Prediction of Job Change

Group 1

Renping Ge, Yuchen Ma, Zongzhu Li

Outlines

- ➢ Introduction
- ➢ Exploration Data Analysis
- ➢ Data Preprocessing
- ➢ Modeling and Prediction
  - Logistic Regression
  - Random Forests
  - Histogram based Gradient Boosting
- ➢ Modeling Evaluation
- ➢ Features Importance
- ➢ Summary and Discussion

Introduction

- ➢ Exploring the probability of a candidate to look for a new job or will work for the company after training
- ➢ Interpreting affected factors on employee decision

Dataset Description

➢ This dataset, provided by kagle, is collected by the company and designed to understand the factors that lead a person to work for the company(leaving their current job).

➢ 19158 entries and 14 columns including features variables and target variable.

```
In[4]: print(data.info())
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19158 entries, 0 to 19157
Data columns (total 14 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   enrollee_id             19158 non-null  int64
 1   city                    19158 non-null  object
 2   city_development_index  19158 non-null  float64
 3   gender                  14650 non-null  object
 4   relevent_experience     19158 non-null  object
 5   enrolled_university     18772 non-null  object
 6   education_level         18698 non-null  object
 7   major_discipline        16345 non-null  object
 8   experience              19093 non-null  object
 9   company_size            13220 non-null  object
 10  company_type            13018 non-null  object
 11  last_new_job            18735 non-null  object
 12  training_hours          19158 non-null  int64
 13  target                  19158 non-null  float64
dtypes: float64(2), int64(2), object(10)
memory usage: 2.0+ MB
```

Dataset Description - Features and Target

- enrollee_id: Unique ID for enrollee.
- city: City code.
- citydevelopmentindex: Development index of the city (scaled).
- gender: Gender of enrollee.
- relevent_experience: Relevant experience of enrollee.
- enrolled_university: Type of University course enrolled if any.
- education_level: Education level of enrollee.
- major_discipline: Education major discipline of enrollee.
- experience: Enrollee 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.

Exploratory Data Analysis

➢ remove unrelated columns

```
data = data.drop(["enrollee_id","city"],axis=1)
```

➢ change some values to be understood easily

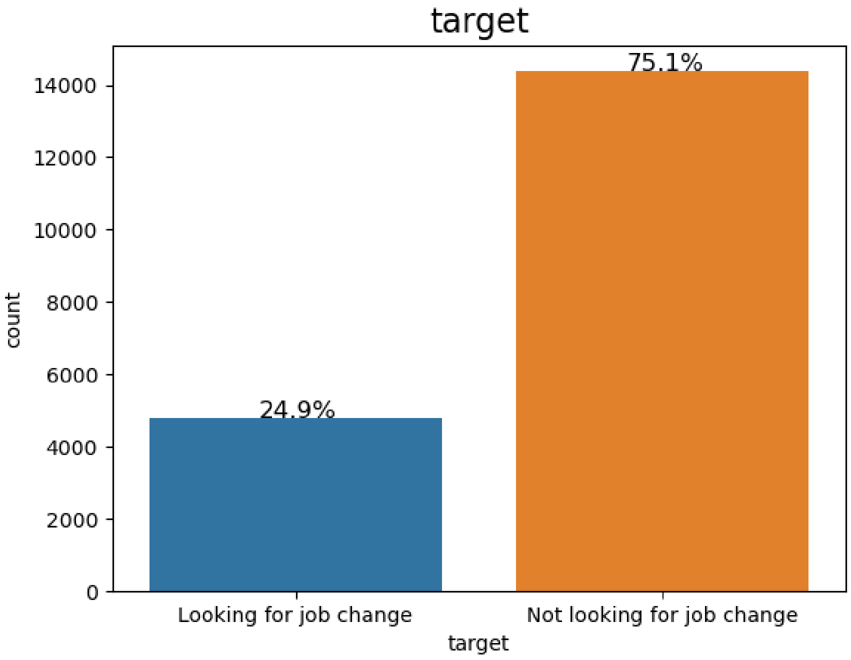
```
data["company_size"].unique()
for i in range(len(data.index)):
    if data['company_size'][i] == '10/49':
        data['company_size'][i] = '10-49'
```

```
data["experience"].unique()
for i in range(len(data.index)):
    if data['experience'][i] == '>20':
        data['experience'][i] = '21'
    elif data['experience'][i] == '<1':
        data['experience'][i] = '0'
```

```
data["last_new_job"].unique()
for i in range(len(data.index)):
    if data['last_new_job'][i] == '>4':
        data['last_new_job'][i] = '5'
    elif data['last_new_job'][i] == 'never':
        data['last_new_job'][i] = '0'
```

