**Consulting Report**

**Table of Contents**

[1. Introduction 2](#_Toc129952452)

[2. Problem Statement 2](#_Toc129952453)

[3. Problem Analysis (Strategy for the Problem Statement analysis) 3](#_Toc129952454)

[4. Sources of Data 3](#_Toc129952455)

[5. Summary of Data Mining 4](#_Toc129952456)

[6. Proposed Solution for Customers (Describe your Analysis in Detail) 4](#_Toc129952457)

[Statistics of Numerical Variable 5](#_Toc129952458)

[Tools Used 5](#_Toc129952459)

## Introduction

**Your client for this project is the HR Department at a software company.**

* They want to try a new initiative to retain employees.
* The idea is to use data to predict whether an employee is likely to leave.
* Once these employees are identified, HR can be more proactive in reaching out to them before it's too late.
* They only want to deal with the data that is related to permanent employees.

**Current Scenario:**

Once an employee leaves, he or she is taken an interview with the name “exit interview” and shares reasons for leaving. The HR Department then tries and learns insights from the interview and makes changes accordingly.

## Problem Statement

**The current process suffers from the following problems:**

* This approach is that it's too haphazard. The quality of insight gained from an interview depends heavily on the skill of the interviewer.
* The second problem is these insights can't be aggregated and interlaced across all employees who have left.
* The third is that it is too late by the time the proposed policy changes take effect.

The company has hired you as data science consultants. They want to automate the process of assigning the drug according to their health report.

**Your Role**

* You are given datasets of past employees and their status (still employed or already left).
* Your task is to build a classification model using the datasets.
* Because there was no machine learning model for this problem in the company, you don’t have quantifiable win condition. You need to build the best possible model.

## Problem Analysis (Strategy for the Problem Statement analysis)

* 1. Exploratory Data Analytics (EDA) is done using Pandas, NumPy, Seaborn, Matplotlib, Pandas Profiling, MySQL connector
  2. Univariate, Bi-Variate and Multivariate analysis is performed

## Sources of Data

**department\_data**

This dataset contains information about each department. The schema of the dataset is as follows:

1. dept\_id – Unique Department Code
2. dept\_name – Name of the Department
3. dept\_head – Name of the Head of the Department

**employee\_details\_data**

This dataset consists of Employee ID, their Age, Gender and Marital Status. The schema of this dataset is as follows:

* employee\_id – Unique ID Number for each employee
* age – Age of the employee
* gender – Gender of the employee
* marital\_status – Marital Status of the employee

**employee\_data**

This dataset consists of each employee’s Administrative Information, Workload Information, Mutual Evaluation Information and Status.

**Target variable**

* status – Current employment status (Employed / Left)
* Administrative information\*
* department – Department to which the employees belong(ed) to
* salary – Salary level with respect to rest of their department
* tenure – Number of years at the company
* recently\_promoted – Was the employee promoted in the last 3 years?
* employee\_id – Unique ID Number for each employee

**Workload information**

* n\_projects – Number of projects employee has worked on
* avg\_monthly\_hrs – Average number of hours worked per month

**Mutual evaluation information**

* satisfaction – Score for employee’s satisfaction with the company (higher is better)
* last\_evaluation – Score for most recent evaluation of employee (higher is better)
* filed\_complaint – Has the employee filed a formal complaint in the last 3 year

## Summary of Data Mining

* 1. There are 29 duplicate records- delete duplicate records
  2. Followig Columns have null values
     1. Dept,gender, martial status-Replace with Mode
     2. tenure, Age- Replace with Median as data is skewed
     3. last\_evaluation,satisfaction- Replace with KNN Imputer
     4. filed\_complaint,recently\_promoted- Replace with 0 assuming no complaint hence null

## Proposed Solution for Customers (Describe your Analysis in Detail)

|  |  |
| --- | --- |
| **Observation** | **Action** |
| 1. More exit from Sales Dept followed by Engineering and then support 2. Unmarried segment has more exit movement | * 1. These dept needs to take appropriate action on employee engagement |
| 1. More exit from unmarried population, whose monthly working hour is 150 and are assigned 2 projects and whose tenure is 3 years and age below 25 | * 1. They are underloaded, not getting enough work, so dept need to work on the proper work allocation |
| 1. People not promoted recently are quitting more than promoted ones 2. **People not promoted recently are having evaluation score of .4 to 1** 3. Low salary grade is quitting more than others 4. People who had filed complaint are quitting more | * 1. See opportunity to standardize salary across org   2. Promotion to be planned for capable employees |

## Statistics of Numerical Variable

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **tenure** | 3.4925 | 1.4533 | 2 | 3 | 3 | 4 | 10 |
| **age** | 32.895 | 9.9735 | 22 | 24 | 29 | 41 | 57 |
| **last\_evaluation** | 0.7163 | 0.1668 | 0.3162 | 0.5715 | 0.7183 | 0.8588 | 1 |
| **satisfaction** | 0.6212 | 0.2492 | 0.0401 | 0.4527 | 0.6483 | 0.8229 | 1 |
| **avg\_monthly\_hrs** | 200 | 50.832 | 49 | 155 | 199 | 245 | 310 |
| **n\_projects** | 3.7781 | 1.2499 | 1 | 3 | 4 | 5 | 7 |

## Tools Used

* + Numpy
  + Panda
  + Matplotlib
  + ScikitLearn