Job Change of Data Scientist

Problem Statement

We will explore a dataset compiled to understand which factors lead a person to leave their current job as a data scientist. Using this data, we'll conduct an exploratory data analysis (EDA) to find patterns, and finally, build a model to predict this turnover and provide insights.

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 are conducted by the company. Many people sign up for their training. Company wants to know which of these candidates 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 is designed to understand the factors that lead a person to leave their current job for HR research 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.

Data Source

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

And consist of three datasets: aug_test.csv, aug_train.csv and sample_submission.csv

Constraints and Scope

The whole data is divided to train and test . Target isn't included in the 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 : Development 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

Criteria for Success (Inspiration)

- Predict the probability of a candidate will work for the company
- Interpret model(s) such a way that illustrate which features affect candidate decision

Approach

Multiple steps will be taken to build a predictive model for this project as well as to analyze the resulting predictions.

1. The aug_test.csv and aug_train.csv will be imported and cleaned via Python. Missing values will be handled appropriately based on specific factors.
2. Categorical variables will all be encoded to numerical variables using Ordinal Encoding, One Hot Encoding, or dummy variable encoding techniques, based on if the features have natural rank ordering or not.
3. The cleaned dataset will be explored visually in order to find interesting trends/correlations in the data. Pairs of columns with correlation coefficient higher than a threshold will be reduced to only one in order to avoid multicollinearity.
4. Multiple models including Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors will be used to train the aug_train.csv dataset after reducing multicollinearity from both the aug_train.csv and aug_test.csv.

Apply the trained models on the aug_test.csv dataset, compare the accuracy of each model and pick up the one with the highest accuracy

Deliverables

The final draft of the project will be presented in the form of a slide deck and formal project report . Jupyter Notebooks will be delivered detailing each step taken and code written for the analysis of the project. A Github repository for the project will be created as well.