Employee Attrition Analysis: Uncovering Insights to Improve Retention at Marvelous Construction

Problem Overview

Marvelous Construction, a major construction firm operating across 35 construction sites in Sri Lanka, has observed a concerning trend of employees resigning. The Human Resources department recognizes the urgency of this matter and seeks data-driven insights to understand the underlying factors contributing to attrition.

The goal of this project is to leverage the available dataset, extracted from the company's ERP system, to perform a thorough analysis and derive valuable insights that will inform strategic decision-making to enhance employee retention. By identifying the key drivers of attrition, we aim to provide actionable recommendations to the CEO and the management team.

Dataset Description

The file “employees.csv” contains details about the employees. It contains 997 records and 19 dimensions. The dimensions are Employee_No, Employee_Code, Name, Title, Year_of_Birth, Gender, Religion_ID, Marital_Status, Designation_ID, Date_Joined, Date_Resigned, Inactive_Date, Status, Reporting_emp_1, Reporting_emp_2, Employment_Category, Employment_Type, Religion and Designation. Date_Resigned and Date_Joined are for training and testing data sets for attrition prediction.

The file “leaves.csv” contains details regarding the leaves taken by employees. It contains 1018 records and 6 dimensions. The dimensions are Employee_No, leave_date, Type, Applied Date, Remarks and apply_type. It is to be noted that older employees have more leaves compared to newer employees.

The file “salary.csv” contains comprehensive salary details of employees. This file includes monthly additions/deductions breakdowns as well. This file contains 9035 records and 57 dimensions.

The file “attendance.csv” contains daily employee attendance. This file contains 224057 records and 10 dimensions. This includes the projects each employee is working on, shift times, in_time, out_time and Hourly_Time too.

Data pre-processing

To begin, I replaced the invalid values '0000' in the 'Year_of_Birth' column with NaN. This ensures better handling of missing values and maintains data integrity.

Next, I addressed invalid data in the 'Status', 'Date_Resigned', and 'Inactive_Date' columns. For all active employees, I set the 'Date_Resigned' and 'Inactive_Date' values to '\N'. Additionally, any instances of '0000-00-00' in the 'Date_Resigned' column were replaced with '\N' to improve consistency.

To handle missing values in the 'Year_of_Birth' column, I calculated the mode of the entire column as well as the mode for each unique 'Designation'. I then filled in missing modes in the 'Designation' groups with the overall mode, ensuring that every designation has a mode value.

To remove inconsistencies in the gender column, I assumed that the title column was correct and made imputations based on that assumption.

Using the mode values by designation, I created a dictionary to impute missing 'Year_of_Birth' values based on each employee's 'Designation'. This approach allows for more accurate imputation based on the specific designation as age is correlated to the designation of the employee.

For further analysis and modeling, I created a copy of the preprocessed DataFrame, named 'chatterbox_new'. This DataFrame included a selected subset of columns: 'Employee_No', 'Title', 'Gender', 'Religion_ID', 'Marital_Status', 'Designation', and 'Year_of_Birth'.

To encode the non-numeric variables, I applied label encoding using the **LabelEncoder()** function to transform 'Title', 'Designation', 'Gender', and 'Marital_Status' into numerical representations.

Next, I divided the dataset into 'train' and 'test' portions with the intention of imputing the missing values in the ‘marital_status’ column. The 'test' DataFrame contained rows with missing 'Marital_Status' values, while the 'train' DataFrame contained rows with known 'Marital_Status' values.

