Machine Learning Project

Project Instruction

InnovateIQ Corporation is a technology company specializing in medical instrument based in New York. Attrition is a significant concern for InnovateIQ Corporation as it can lead to a loss of valuable talent and increased recruitment costs. As a data scientist at InnovateIQ Corporation, your task is to generate a machine learning model to predict attrition of employees using historical data, including employee demographic information and job-related details, and evaluate the model's performance using appropriate metrics. Dataset is available at `project_dataset.csv`.

Description of the variables:

- `EmployeeID` - unique identifier for each employee
- `age` - Age of the employee
- `BusinessTravel` - represents the frequency of travel required for an employee job
- `MonthlyIncome` - income earned per month by an employee
- `JobSatisfaction` - level of employee satisfaction with their job (higher is better)
- `Bonus` - Additional financial compensation as reward given to employee
- `Department` - division of an organization that an employee works for
- `DistanceFromHome` - variable representing distance of employee's home to office
- `Education` - level of education of employee
- `EducationField` - field of study in which employee was educated
- `EnvSatisfaction` - level of employee satisfaction with working environment
- `Gender` - employee's gender
- `JobRole` - role of the employee within the company
- `MaritalStatus` - employee's marital status
- `PerformanceRating` - employee's performance evaluation rating
- `TrainingTimeLastYear` - number of training hours employee received last year
- `YearsAtCompany` - number of years employee has worked at company
- `YearsSinceLastPromotion` - number of years since employee's last promotion
- `Overtime` - whether or not employee works overtime
- `Attrition` - whether or not employee left the company

A. Coding Activities

As a data scientist, your tasks are to perform the following tasks using Python programming language:

1. Import and check the quality of the dataset. Treat problem if any.
2. Perform Exploratory Data Analysis into the dataset to obtain insights and relevant information as much as you can. Use the findings for features selection.
3. Pre-processing the dataset and make it ready to train the algorithm.
4. Train relevant machine learning algorithm to generate the predictive model to forecast relevant output.
5. Evaluate and discuss the performance of the model.
6. Finally, save your Jupyter Notebook work file as `firstname_project.ipynb` for submission.

B. Analytics Report

Prepare a presentation slide using PowerPoint as your analytics complete report of your findings, insights and model performance. Save your report as `firstname_report.pptx`. Your report should contain the following insights:

1. Introduction
2. Describe your analytics process
3. Discuss findings of univariate analysis
4. Discuss findings of multivariate analysis (response vs input's candidates)
5. Discuss about the performance of your model
6. Add any information that you think necessary for this report

C. Presentation

As part of your evaluation, you will be required to present your key findings and analytics process of your project. You only been given for 10-minutes to present your data science findings to the panel of assessors. Schedule of your presentation will be announced by the organizer. You are requiring preparing a 10-minutes presentation slide (Around 8 to 10 pages) on your key findings for live presentation, as suggested follows:

1. Describe the process of your analytics work.
2. Major findings of your EDA (Univariate of output variable, Features Selection, and etc.)
3. Discuss about your model, including performance and how to improve.

What You have to Submit

You are required to submit the following files:

1. Jupyter Notebook file (`.ipynb`). Name your file as `firstname_project.ipynb`.
2. Project report in PowerPoint file. Name your file as `firstname_report.pptx`.

Student are required and compulsory to submit ALL these files via Google Classroom before or at **11:59PM on 15 August 2024 (Thursday)**.