October 23, 2023

Srinvas poluchalla

People analytics

# Introduction

The article explores and evaluates the People analytics dataset for an organization. The source and authenticity of the data is not known. It is part of the course exercise to understand features contributing positively and negatively to employee attrition. The dataset size is very small and has only 150 rows of data in that 59 rows belong to the employees who quit. Firm conclusions drawn from this analysis are questionable. Hence this article should only be treated as an exercise document*.*

# Problem

High employee turnover costs organizations to retrain and recruit.

# Causes

The analysis explores various features that contribute an employee to quit.

# Solutions

* Identify at-risk employees
* Plan corrective measures
* Design products and jobs to match employee’s expectations

# Raw Dataset

# 

# Hypothesis questions

The data analysis is performed with hypothesis questions in mind to avoid confirmation bias.

1. Would the employee's age play a significant role in quitting?
2. How bad is job monotony?
3. Does pay matter?
4. What are the positive and negative features influencing decision-making to quit?

# Assumptions

1. Bachelor’s degree completed at the age of 21 years, and those who have no bachelor’s degree joined the workforce at the age of 18 years.
2. Column Quit this year, 1 means Quit, 0 means stay.
3. Column Applied for the internal job last year?,

1 means Yes, 0 means Not applied

1. Column Bachelor’s degree, 1 means Yes, 0 means No bachelor’s degree.
2. Column Performance rating, Manager Performance rating, 0 = No ratings, 1= development needs, 2 = Below expectations, 3 = Meets expectations, 4= means exceeds expectations, 5 = Outstanding

# Visuals

## #1 Hypothesis: Age pool

H0 = Bachelor degree holders and non-bachelor degree holders’ ages are the same.

Ha = Statistically failed to conclude the means are of the same age

With assumption1, an age column has been added to the dataset. The age is calculated, for bachelor’s degree holders 21 years + prior experience + organizational tenure, similarly for non-bachelor’s degree holders 18 years + prior experience + organizational tenure has been added.

Chart, box and whisker chart

Description automatically generated

A T-Test was performed to compare means/ variances of ages between bachelor’s and non-bachelor’s degree holders.

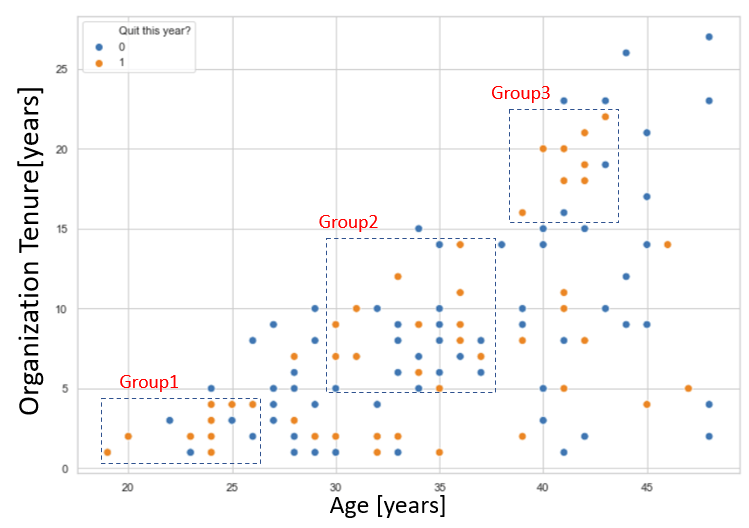
The test failed to conclude the null hypothesis, thus the age pool below the same population. [ The data drawn from the organization belong to the traditional industry and not from the start company where the age demography seems to be similar age pool]

The next logical question comes to mind: Does age play a role in deciding to quit or stay?Chart, box and whisker chart

