, we trained support vector machines (SVMs) with different kernel functions. Since training SVMs and tuning parameters for each model with Cross Validation takes a significant amount of time, we had to get creative. First, we reduced the amount of features using Principal Component Analysis (PCA) which captures information about the most informative features. We used Randomized Search Cross Validation instead of Grid Search to speed up the tuning process. The best SVM used a radial basis function and had training and testing accuracies of 65.4% and 62.4% respectively.

Key Findings

Penn's workforce is definitely being affected by the Great Resignation. If we look at the rates of resignation in the U.S. and at Penn, we see that they follow the same trend. They both plummet a little after 2020 and sky rocket after 2021 leading into 2022. Due to the difference in population size the rates aren't the same, but identical trends suggest that the Great Resignation exists at Penn (see Figures 1 and 13).

When we separate resignees by race and sex, we found that the trends are very identical. Both sexes resigned at the same rates at the same time with very slight differences. Sex A seems to resign at a marginally higher rate than sex B, and sex A has a 3% difference in resignation rate in January 2022 (See Figure 2). We find the exact same phenomenon when looking at resignees by race. They follow the exact same trend with small differences. Race C peaks at about 9% in the summer of 2021. Thus, it seems as though sex and race have very little, or only marginal effects on resignation rates among employees.

Grouping resigned employees by job type yielded interesting results. First, we saw that certain job types have higher resignation rates than others. There are two rationalizations for this phenomenon. The first is some jobs at Penn are more secure and beneficial than others, so employees are hesitant to resign. Jobs with higher resignation rates could also be for workers that are contract-based and have less obligations to stay beyond the normal school semester. The other rationale is that higher paying jobs provide more financial freedom, which could enable those employees to risk leaving Penn's workforce. We can't say for sure since the job types are being kept ambiguous, but later looking at the data separated based on salary decile provides more insight into this.

We also looked at time series data for the four job types with the highest rates of resignation. These jobs saw a rising behavior of voluntary terminations around the same time that the Great Resignation saw the trend at Penn and in the U.S.

Resignation rates separated by salary decile (Figure 7) allow us to piece together that jobs with higher salaries have lower resignation rates and this holds when comparing any pair of salary deciles, indicating a strong correlation between pay and employee attrition. We can see in Figure 8 that almost regardless of the time of the year, the lowest two salary deciles have significantly higher voluntary termination rates than those in the highest two salary deciles. There are noticeable peaks of resignations for the lowest paid employees every July, but the more recent July peak is much higher. We notice that for the higher paid employees, this peak is much less noticeable or almost non-existent, indicating that perhaps the Great Resignation has had a much less noticeable effect on the highest paid at Penn.

Voluntary termination rates grouped by age ventile was the last grouping we analyzed (Figure 5 and 6). Here, we see that generally, younger employees had much higher termination rates than older employees. In Figure 6, we clearly see large peaks again in July for the younger employees while the older employees have much smaller resignation waves. Once again, we can see significantly more resignations last July than the previous year, reflecting the Great Resignation trend.

The most interesting insights came from our two best classifiers, the Random Forest Classifier and Gradient Boosting Classifier. We ranked the importance of each feature in each model. For both classifiers the most important predictors of resignation were age, salary, and the highest degree obtained in that order (see Figures 10 and 11).

Recommendations

Based on our model results as well as the visualizations of the data grouped by various characteristics, we can draw a few conclusions from the data that can inform future decisions by the HR department at Penn. First, we can see that Penn is likely feeling the effects of the Great Resignation on its workforce. We can see that even when comparing resignation rates between the same times of the year, 2021 showed much higher resignation rates among almost all groups of employees compared to 2020. What this means is that if Penn does not provide better incentives to their employees, they likely will have a difficult time retaining them moving forward.