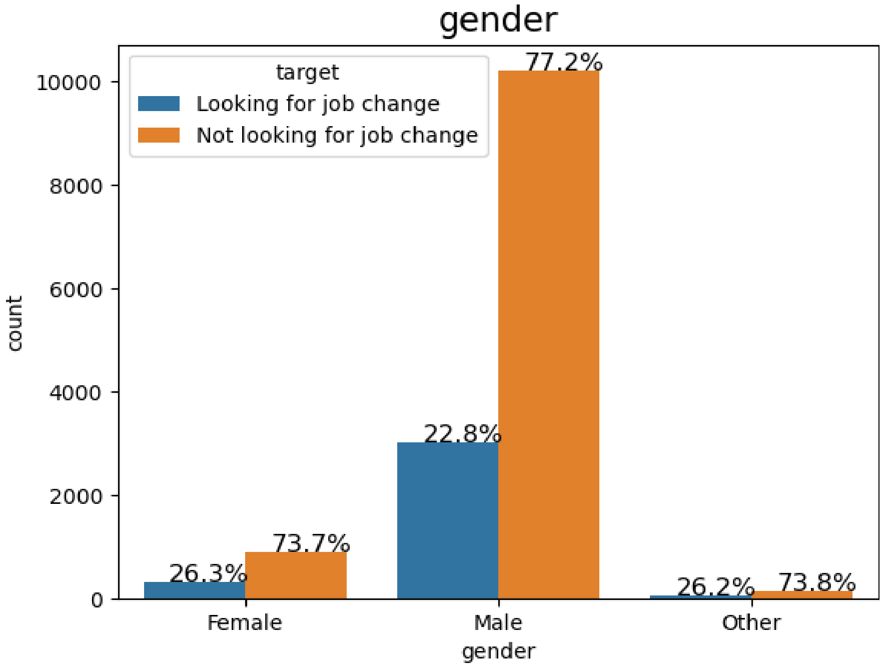
```
retarget = {0.0: 'Not looking for job change',
            1.0: 'Looking for job change'}
data['target'] = data['target'].map(retarget)
```

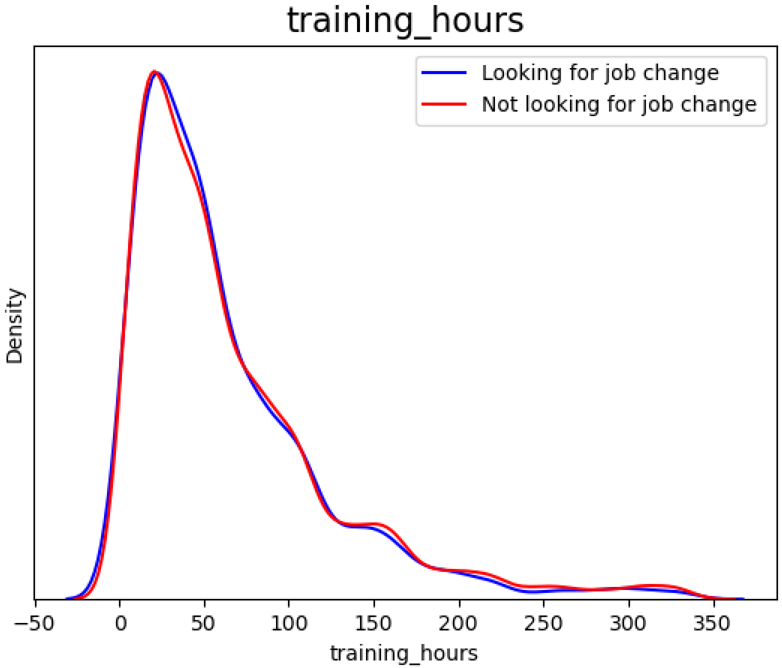
EDA - Counts for target



➢ There are 24.9% enrollee is looking for job change and 75.1% enrollee is not.

