Employee Attrition in Marvelous Construction

Final Report

Group 18

Group Members : *210273B, 210685N, 210333K, 210401T, 210352R*

Please note that we have mistakenly uploaded the preprocessed dataset as the 1st submission and not the code. We apologize for the mistake and uploaded the code with this submission, the link for the code:
https://github.com/RavinduWeerakoon/Employee-Attrition-Analysis/blob/main/18.py

1. Problem Overview

Marvelous Construction is a major construction firm with 35 construction sites in different areas in Sri Lanka. The Human Resources department of Marvelous Construction has recently noticed many employees resigning. Since employee attrition is an alarming situation, We have decided to analyze these unusual employee resignations. We were provided a dataset containing employee details, attendance, leaves, and salary extracted from the ERP of Marvelous Construction. Our task was to analyze the given dataset and derive valuable insights that would be useful for the CEO of Marvelous Construction to make strategic decisions to improve employee retention.

2. Dataset description

- Data Source
  The dataset used in this analysis was obtained from the Human Resources department of Marvelous Construction, a major construction firm in Sri Lanka.

- Data Collection Method
  The data was collected internally through the company's Enterprise Resource Planning (ERP) system, which tracks employee details, attendance, leaves, and salary.

- Data Format
  The dataset is provided in a structured format as CSV (Comma-Separated Values) files.

- Data Size
  The dataset consists of four CSV files:
  - Employee Dataset: 631 records, 17 attributes
  - Leaves Dataset: 237 records, 6 attributes

    - Salary Dataset: 2632 records, 4 attributes
    - Attendance Dataset: 60354 records, 10 attributes

- Data Quality
  The dataset exhibits various quality issues, including missing values, redundant attributes, unknown attributes, categorical variables, outliers, and more. These issues were addressed through data cleaning and preprocessing techniques.

3. Data pre-processing

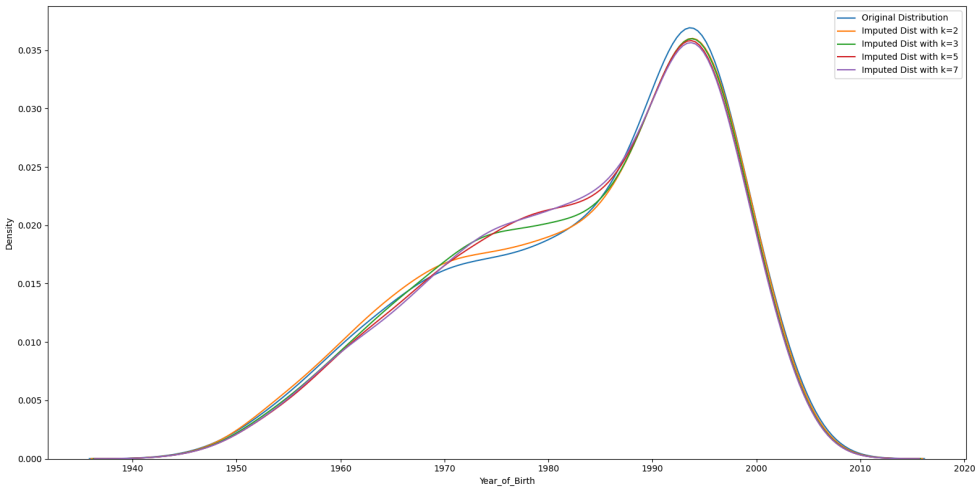
- Data Integration
  Merged necessary features of four CSV files (employee, attendance, leaves, salary) into a single dataset named "marvelous", facilitating comprehensive analysis.

- Remove Redundant Attributes
  Identified and removed redundant attributes such as "Religion" and "Designation" to simplify the dataset and improve model efficiency.
  The title attribute is removed since it can be generated with combining the “Marital Status” and the “Gender” attributes.Introduce the new feature ‘ages_in_the_company’, taking the difference between the existing columns ‘Date_Resigned’ and ‘Date_Joined’ and removing them.