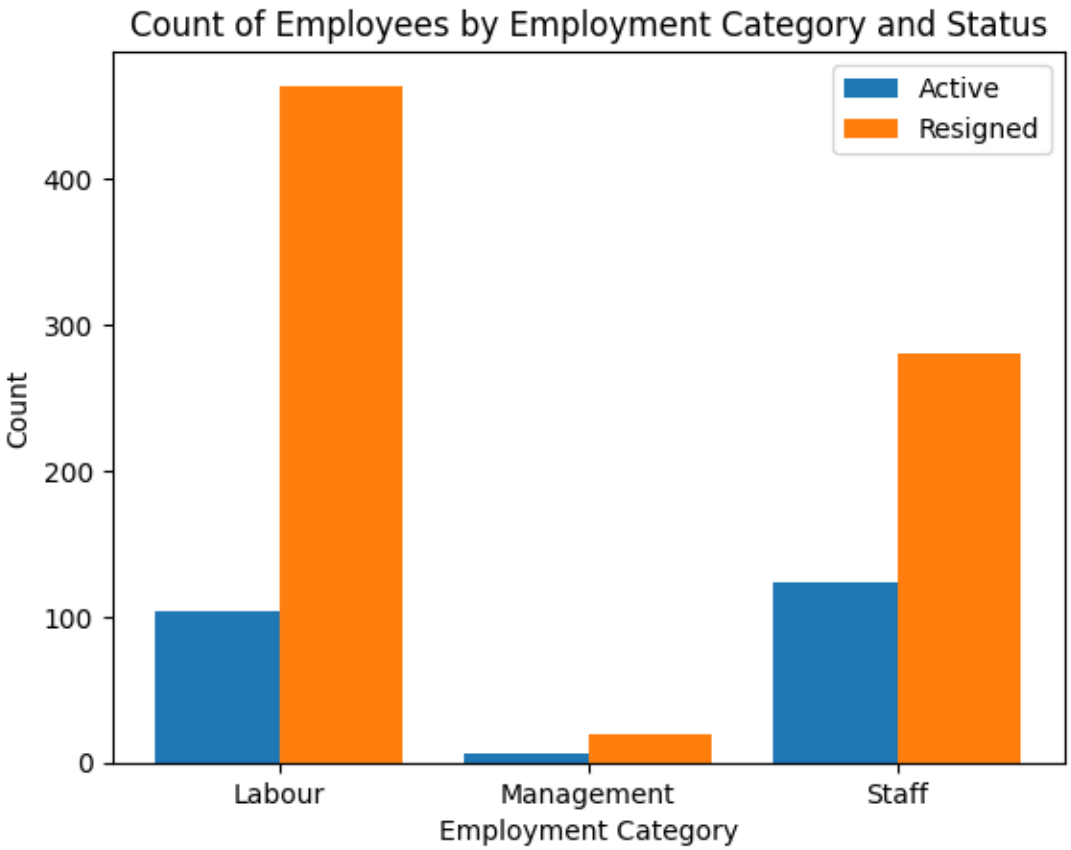
I then trained a Random Forest Classifier model on the 'train' DataFrame, using the features in 'X' and the target variable 'y', to predict missing 'Marital_Status' values in the 'test' DataFrame. This decision was taken due to the observed correlation between dimensions. This model yielded more than 75% of accuracy, which I considered acceptable.

After predicting the missing values, I remapped the numerical predictions back to their corresponding labels, 'Married' and 'Single', in the 'Marital_Status' column.

To incorporate the imputed values back into the original 'chatterbox' DataFrame, I assigned the imputed 'Marital_Status' and 'Year_of_Birth' values from 'chatterbox2' to their respective columns.