Description automatically generated

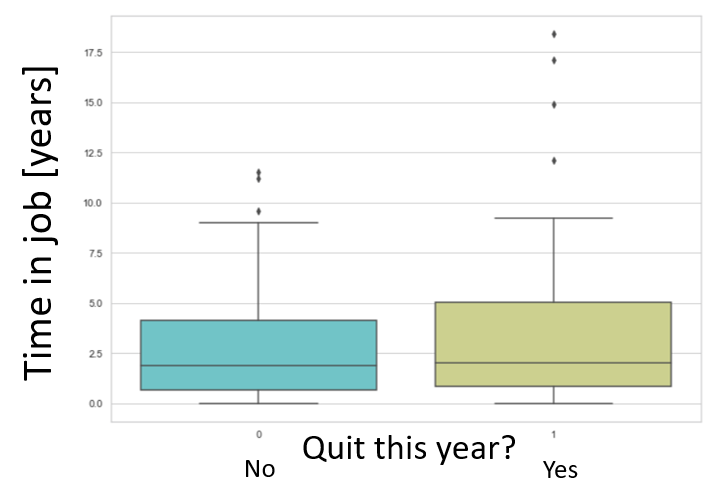
From the T-test, we can conclude that age does not play a role in deciding to quit or stay within the organization.

Another way of looking at the data is to see which age groups have high employee turnover. The figure below shows orange dots in various groups; the Group 1 has just joined the organization, and the employees are getting accustomed to the culture.

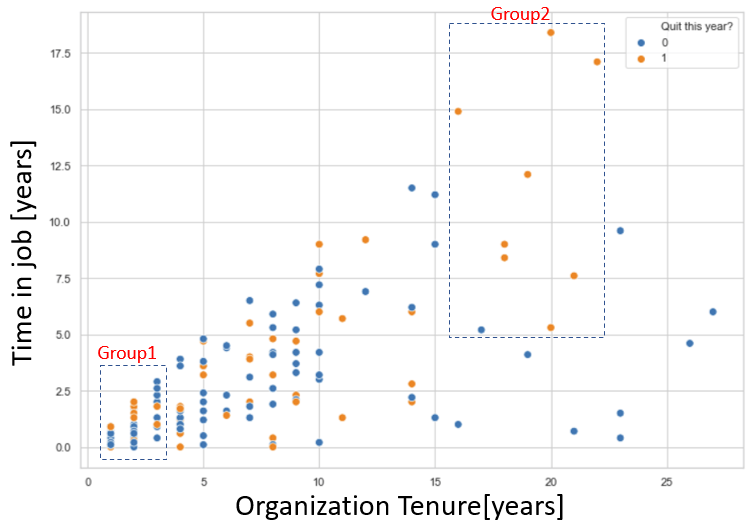
Clusters in groups 2 to 3 are mid-career with considerable experience within the organization. Keeping groups 2 &3 employees within the organization is essential.

## # 2 Hypothesis: How bad is job monotony?

Longer the “time in job”, could lead to job monotony, no recognition, and excellent job opportunities outside of the organization.



Statistically, not able to conclude job monotony leads an employee to quit. Partly because there are several outliers of longer “time in job” [high variance in the data] who have made the decision to quit.



Group1 data cluster, employees who came into the organization and started a new job, decided to quit within three years.

Insight questions:

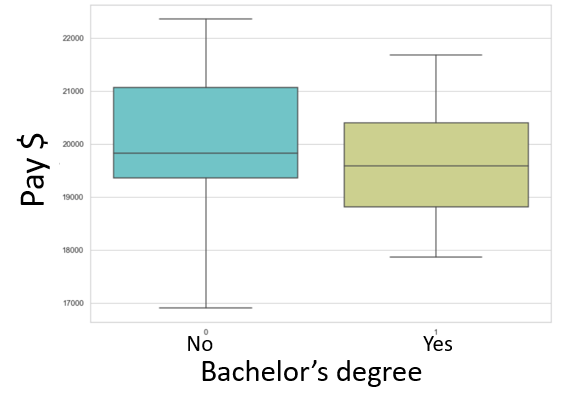
* Belonging issue or cultural mismatch?
* Onboarding and mentoring could play a role.
* Training and team-building activities
* Opportunities to shine
* Are they not challenged?