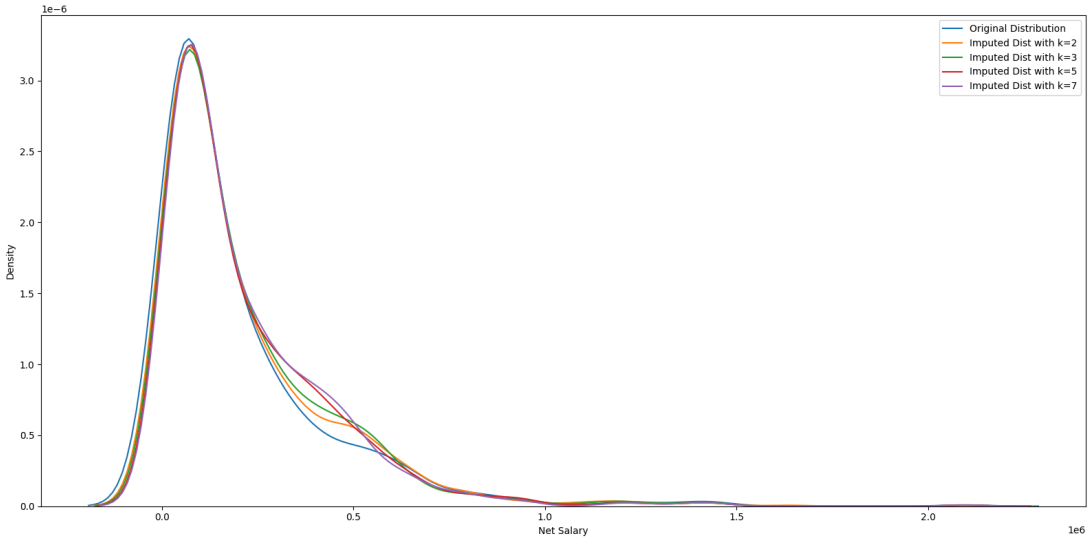
- Remove Irrelevant Attributes
  Removed irrelevant attributes including "Employee_No", "Employee_Code", "Name", "Title", and others to enhance performance and reduce overfitting.

- Categorical Columns Encoding
  Employed binary, ordinal, and one-hot encoding techniques to encode categorical columns such as Marital Status, Employment Type, and Gender into numerical values. Apply ordinal encoding for Gender, Status, and Employment_Category columns and one hot encoding for Marital Status, Employment Type, and Religion.

- Impute Missing Values
  Initially replaced missing values in the "Year_of_Birth" column with zero and converted it to numerical format to maintain data integrity. Then we used the KNN imputer to impute the year of birth column and we used multiple k values and checked the distribution of the data after being imputed and we have selected the k which gives a slighter deviation to the original distribution.



Initially replaced missing values in the ‘Net Salary’ column with zero and applied the KNN imputer to impute missing values we used multiple k values and checked the distribution of the data after being imputed and we selected the k which gives slighter deviation to the original distribution.

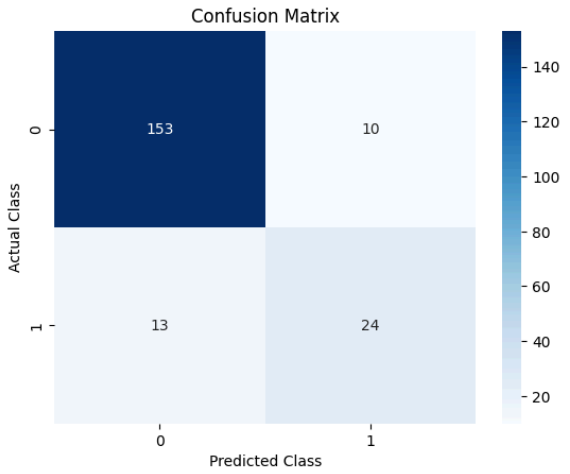


- Data transformation
  Applied Min-Max Scalar to ‘Year_of_Birth’, ‘ages_in_the_company’, and ‘Net Salary’ columns which were reshaped within the range [-1,1].

Predictive analysis for validating the preprocessing quality

To assess the quality of our data preprocessing steps, we employed an XGBoost classification model. The model achieved an accuracy of 0.89, indicating a strong ability to learn from the prepared data.