Another insight is that younger employees are the most likely to leave Penn for work elsewhere and especially those who are the lowest paid. One possible explanation for this trend is simply that younger employees have more mobility and flexibility in terms of the jobs they take on and have less difficulty shifting careers. Thus, it is easier to draw them away to some other job or career offered elsewhere. To combat this beyond simply raising wages, Penn HR needs to look at ways to improve career advancement and mobility among its employees, especially those early in their careers. Look to create mentorship programs and accelerated career development pipelines that will provide greater opportunities for employees to advance in their careers. This will make it more enticing for new workers to stay at Penn longer and reap the benefits of their continued involvement with the community, rather than job-hop to another company. Job security and concrete advancement plans will benefit both the university and its employees, as the employees will have clear directives on how to perform well in their job and quickly move up as they gain more experience and the university will not have to spend as many resources hiring new workers and training them to the level of the previous employees we resigned.

Appendix

Figures and Visualizations

Figure 1

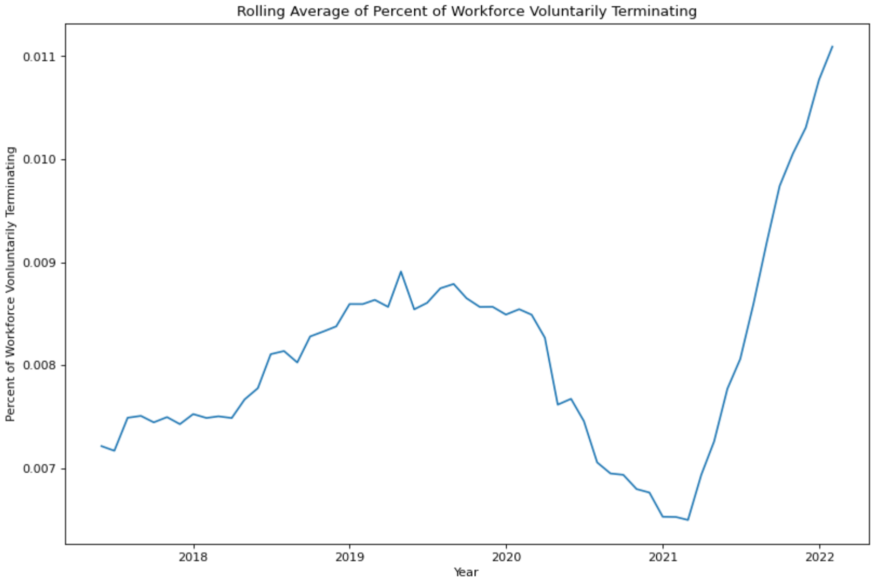


Figure 2

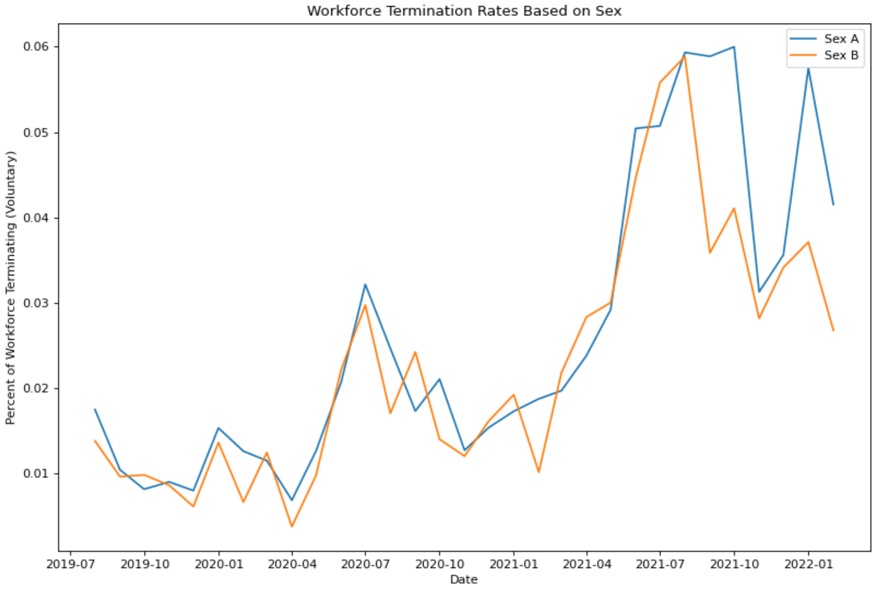


Figure 3

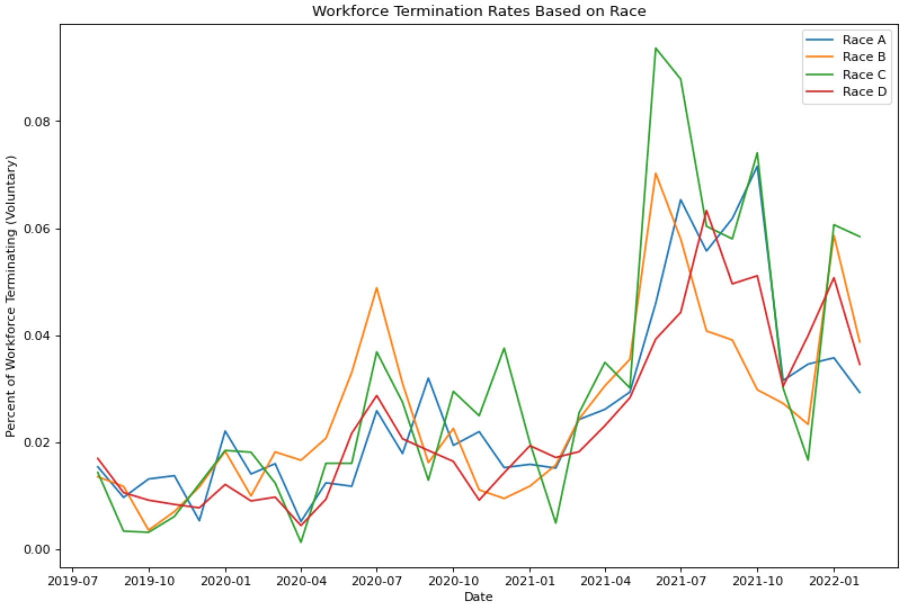


Figure 4

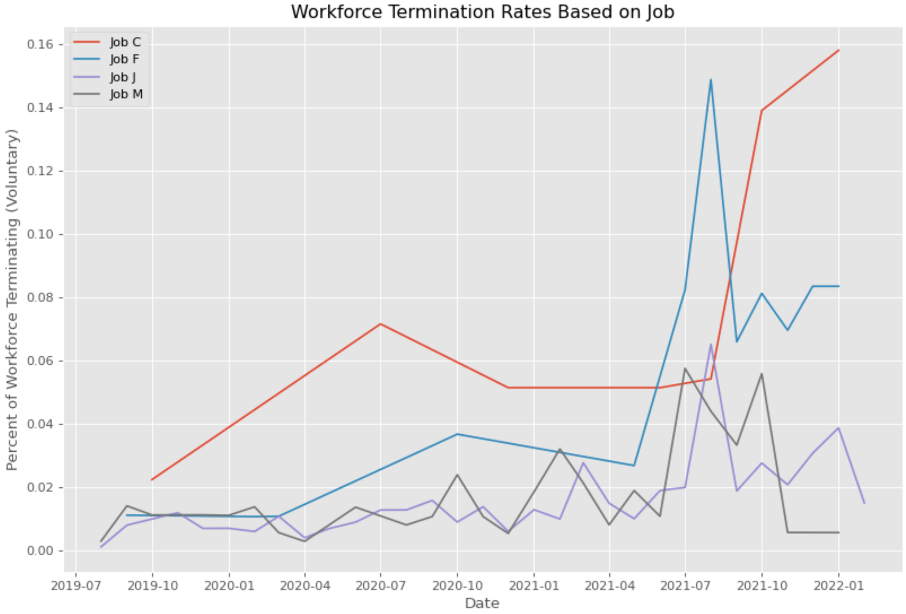


Figure 5

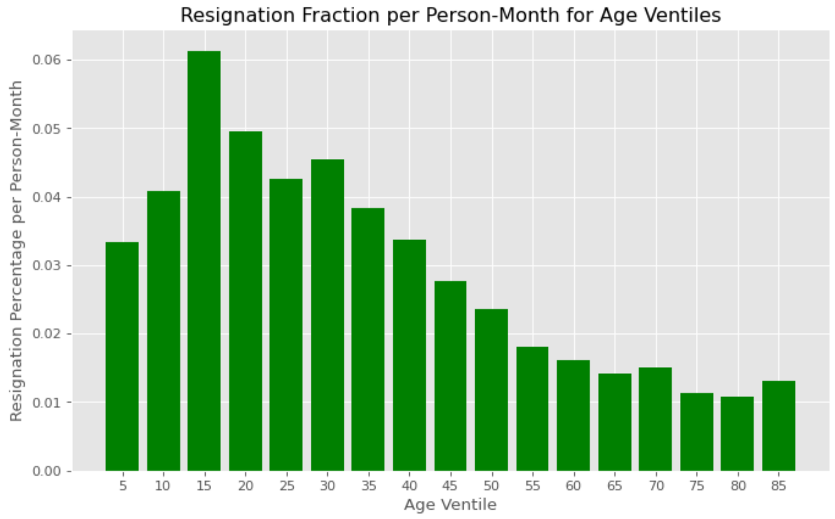


Figure 6