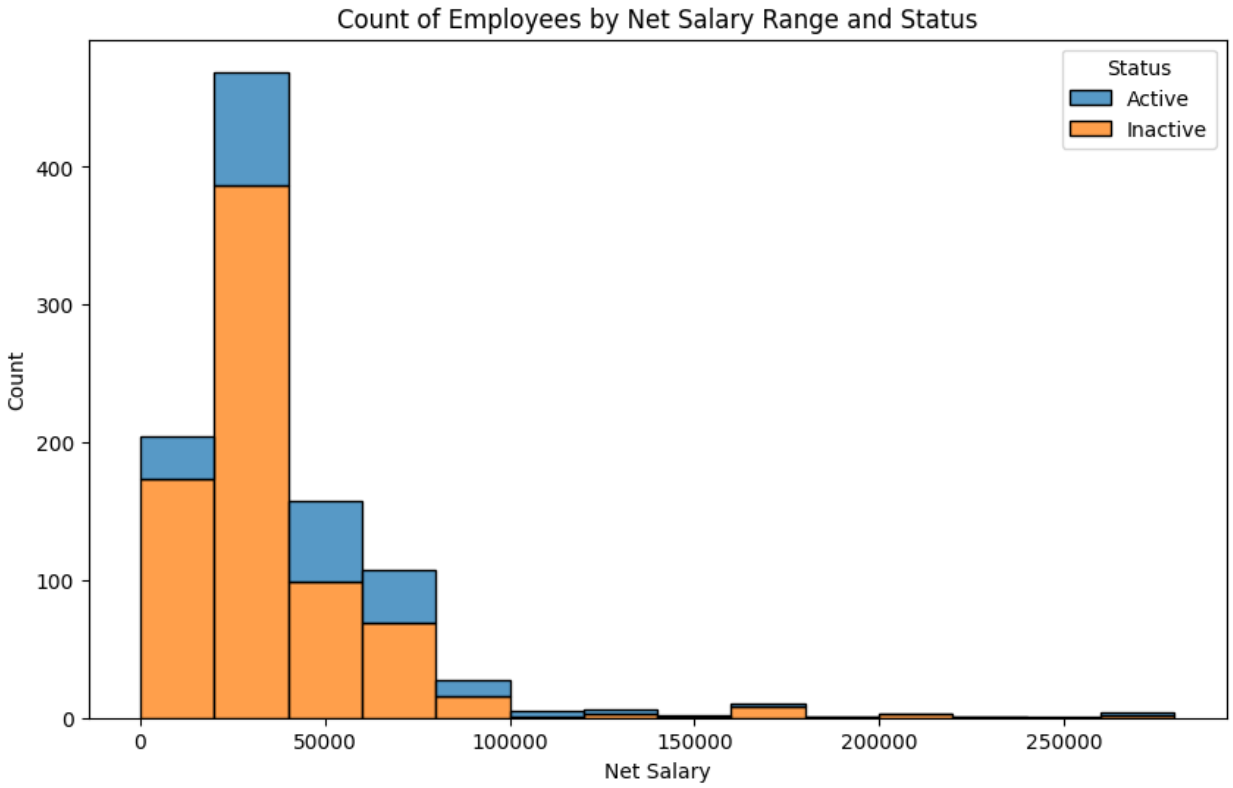
Insights

1. A higher proportion of laborers have resigned.



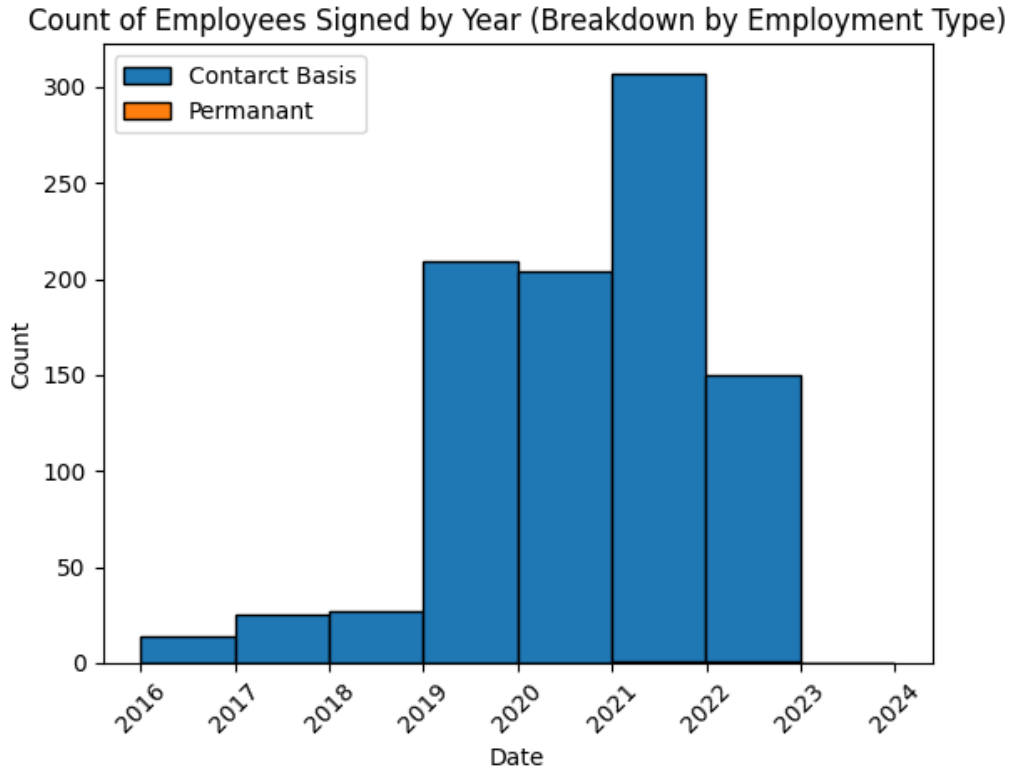
As evident in the above graph, a higher proportion of labor category employees have left the job. Several reasons could be behind this. Low pay, an unsuitable working environment could be among them. It is important to investigate this situation further to pinpoint the exact reasons why a higher proportion of laborers a leaving the job compared to other categories. A high proportion of the staff have also left their positions. Salary and poor management could be reasons for this.