Group2: group has a strong tendency towards job monotony, lack of other opportunities to create impact, better market conditions, and the essential employee needs are addressed. At this age, the employee is looking for self-esteem or self-actualization.

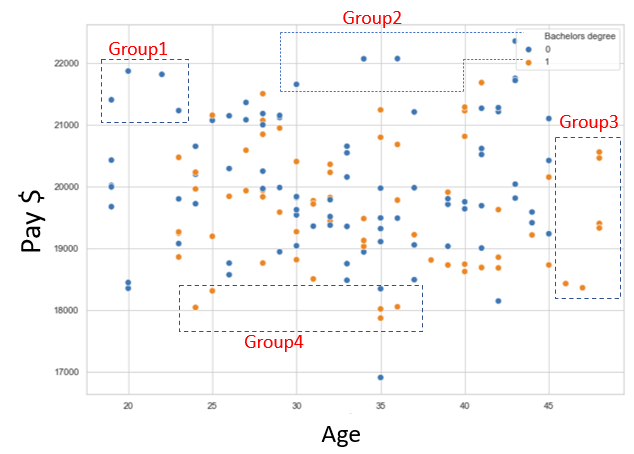
## # 3 Hypothesis: Does pay matter?

H0 = The mean pay between bachelor’s degree holder and non bachelor’s is same.

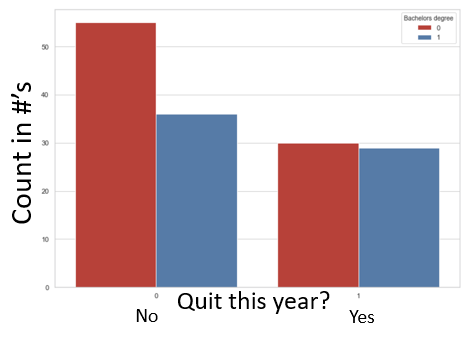
Ha = Statistically failed to conclude the means are of the same



There is a statistically significant difference in pay between nondegree holders and degree holders. The degree holders are paid lower!



Look at groups 1, and 2 with no bachelor’s degree are paid more than bachelor’s degree holders. Similarly, group 3 and 4 shows that bachelor’s degree holders are underpaid.



Another aspect of looking at the data is who quit the most. Do degree holders or non-degree holders [ Caution: - There is an imbalance in counts between degree and non-degree holders]

# Visuals

Before starting any analysis, it is essential to look at the correlation between each parameter in the form of a heatmap.

The number of features that show a strong correlation with each other can be removed from the analysis.



The features that show some correlations are age, Prior Experience, and Organizational tenure. The age is derived from the data of prior experience and organizational tenure hence co-relation is stronger.

The other stronger correlation is between ‘Quit this year?’ and ‘Applied internal job last year?’

Since none of them show a strong correlation, hence all the features need to be considered for further analysis.

# Linear regression analysis

Linear regression analysis is performed, the features with p-value <.05 are shown in the figure below.

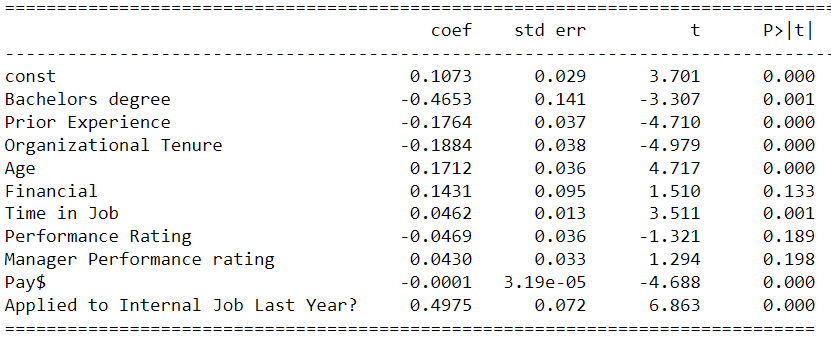
The goodness of fit, R Squared value = 0.376, is low. [Caution: Avoid drawing strong conclusions from the analysis]

Features that make an employee quit [Positive features]

* Intercept [constant]
* Age
* Time in job
* Applied to internal job last year?