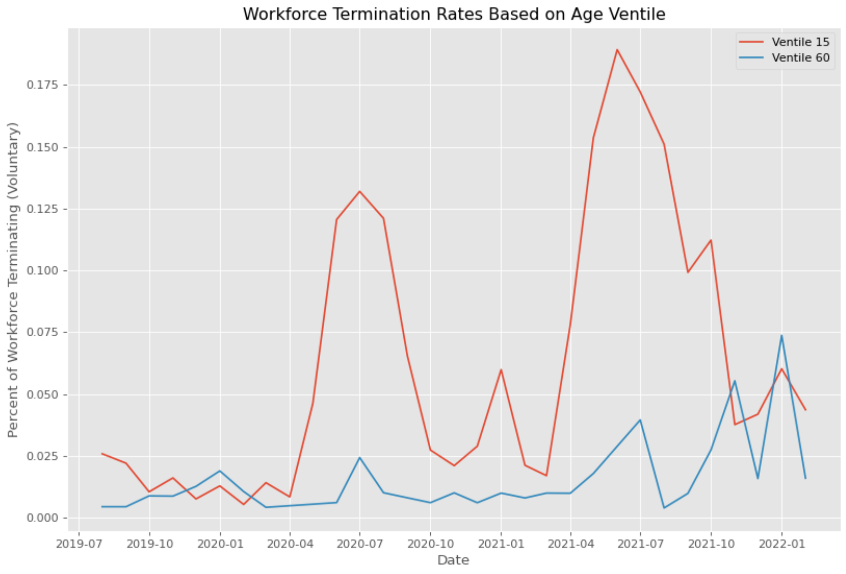


Figure 7

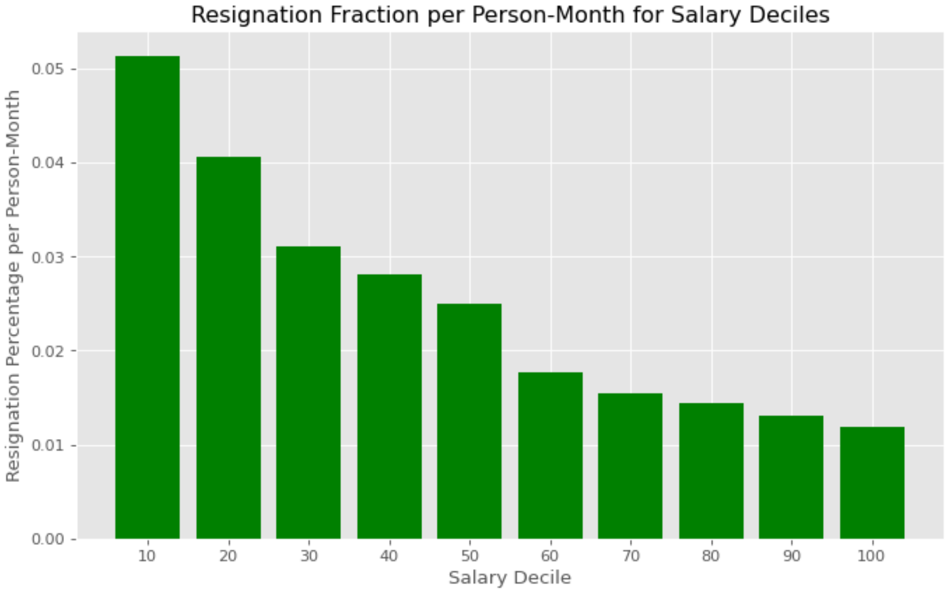


Figure 8

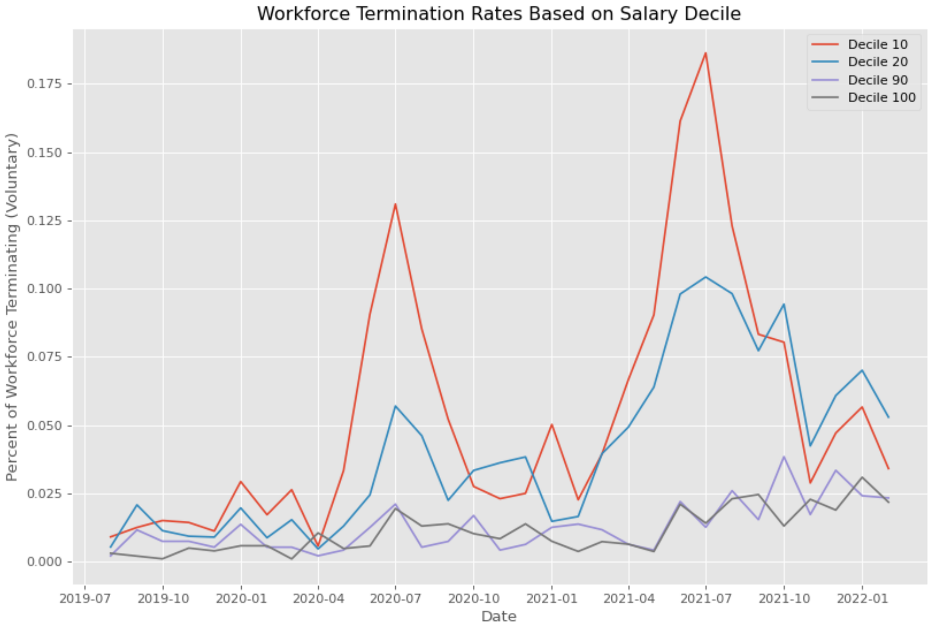


Figure 9

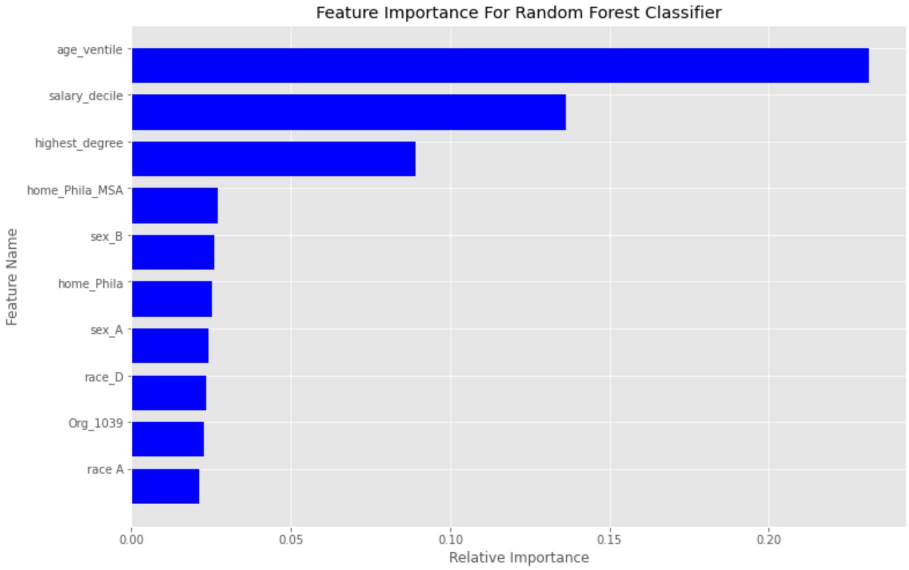


Figure 10

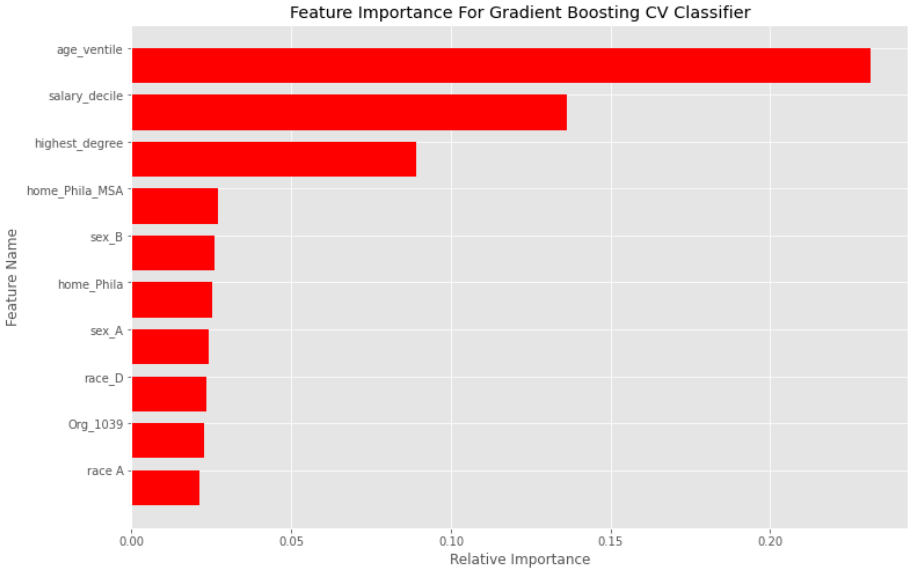
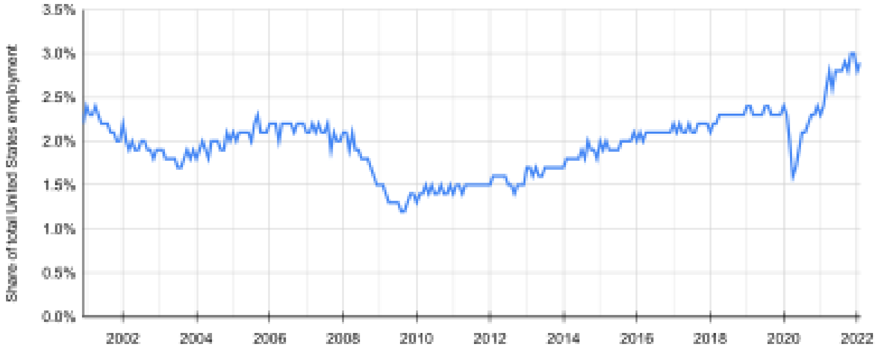


Figure 11 (from: https://en.wikipedia.org/wiki/Great_Resignation)


Share of total United States employment