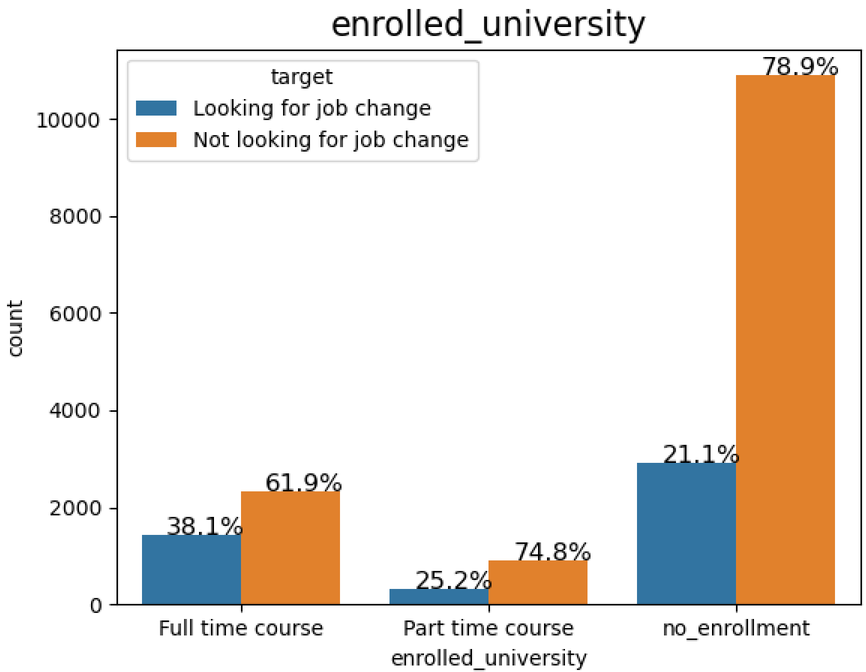
EDA - Distribution of job change by gender and training hours

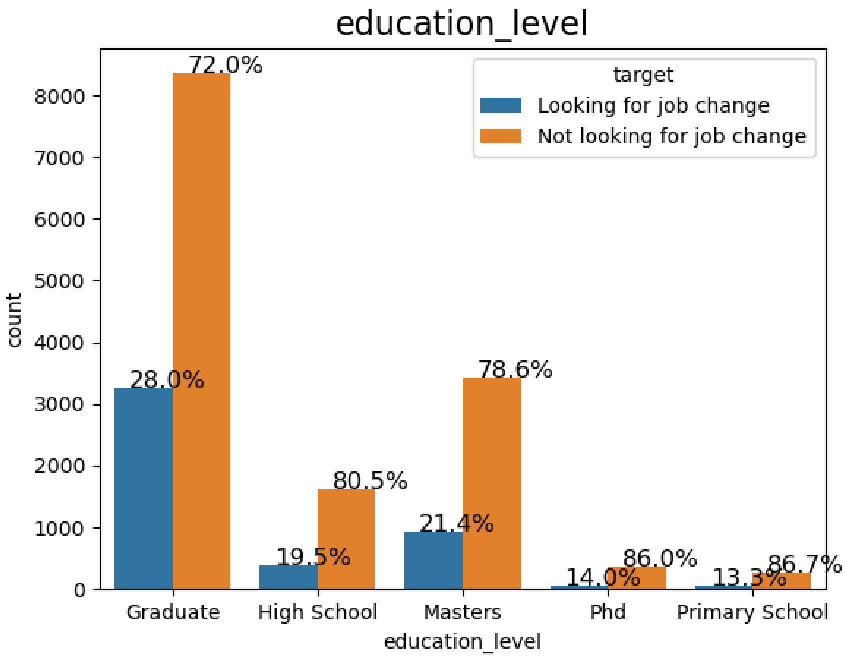




- ➢ People with different gender shows comparable rate of looking for a new job.
- ➢ People with different training hours shows comparable rate of looking for a new job.

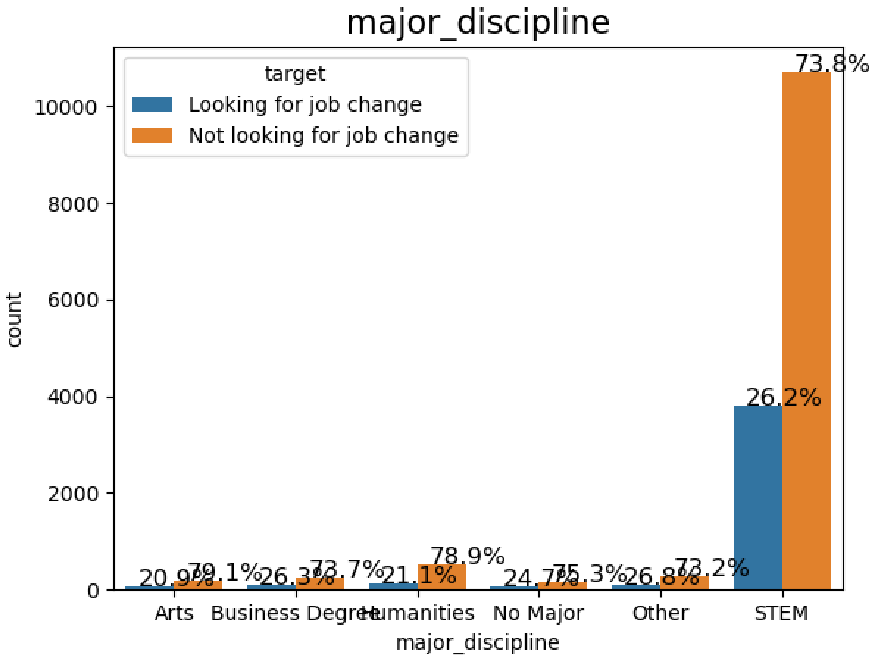
EDA - Distribution of job change by enrolled_university and education_level