2. **Employees in the lower salary ranges have resigned more.**



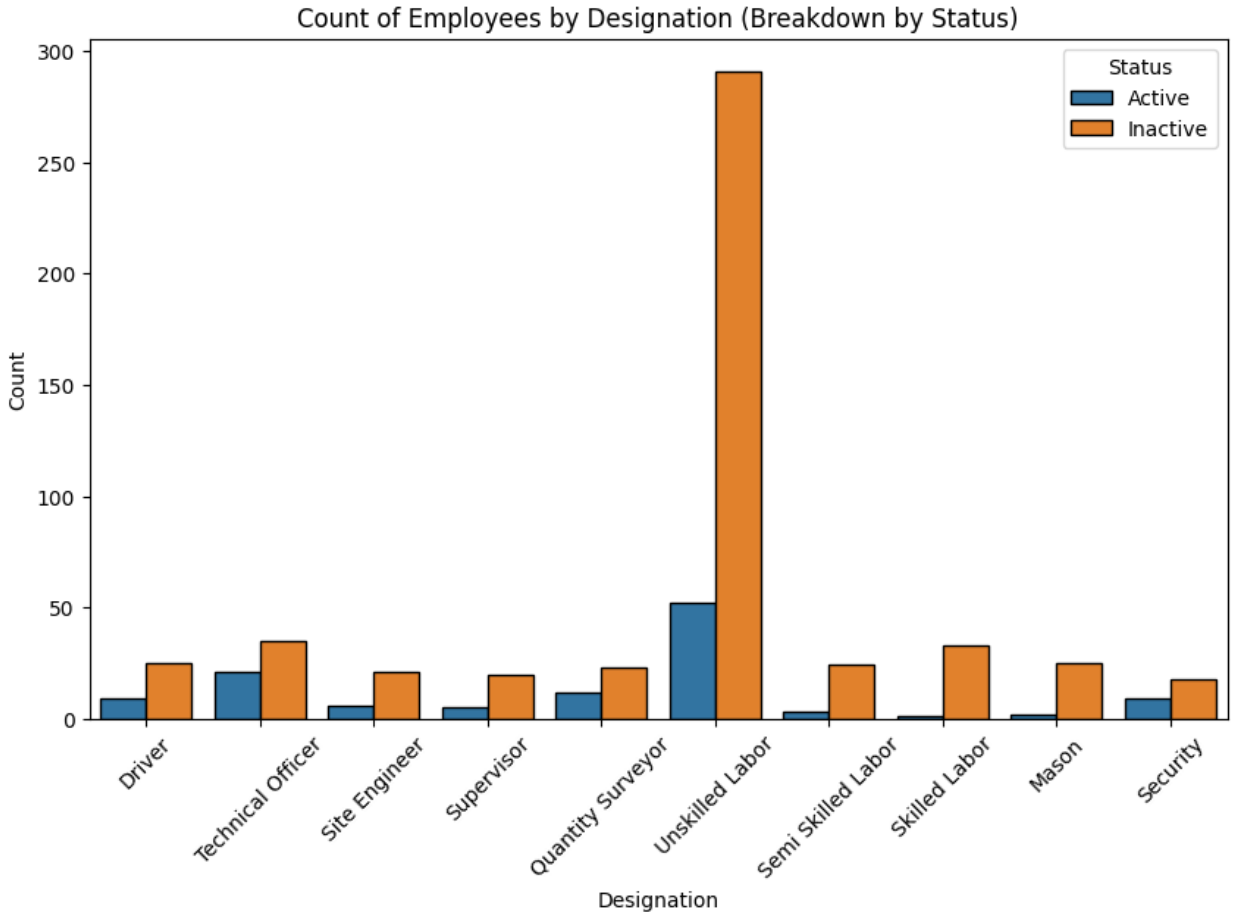
As you can see in the above graph, employee attrition among employees who receive salaries of the lower end is very high. To remedy the situation, the company needs to re-evaluate the market rates for employees of the same caliber in other companies. The reason why many employees leave could be due to the attractive salaries offered by competitors. In addition to that, poor management can also be a reason behind this.