Features that make an employee stay in the organization [Negative features] include

* Bachelor’s degree
* Prior experience
* Organizational tenure
* Pay



# What can I do now?

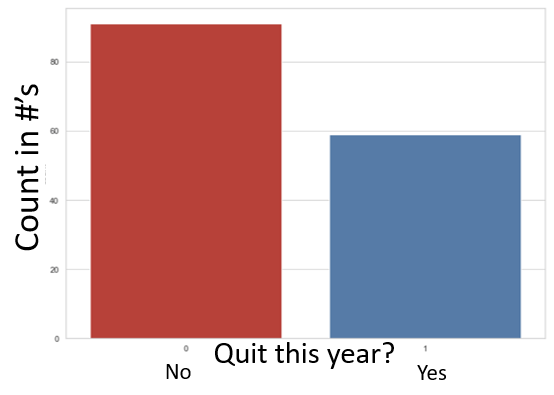
First

* Identify potential employees who are at risk of quitting.

Second

* Adopt corrective measures to improve positive features

# Baseline model:



Probability of predicting an employee would quit = [# no of employees who quit/ Total #of employees] 100%

In this case, the baseline probability of predicting an employee would quit = 39.33%

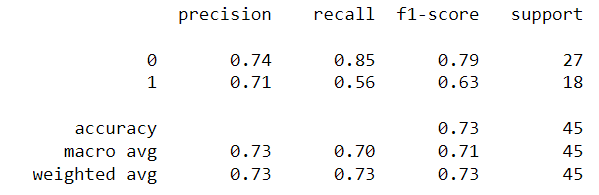
The classification model needs to predict better than 39.33%.

# Logistic regression to classify

The existing data can be split into train & test groups [80% train & 20% test]. The train data is fed into the model inorder to fit the data.

Once the model is trained, the test data can be used to evaluate the performance of the model.

The classification summary report is used to measure the quality of predictions from the algorithm.



Precision: The ratio of true positives to the sum of true and false positives.

Recall : The ratio of true positives to the sum of true positives and false negatives

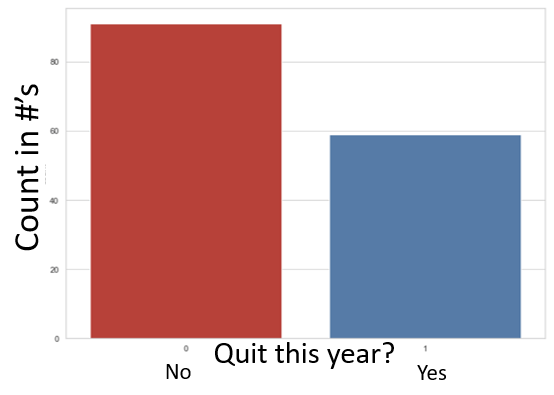
F1 score: Harmonic mean of precision and recall.

From the above summary report, the algorithm’s accuracy in predicting an employee would quit is close to 73%.