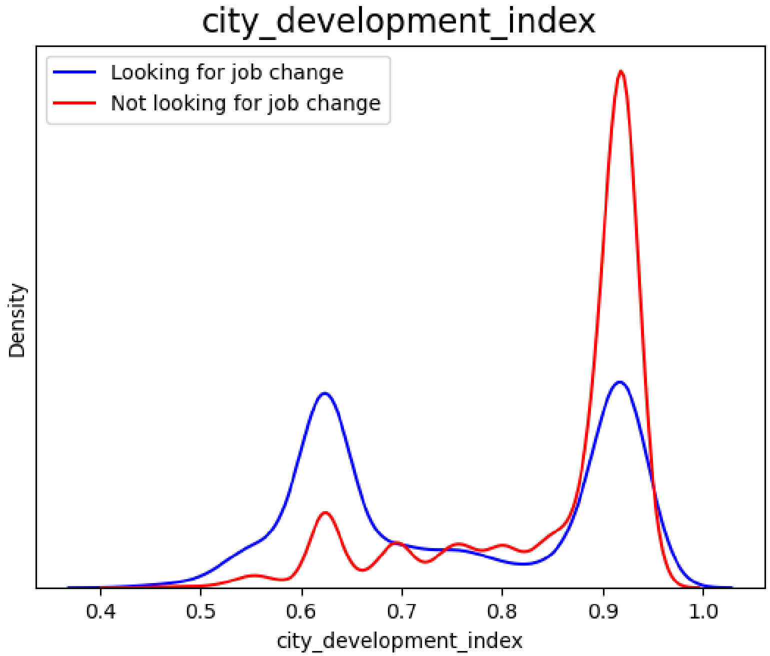




- ➢ People who took the full time course are more likely to look for a new job compared to others.
- ➢ People with graduate education level are more inclined to look for a new job.

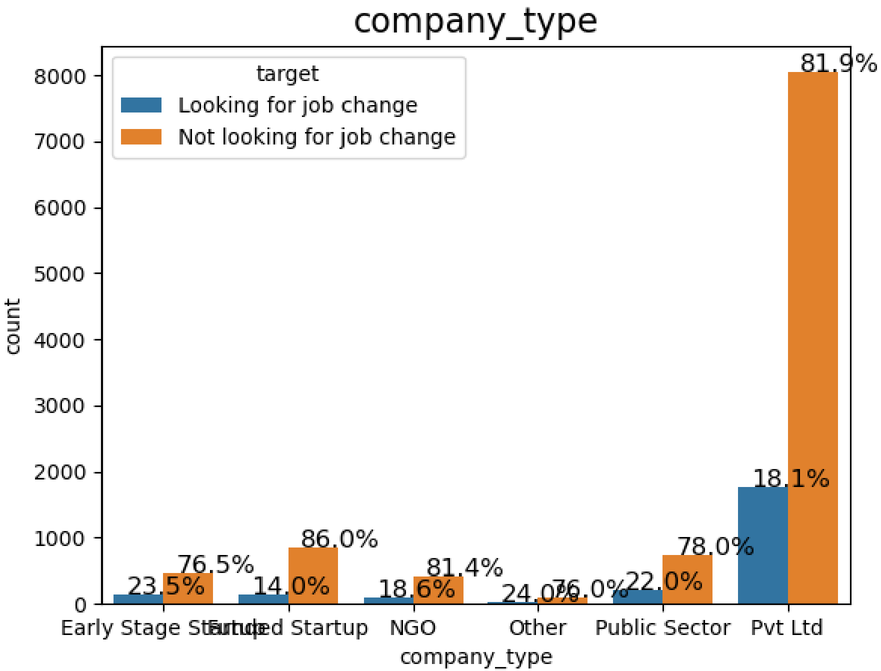
EDA - Distribution of job change by major_discipline and city_development_index

