CS3121 Introduction to Data Science

Employee Attrition in Marvelous Construction
Project Description

Deadlines:

Pre-processed dataset - 5th April 2024 at 11.59 p.m.

Final Project Report - 9th May 2024 at 11.59 p.m.

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, the CEO of Marvellous Construction has decided to hire a data scientist to analyze the data within the company to understand the situation. Suppose you are the data scientist hired for the job.

You are provided with a dataset that contains employee details, attendance, leaves, and salary extracted from the ERP of Marvelous Construction. More details about each data file are given below.

File Name	# Records	Dimension	Remarks
employee	631	Employee_No, Employee_Code, Name, Title, Year_of_Birth, Gender, Religion_ID, Marital_Status, Designation_ID, Date_Joined, Date_Resigned, Reporting_emp_1, Reporting_emp_2, Employment_Category, Employment_Type, Religion, Designation	Join date, Resigned date are special fields for training and testing data set for attrition prediction
leaves	237	Employee_No, leave_date, Type(Half day/Full Day), Applied Date, Remarks, apply_type(Annual/Casual)	New employees have fewer leaves, while old employees have a higher no of leaves.
salary	2632	Employee_No, Amount, month, year, <<different factor names>>	Monthly addition/deduction breakdown is included
attendance	60354	id, project_code, date, out_date, employee_no, in_time, out_time, Hourly_Time, Shift_Start, Shift_End	Late minutes = in time - shift start time

Your Tasks

Your goal in this project is to analyze the given dataset and derive valuable insights that would be useful for the CEO of Marvellous Construction to make strategic decisions to improve employee retention. For this purpose, analyze the given dataset with the knowledge you gathered in the class. Before analyzing the data, as usual, pre-process the dataset.

For this project, use Python 3.6 or later. Also, you can use any Python library required for data pre-processing and analysis, e.g., Pandas, Scikit-learn, Numpy, Matplotlib, Seaborn, etc.

1. Data Pre-processing

1. Load the provided CSV file into a Data Frame object named *Marvellous*.
2. Identify the data type of each variable.
3. Impute the "Year_of_Birth" and "Marital_Status" missing values in the "employee.csv" file using an inference-based approach. You may use information in other files if you want.
4. In addition, identify whether there are data quality issues evident in the dataset and handle them.
5. Conduct necessary data transformations.

2. Data Analysis

Conduct descriptive, exploratory, and predictive data analysis on the provided dataset. You may use predictive analytics to validate the pre-processing quality.

3. Hypothesis Testing

Come up with 5 meaningful hypotheses about the data and test them.

Submissions

All submissions are group submissions

Data Cleaning

Create a Python script named `<<Your_Student_group_Id>>.py` to clean the "`employee.csv`" file. Note that we will only evaluate "`employee.csv`" under data cleaning.

Use Python 3.6 for this project, and you can use the following libraries for data cleaning (Please use the specified versions).

- Pandas - 0.24
- Numpy - 1.17
- Scikit-learn - 0.22

The data cleaning section of the project will be evaluated automatically. When your submitted Python script is run, it should read the original data files in the same folder and write the cleaned "`employee.csv`" file to the same folder. The cleaned file should be named `employee_preprocess_<<Your_Student_group_Id>>.csv`. Before submission, run and test your script properly and ensure it is bug-free.

Final report

The final report should describe and justify the pre-processing steps that you have employed. Further, report the five most significant insights you derived from the data analysis. Describe how you arrived at each insight with supporting visualizations and analysis.

The suggested report format is as follows.

- Problem overview
- Dataset description
- Data pre-processing
- Insights from data analysis
    - Insight 1
        - Description of the approach with supporting visualizations and analysis.
    - Insight 2
        - Description of the approach with supporting visualizations and analysis.
    - …
- Results of Hypothesis Testing

The report should be in PDF format and **no longer than eight pages**. It should be named using your student ID (e.g., Group_99.pdf). Upload the report to Moodle before **9th May 2024 at 11.59 p.m.**