3. **More employees have joined the company on a contract basis in recent years.**



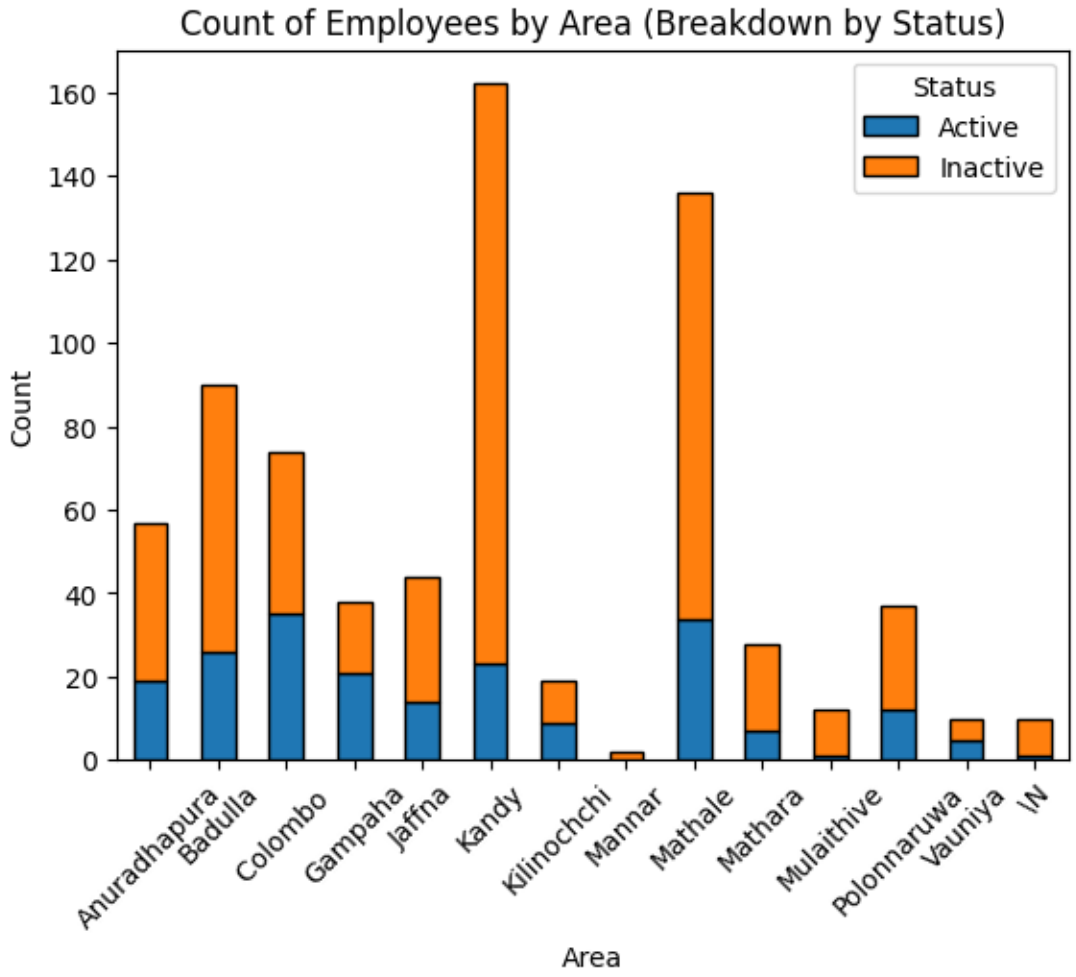
Another reason for higher employee attrition could be related to the type of employment. As evident in the above graph, almost all the employees signed after the year 2016 are on contract basis. On the contrary, most of the permanent employees were signed much earlier. This opens the opportunity for employees to resign after the termination of the contract.

4. The overwhelming majority of employees are unskilled labor.



As observed in the above bar chart, an overwhelming majority of employees fall into the unskilled labor category. And most of the unskilled labor have resigned. This could be resulted by most unskilled laborers being contract basis workers. Apart from unskilled laborers, skilled and semi-skilled labor attrition rates are high too. Implementing targeted strategies to tackle these issues can effectively decrease attrition rates within the company.

5. Geographical influences



As you can see in the above plot, the districts Kandy, Matale and Badulla show the highest rate of employee attrition. All 3 of them are from the central hill area of the country. Several reasons can cause this phenomenon. One of which can be the fact that construction work in hills can be very challenging and taxing on the body that it requires longer recovery times. Apart from a select few areas, most of them showcase a similar trend of higher employee attrition rate.

Final Remarks

Our analysis focused on understanding the factors and reasons behind employee turnover at Marvelous Constructions. Based on our findings, we anticipate that employees in the Labor category, individuals signed on a contract basis, individuals engaged in unskilled labor, employees earning salaries below the average, and individuals from regions such as Kandy, Matale, and Badulla are more prone to leaving the company compared to other employees. By implementing careful planning and effective management strategies, it is possible to gain control over this situation and potentially enhance it.

Google colab notebook of the workings:

https://colab.research.google.com/drive/1F2WY1UoVFVAZaXbiD9YzUWCKYu3-hQw7?usp=sharing

Akmal Ali Jasmin
200238N
2023/07/16