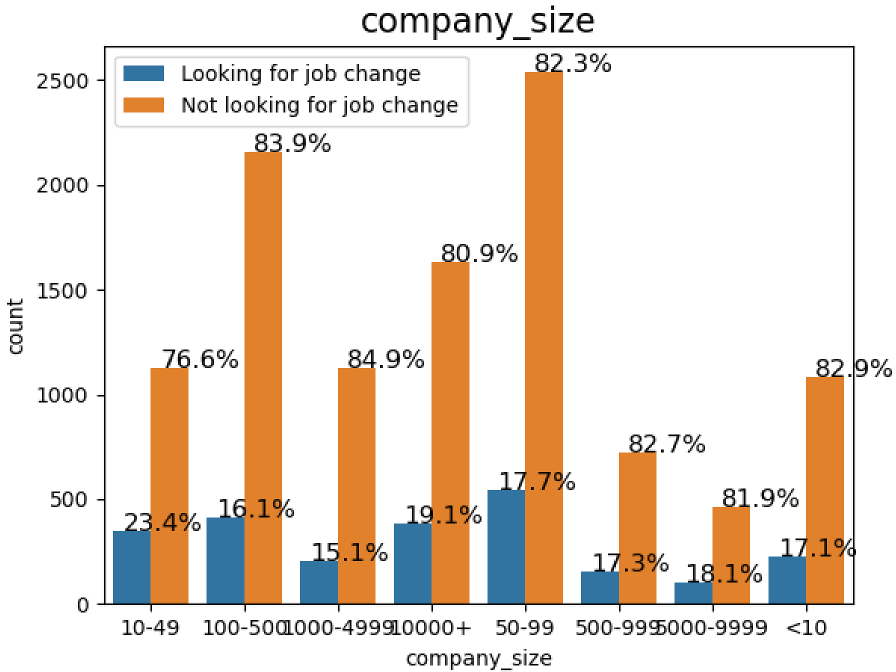




- ➢ People with different major discipline shows comparable rate of looking for a new job.
- ➢ In the cities with lower city_development_index, more people is likely to look for a new job.

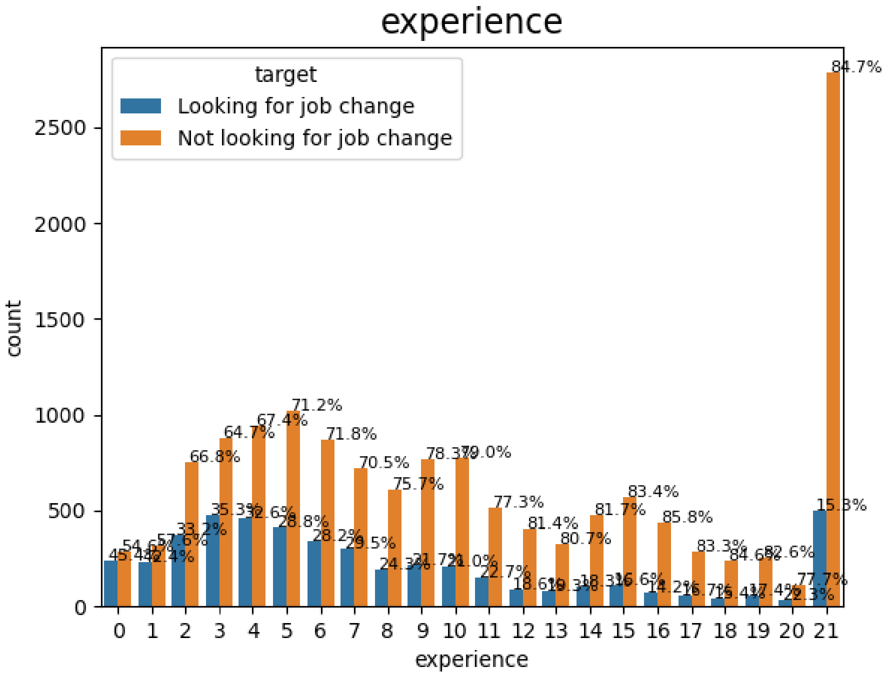
EDA - Distribution of job change by company_type and company_size