# Bias

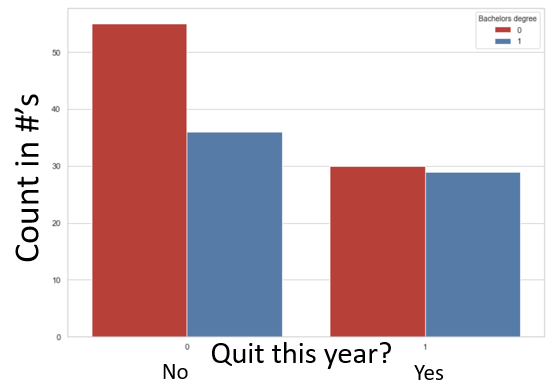
### Imbalance data Quit/ Stay

The number of quit to stay ratio is not 50:50, which is ideal to have for classification problem. However, since it is a small population [150 only], drawing hard conclusions using this dataset is not advisable.



### Imbalance in Bachelors/ non bachelor’s

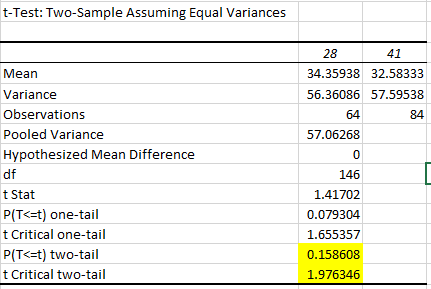
Sample size for non-bachelor’s is higher compared to bachelors.



# Appendix

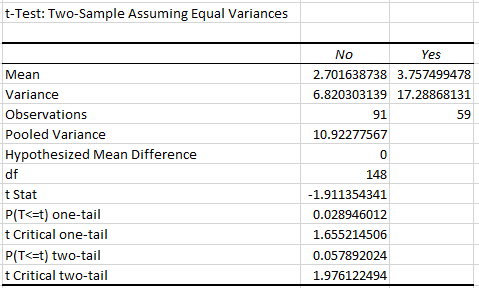
## # 1 Hypothesis

Age comparison between bachelor’s degree holders and non-bachelor’s degree holders



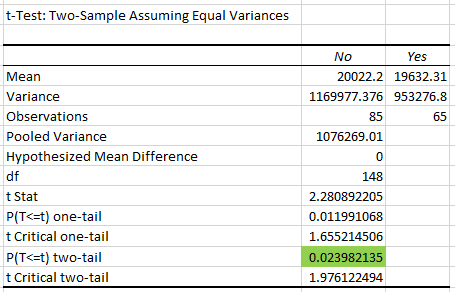
# #2 Hypothesis

Does Job monotony play a role for an employee to quit?

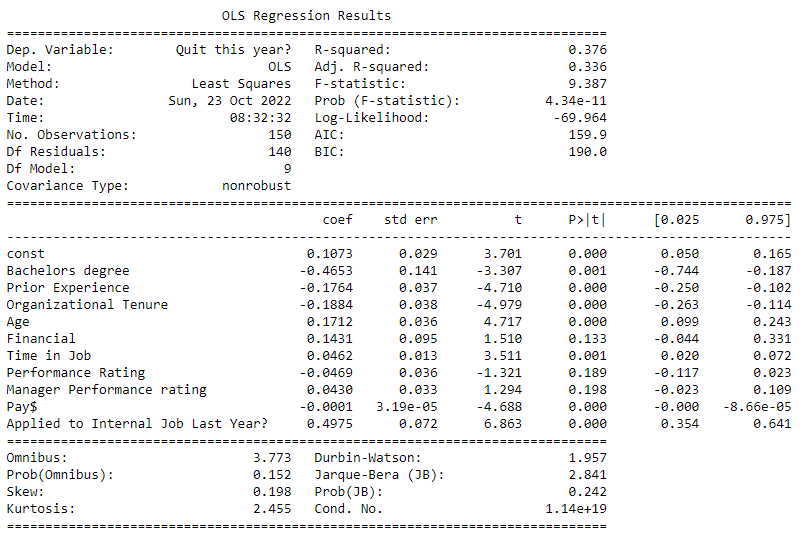


# #3 Hypothesis

Is there a pay difference between bachelor’s degree & non-bachelor degree holders?



## Linear regression



## Logistic regression

