Introduction

In this assignment you will have to perform the required exploratory data analysis on the give data set. For each of the questions, you must have to put the question as a text cell and then in the following one or more cell you will answer the questions in terms of code and discussion.

Dataset

hrdata.csv.

Context

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 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.

Although the main objective of this kind of data is the build predictive model to classify a candidate to find the probability whether the candidate will work for them or not, your main goal in this assignment will be to do exploratory data analysis (We have learned various concepts of EDA in the class already in the class). During this process you will learn more about this data, missing values, outliers, correlations, distributions, filling out missing values or removing records, combining features, removing unnecessary features, etc. You will also find other issues such as whether the data set is imbalanced and if yes, how to balance it, etc.

In order to answer the questions, you might need to use your python and EDA knowledge from data camp courses, lecture notes, and for some python syntax and libraries you might need to google or use python documentation and examples.

List of 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

* lastnewjob: Difference in years between previous job and current job
* training_hours: training hours completed
* state: State of the candidate
* city_development_matrix: An indicator of the city development
* target: 0 – Not looking for job change, 1 – Looking for a job change

Tasks:

1. import libraries: pandas, numpy, matplotlib (set %matplotlib inline), matplotlib’s pyplot, seaborn, missingno, scipy’s stats, sklearn (1 pt)

2. import the data to a dataframe and show how many rows and columns does it have (1 pt)

3. call the describe method of dataframe to see some summary statistics of the numerical columns. (1 pt)
I. Explain in words if you find any column’s statistics interesting and good to know (1 pt)

4.Show the top 5 rows and last 5 rows of the data frame (1 pt)

5. List all the numerical columns (1 pt)

6. List all the categorial columns (1 pt)

7. Examine missing values: (2 + 2 + 2 + 5 = 11 pt)
I. Show a list with column wise count of missing values and display the list in count wise descending order
II. Show a list with column wise percentage of missing values and display the list in percentage wise descending order
III. Display a bar plot to visualize only the columns with missing values and their count. The plot should display from less missing value columns in the left and then more missing value columns to the right side of the plot
IV.Use missingno’s bar plot, matrix plot with 200 sample, and heatmap.
1. Interpret any interesting information you found in the heatmap and any one plot

8.Understanding Categorical attributes (this part may require you to make 20+ plots) [26 pts]
I. For each categorical attribute perform the following:
II. Use seaborn bar plot for the categorical feature to see different values and count
III. Use seaborn countplot for the categorical feature against the values of the target
IV. Interpret any interesting information and any information that might help you to make any decision on combining, removing, or adding features based on that, or any resampling maybe needed.

9.Understanding Numerical attributes (16 pts)
I. For each numerical features, perform the following:
II. Plot their distributions using histogram (removed the group by word)
III. Plot the distribution using seaborn distplot
IV. Interpret any interesting information

10. Correlation: (15 pts)
I. For the numerical attributes, use heatmap to show the correlation
II. If you find any interesting short list of columns, create another heatmap with them and show the correlations inside the heaptmap as well
III. Show scatter plots between columns to show the relationships with the target
IV. Interpret and explain any finding and next course of action from there

11. Outliers: (5)
I. Use boxplot or any other strategies to find outliers

12. What are the different values of experience, can you categorize them in to 0, 1, and 2? (5 pts)

13. Summary and discussion: (15 pts)
I. Finally after all the above EDA, summarize your finding, next course of action such as we may need to transform distribution because of right skew etc, need to remove a particular columns for any reasons, remove records for any reasons, need to rebalance data and what are the rebalancing options (if needed), and any other finding.