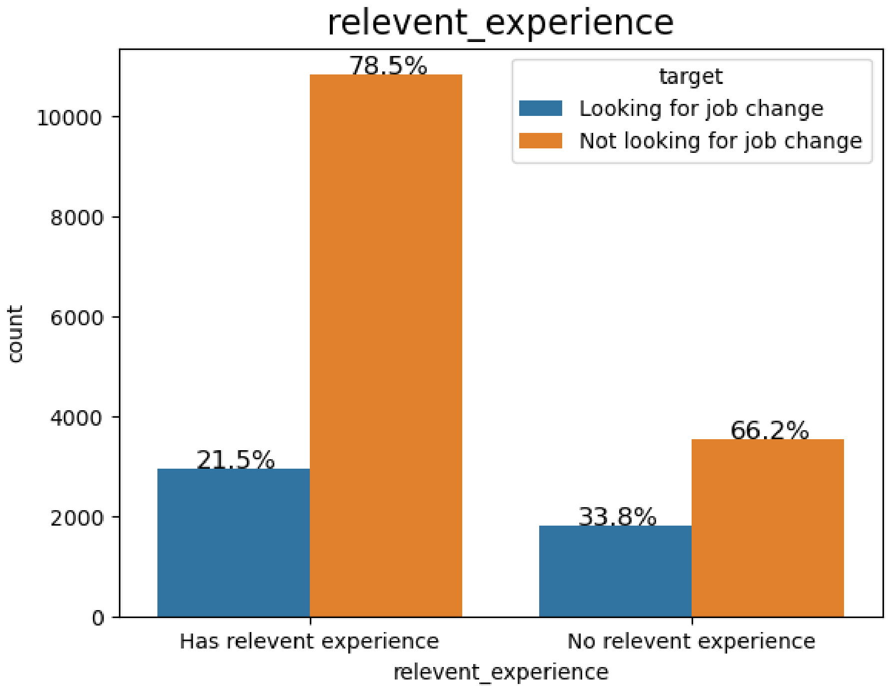




- ➢ People working in the Pvt Ltd, NGO and Founded Startup are less likely to look for a new job.
- ➢ People working in the company with size of 10-49 are more inclined to look for a new job.

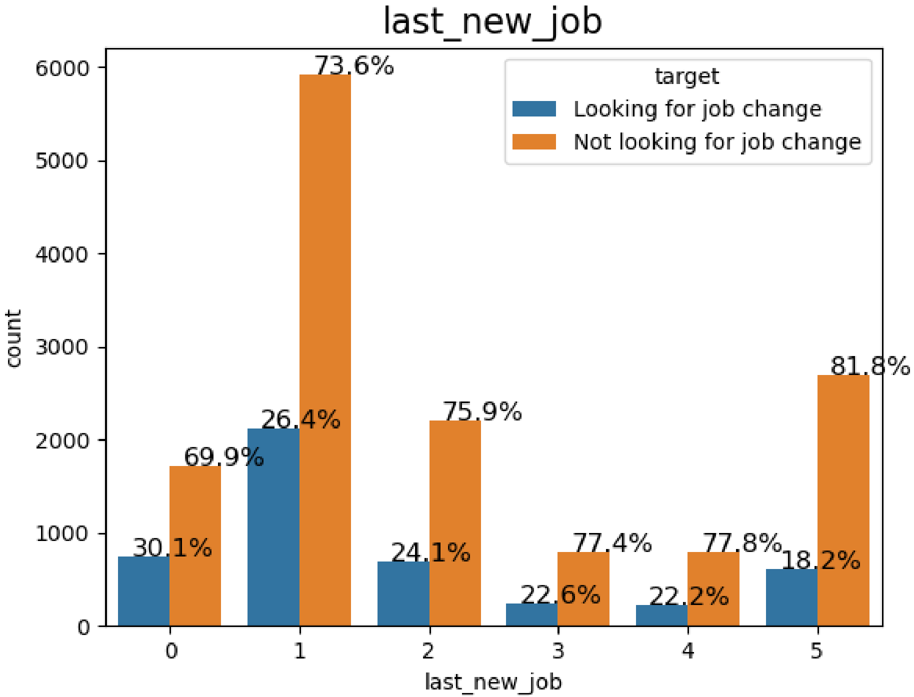
EDA - Distribution of job change by experience_years and relevent_experience





➢ People with less working experiences and with no relevant experience are more likely to look for a new job.

EDA - Distribution of job change by last_new_job

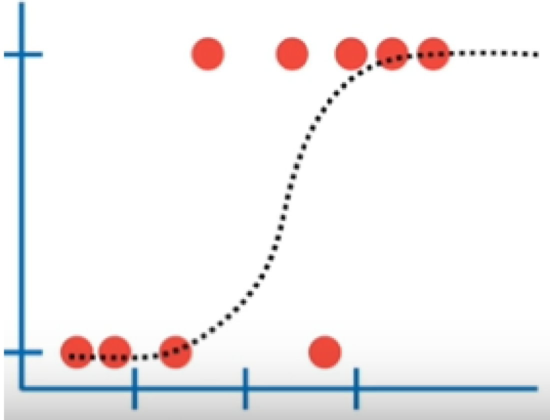


➢ The difference of 1 year and zero year shows a significant higher rate of looking for a new job.