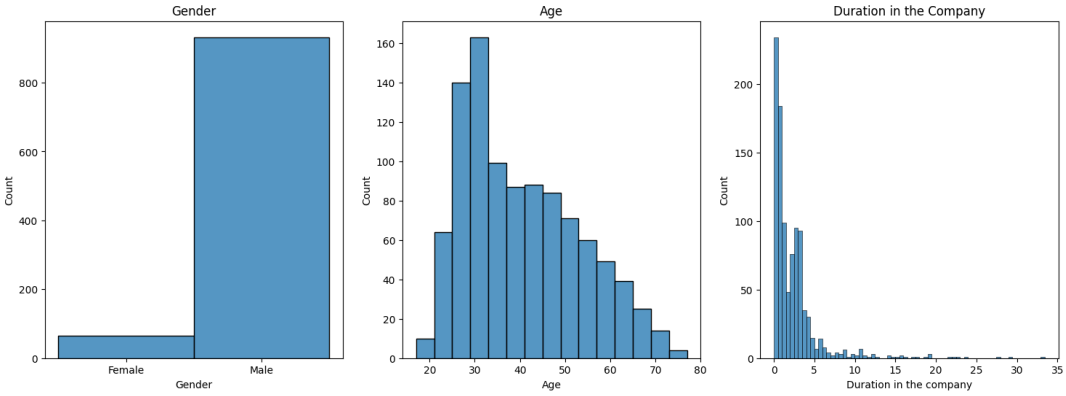
We further evaluated the model's performance using additional metrics beyond accuracy. This included metrics like the F1-score, which provides a balanced view of precision (correct positive predictions) and recall (capturing all true positives). The scikit-learns classification_report offered a comprehensive breakdown of these metrics for each class, allowing us to identify any potential class imbalances.



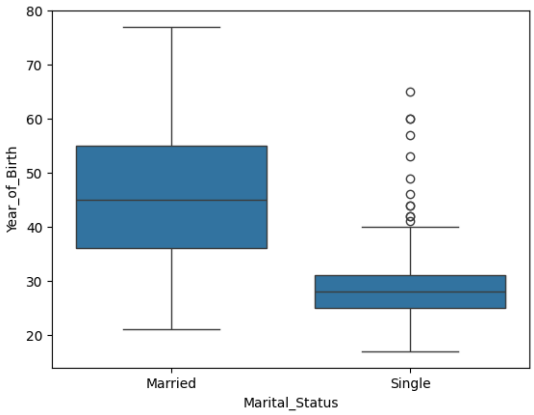
From these promising results, we can say that the data cleaning and preprocessing step has gone well and we have created a clean atmosphere to do our analysis

4. Insights from data analysis

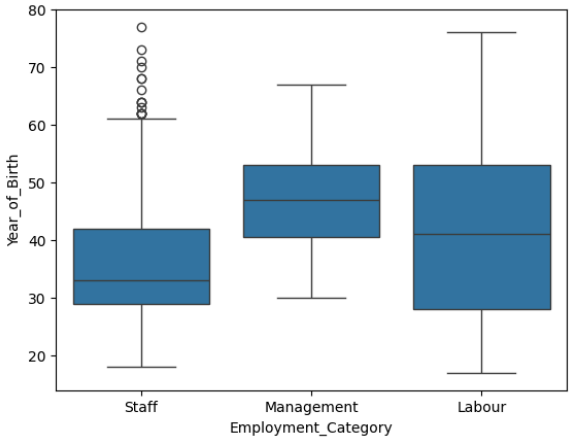
- We can see that the workforce of the company is relatively young and shows a skewed distribution of age. (Consider that age is scaled between 0 and 1)



