Artificial Intelligence
Datathon Assignment

Introduction

You are a member of the **Data Science & AI Team of CGI**.

CGI (https://www.cgi.com/us/en-us) was founded in 1976 and is one of the largest IT and business consulting companies. The company knows its industries and customers, acts with a focus on results, and helps customers increase their return on investment in business and IT.

Currently, you are working on the further development of your HR and Sales strategy. Recently, you received a HR-based data file with around 1.250 data points from sales reps. The data file is containing data on general information about sales reps, salaries, positions, roles, main expertise areas, seniority, etc.

The consistency of a data point depends on the quality of data gathering in various HR units, and other factors. Thus, you have to check the consistency and completeness of the data.

The Head of HR is demanding a report about the general effectiveness of the HR strategy regarding sales reps, salary development and distribution, relevant insights for recruiting, potential measures for attrition management, and other relevant factors.

Furthermore, the Head of HR is interested in AI and keen to know how to use data for additional digital services or recommendations for HR professionals.

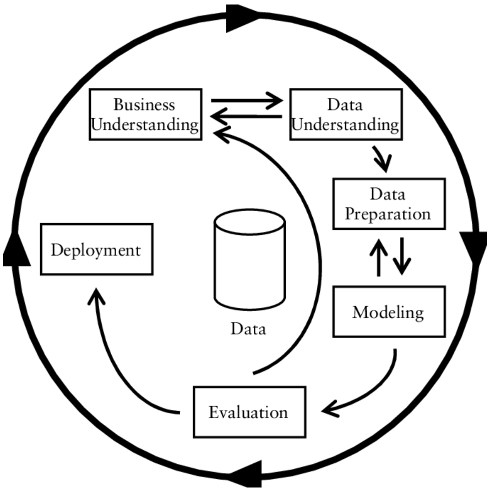
Your board presentations are starting at 15:00.
Prepare your files* for the presentation.

You have a max of 10 minutes time for presentation.
Additional 10 minutes are reserved for questions.

*Mandatory submissions:
1) Presentation file as PDF,
2) All data files and notebooks related to the presentation

Procedure

Your team is working around the CRISP-DM cycle.



Stage 1: Business Understanding

The data set contains multiple variables.

Generally, you are interested in machine learning algorithms based on your data that can predict relevant target metrics for HR and/or Sales based on typical objectives in this functional areas (e.g. salary, acquisition strategies, attrition management, etc.).

CGI would like to optimize their HR strategy with respect to sales employees. Particularly, they would like to focus on keeping high performers engaged and loyal to the company.

*Mandatory submissions:
1) Presentation file as PDF,
2) All data files and notebooks related to the presentation

Stage 2: Data Understanding

The data set is containing the following variables. All variables are self-explaining. Information on scales of variables is not available.

- Index
  = Continous index for identification of employee
- Age
  = Current age of employee
- Gender
  = Gender of employee
- City
  = Main location of business office
- Position
  = Current job role of employee
- Quit
  = Shows whether employee left CGI this year.
- YearsOverall
  = Years of overall working experience
- YearsGer
  = Years of working experience in Germany
- Seniority
  = Shows level of current hierarchy position
- MainTech
  = Main technology expertise area of employee
- OtherTech
  = Other expertise areas of employee
- BasicSalary
  = This year's basic salary of employee
- SalaryBonus
  = This year's salary bonus of employee
- LY_Salary
  = Last year's basic salary of employee
- LY_Bonus
  = Last year's bonus of employee
- VacDays
  = Number of average vacation days of employee at CGI p.a.
- EmplStatus
  = Contract status of employee
- Contract
  = Contract format of employee (limited, unlimited, temporary)
- Language
  = Mother language of employee
- CusSize
  = Size of the CGI clients for which the employee works
- CusArea
  = Industry of customers for which the employee works

*Mandatory submissions:
1) Presentation file as PDF,
2) All data files and notebooks related to the presentation

Stage 3: Data Preparation

The consistency of a data point depends on the quality of data gathering in various HR units, and other factors. Thus, you have to check the consistency and completeness of the data, e.g. missing values, duplicates, obviously wrong values, outliers, introduction of dummy variables, distribution of data, feature scaling.

You should use a data preparation approach powered by Python and Jupyter Notebooks (on your local device or IBM cloud) – optional you may use Data Refinery (on IBM Cloud). Report your findings regarding data preparation and data quality. Present relevant descriptive statistics on relevant variables.

Stage 4: Modeling

Modeling depends on the goals of your analysis. You are requested to use state-of-the-art statistical methods and present your findings based on different types of data visualizations and different types of data analysis.

Make full use of the data and develop different models for different business problems, i.e. different target variables.

Overall, you should develop at least two different models with cluster, regression or classification analysis, e.g. one classification model and one regression model.

Stage 5: Evaluation

- Evaluate the quality of your models based on relevant analysis and metrics.
- If you run classification models, report accuracy, precision, and recall and find arguments for the appropriateness of such metrics.
- Report the fit of your models on the given data. Provide evidence that you prevent overfitting on the data.
- Also, report the different features and weights included in your model. Give an interpretation on the overall quality of the model and the business interpretation of different features and weights.
- If your model performance is poor, reflect on possible reasons and mention areas for improvement.
- Consider ethical issues in your evaluation and presentation. Take into account that you are working with HR data.

*Mandatory submissions:
1) Presentation file as PDF,
2) All data files and notebooks related to the presentation

Stage 6: Deployment

You do not have to create model deployments but the Head of HR wants to get answers for the questions below.

- How do you plan to eventually deploy the model?
- How can the deployed model be integrated into existing IT infrastructures and digital services?
- What are potential limitations and assumptions?

Technical Support

Each team will receive technical support.

Results

Prepare a compelling presentation for the Head of HR about your working cycle, findings, and recommendations, and cover the following aspects:

- Specific problems and questions you want to solve.
- Data exploration and preparation.
- Suitable machine learning approaches.
- Modelling and evaluation.
- Business value of machine learning and limitations.
- Possible integrations with existing and new digital services.
- Information about your teamwork.

Good luck!

*Mandatory submissions:
1) Presentation file as PDF,
2) All data files and notebooks related to the presentation