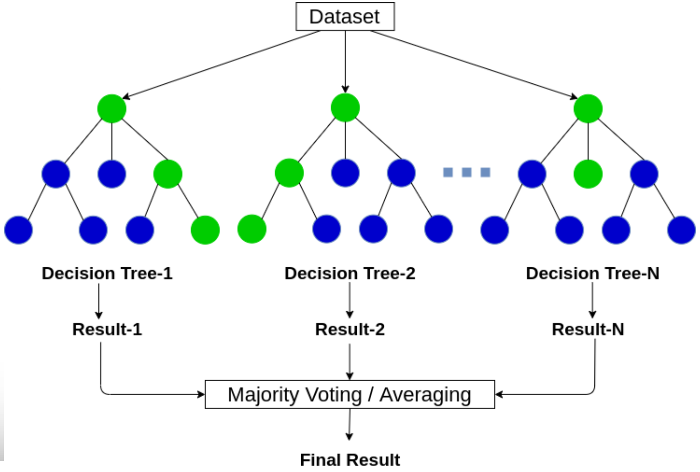
Data Preprocessing

- ➢ specify the predictors and target variable
- ➢ fill na
- ➢ standardization and centralization for numerical variables with StandardScaler()
- ➢ encoding categorical features with OneHotEncoder()
- ➢ label target variable with LabelEncoder()
- ➢ split the dataset into train (70%)and test (30%)