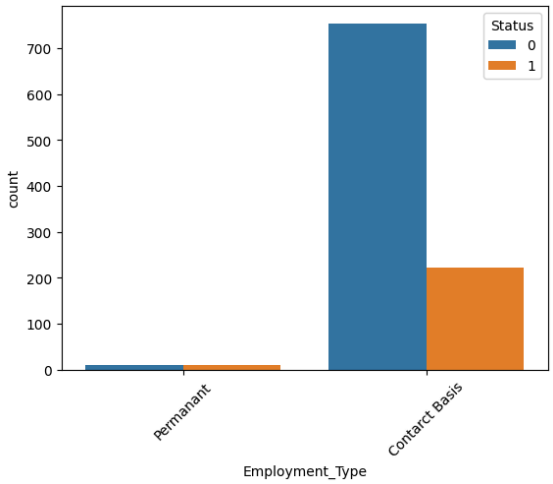
- Most of married people have higher ages compared with unmarried



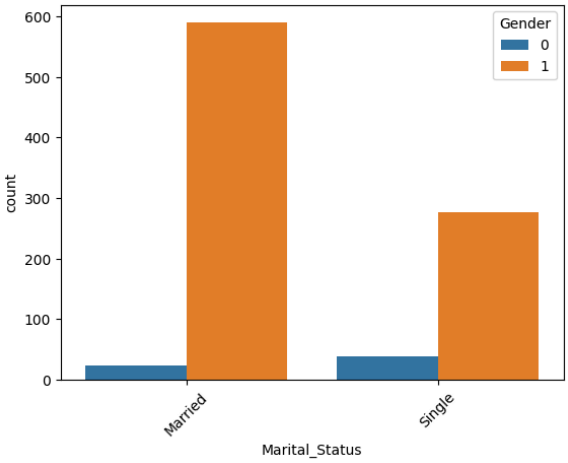
- Management people are of higher ages when compared with the others.



- Most of the inactive people are contract-based workers. And this is obvious since the entire workforce of the company is likely to be contract based and to check the dependency with the status we have to think about the proportions and do further analysis.

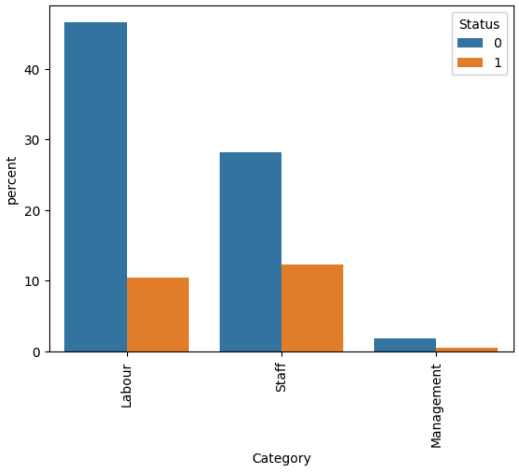


- It shows that higher proportion of males are married when compared with the female employees of the company and the married proportion is large when compared with the unmarried proportion



5. Results of Hypothesis Testing

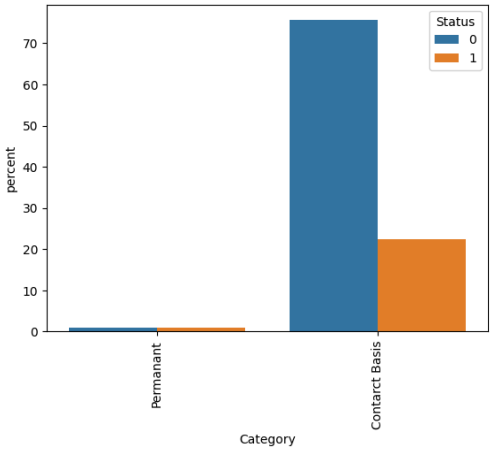
- Hypothesis 1: Employment Category does not have an impact on the employment status



Chi-square statistic: 19.420488818889876
p-value: 6.065888667228183e-05
Degrees of freedom: 2
Reject the null hypothesis at 5% level of significance.

