Proposal for Mid Project

Project Name:

Analytics Job Change of Data Scientists

Dataset Source:

https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists/tasks?taskId=3015

Data Files:

- Data split in three files 1-train, 2-test without result & 3-result only for the test data
- contains 19,158 rows & 14 columns before cleaning

Objective:

To combine the three files in one file and make the EDA , cleaning, handle missing values & outliers and make necessary analysis with suitable visualization then to proceed the preprocessing steps to extract the features then scaling and finally to split the data into train & test to prepare the data to the machine learning model.

About Dataset

Context and Content

A company which is active in Big Data and Data Science wants to hire data scientists among people who successfully pass some courses which conduct by the company. Many people signup for their training. Company wants to know which of these candidates are really wants to work for the company after training or looking for a new employment because it helps to reduce the cost and time as well as the quality of training or planning the courses and categorization of candidates. Information related to demographics, education, experience are in hands from candidates signup and enrollment.

This dataset designed to understand the factors that lead a person to leave current job for HR researches too. By model(s) that uses the current credentials, demographics, experience data you will predict the probability of a candidate to look for a new job or will work for the company, as well as interpreting affected factors on employee decision.

The whole data divided to train and test . Target isn't included in test but the test target values data file is in hands for related tasks. A sample submission correspond to enrollee_id of test set provided too with columns : enrollee _id , target

Note:

- The dataset is imbalanced.
- Most features are categorical (Nominal, Ordinal, Binary), some with high cardinality.
- Missing imputation can be a part of your pipeline as well.

Features

- enrollee_id : Unique ID for candidate.
- city: City code.
- city_ development _index : Developement index of the city (scaled).
- gender: Gender of candidate
- relevent_experience: Relevant experience of candidate
- enrolled_university: Type of University course enrolled if any
- education_level: Education level of candidate
- major_discipline :Education major discipline of candidate
- experience: Candidate total experience in years
- company_size: No of employees in current employer's company
- company_type : Type of current employer
- last_new_job: Difference in years between previous job and current job
- training_hours: training hours completed
- target: 0 – Not looking for job change, 1 – Looking for a job change

---

Conclusion

- Illustrate which features affect candidate decision.
- Prepare the data to the machine learning model which will predict the probability of a candidate will work for the company.