Context and Content

A company active in Big Data and Data Science wants to hire data scientists among people who successfully pass some courses conducted by the company. Many people signup for the course. The company wants to forecast how many people will decide to work for the company and how many will look for a new job somewhere else once the training completes.

This dataset is designed to understand the factors that lead a person to leave their current job for HR research. By model(s) that uses the existing credentials, demographics, experience data, I will predict the probability of a candidate looking for a new job or continue working for the company and interpret factors that affect their decisions.

The whole data divided into train and test. The target column is excluded in the test and saved in a separate file with the corresponding enrollee\_id.

Note:

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

Features

* enrollee\_id: Unique ID for the candidate
* city: City code
* city\_ development \_index: Development index of the city (scaled)
* gender: Gender of the candidate
* relevent\_experience: Relevant experience of a candidate
* enrolled\_university: Type of University course enrolled if any
* education\_level: Education level of candidate
* major\_discipline: Education major discipline of the 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 – will not work for the company, 1 – will work for the company

Inspiration

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

Hypothesises

1. Candidates from a small company with high training hours will highly likely decide to work for the company.
2. Candidates from a disadvanced city with high education level will highly likely decide to work for the company.
3. Male Candidates with fewer years of experience will less likely decide to change their current job (will not work for the company)

Please refer to the following task for more details:  
<https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists/tasks?taskId=3015>

**Predict the probability of an candidate looking for a new job**

[Möbius](https://www.kaggle.com/arashnic) · 35 Submissions · 2 months to go

keyboard\_arrow\_up68

**Task Details**

This dataset designed to understand the factors that lead a person will work for the company(leaving current job) ,and the goal of this task is building model(s) that uses the current credentials,demographics,experience to predict the probability of a candidate looking for a new job or will work for the company.

The whole data divided to train and test . Also sample submission has been provided correspond to enrollee\_ id of test set ( enrolle\_ id | target)

* Note:
  + The dataset is imbalanced so it might affect your result if you dont handle it
  + Most features are categorical (Nominal, Ordinal, Binary), some with high cardinality so encoding methods and techniques will help to boost models performance
  + Missing imputation strategy might affect the results so it can be a part of your pipeline as well.

**Expected Submission**

Solution Notebook with following mandatory outputs:

* Prediction: for each row in test data the predicted probability of candidate looking for a job, (probability of the class with the greater label).Format is given in sample\_submission.csv:
  + enrollee\_id : Unique ID for enrollee
  + target : probability of an enrollee looking for a job change
* Model Evaluation: roc\_ auc\_ score of (answer,prediction probability of rows in test data)  
  \*answer: test target values(download from : <https://www.kaggle.com/arashnic/job-change-dataset-answer>)
* Model Interpretation: Interpret model(s) such a way that illustrate which features affect candidates decision

**Evaluation**

The evaluation metric is [area under the ROC curve](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html) score.

You have better first evaluate your model(s) with your selected strategy (KFold CV, hold\_ out,…) and then predict test (sample\_submission) and finally compare the result for your own with test target values and find out your model performance on unseen data and your model fitness.

*The pipeline you apply for getting result (EDA, preprocessing, feature engineering and selection , model building and evaluation strategy and so on, will be evaluated by your model performance on unseen data(aug\_test.csv)*