Employment Category has a statistically significant impact on the employment status

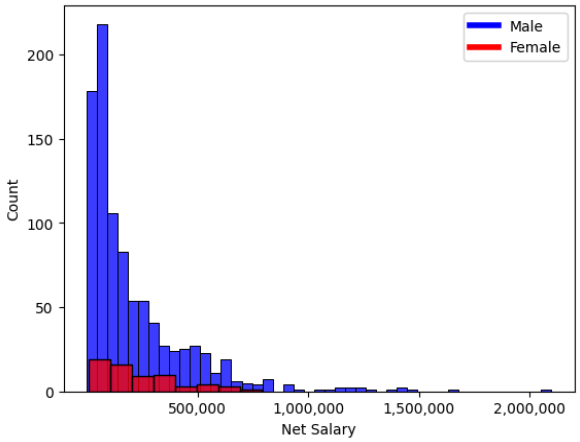
- Hypothesis 2: Employment Type does not have an impact on the employment status



Chi-square statistic: 6.635632103042585
p-value: 0.00999587204002437
Degrees of freedom: 1
Reject the null hypothesis at 5% level of significance.

Conclusion:
Employment Type does has a statistically significant impact on the employment status

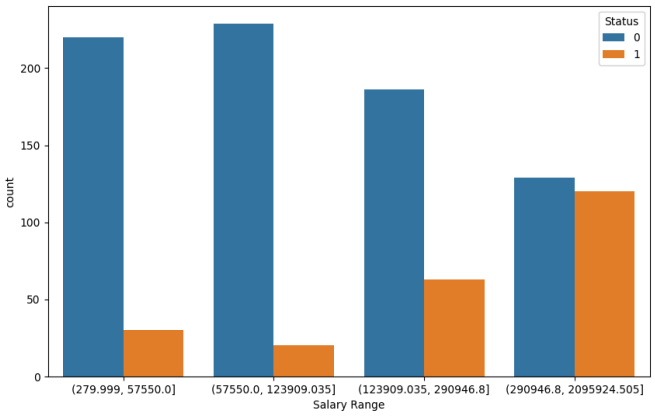
- Hypothesis 3: Gender affects the salary of employees



F-statistic: 0.8736340125889004
P-value: 0.35017799611564016
Fail to reject the null hypothesis at 5% significance level.

Conclusion: There is no significant difference in salary between genders

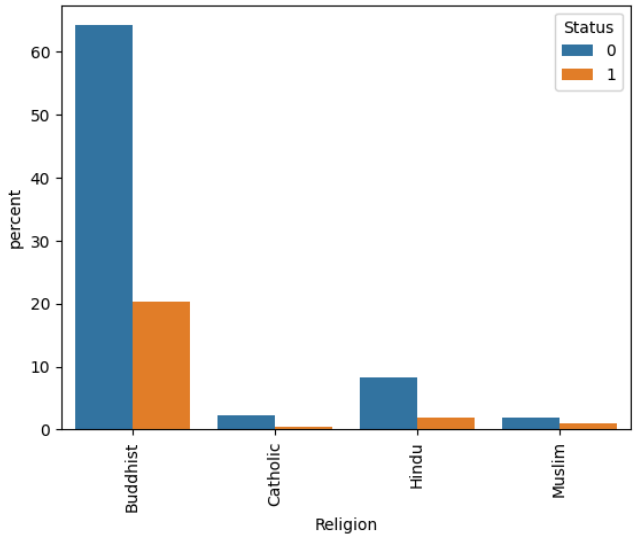
- Hypothesis 4: Net Salary does not significantly affect the employment status



Chi-squared statistic: 136.94707535171807
p-value: 1.7205769695162654e-29
Reject the null hypothesis at 5% significance level

Conclusion:There is a statistically significant relationship between salary and employment status.

- Hypothesis 5: Religion does not significantly affect the employment status



Chi-square statistic: 4.0461987006202955
p-value: 0.256518523656958
Degrees of freedom: 3
Fail to reject the null hypothesis at 5% level of significance.

Conclusion: Religion does not have an impact on employment status.

CONCLUSIONS:

Since the CEO of Marvelous Construction is looking to minimize people resigning their jobs, we can make the following recommendations based on the data analysis.

1. Hiring people on a permanent basis - This has shown to drastically decrease the rate of people leaving
2. Increasing the salary of employees - While this seems straightforward, the data also suggests that increasing the salary will lead to better employee retention
3. Look at ways to retain employees of the “Labour” category. Most of the employees leaving the job happen to be from the “Labour” category. Thus providing them with more incentive to remain at the company could drastically increase the overall employee retention rate.