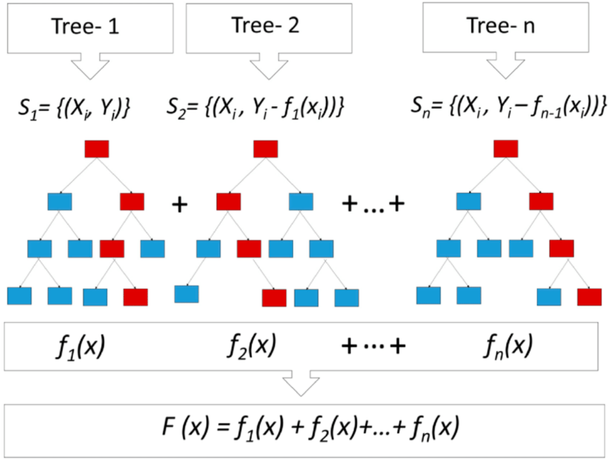
Models



Logistic Regression



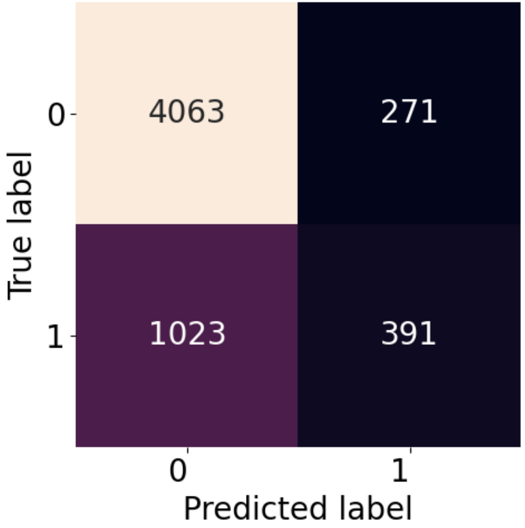
Random Forest



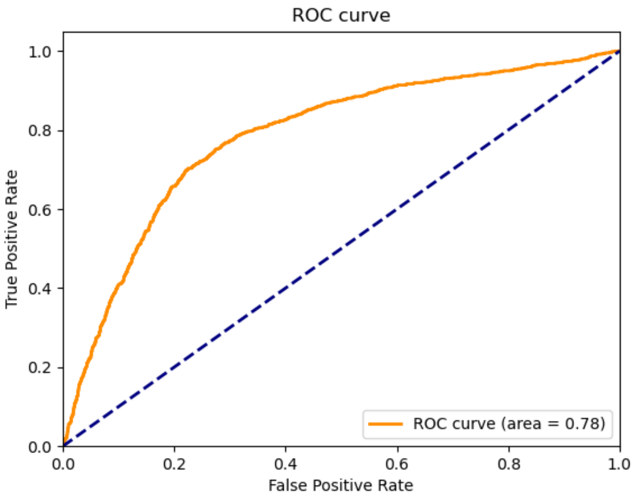
Gradient Boosting

Modeling Evaluation - Logistic Regression

confusion matrix



ROC area under curve



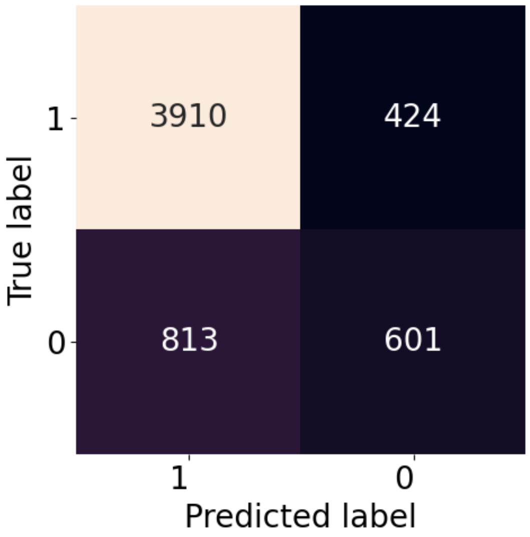
report

	precision	recall	f1-score	support
0	0.80	0.94	0.86	4334
1	0.59	0.28	0.38	1414
accuracy			0.77	5748
macro avg	0.69	0.61	0.62	5748
weighted avg	0.75	0.77	0.74	5748

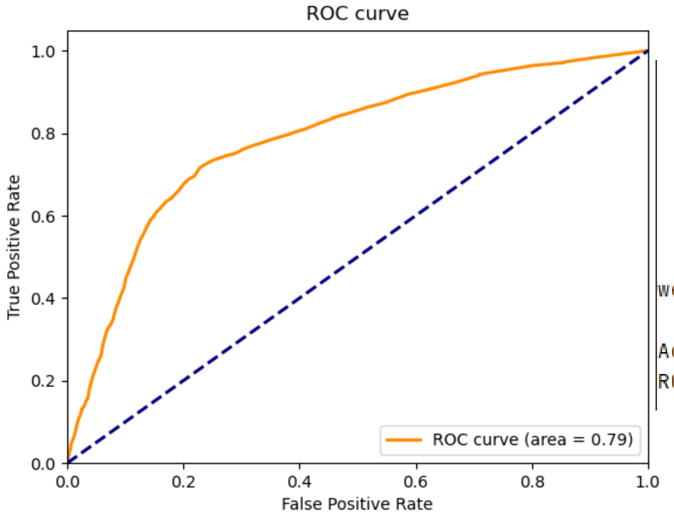
```
Accuracy :  77.48782185107864
ROC_AUC :  78.41463080318185
```

Modeling Evaluation - Random Forest

confusion matrix



ROC area under curve

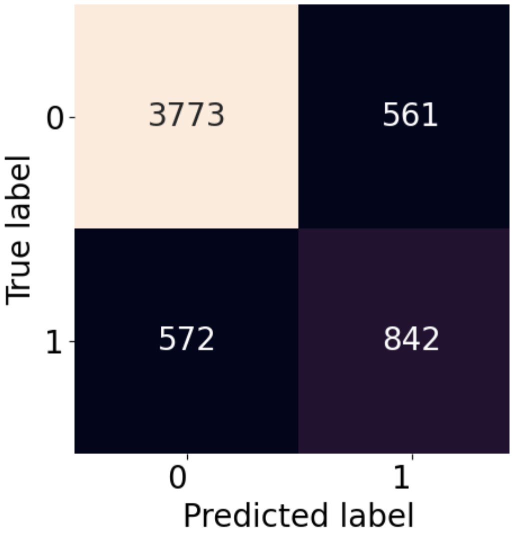


report

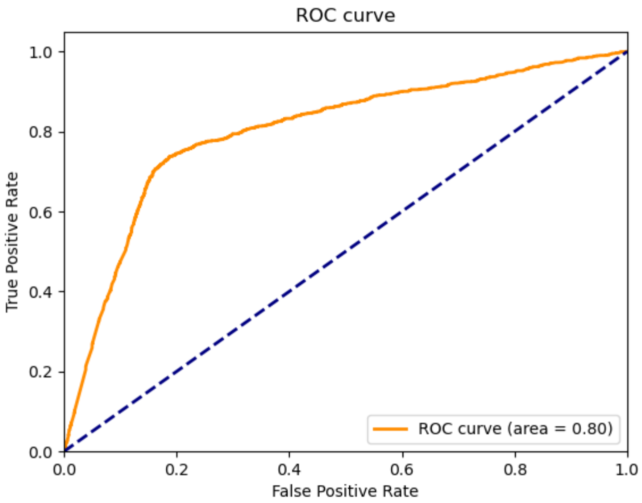
```
              precision    recall  f1-score   support

           0       0.83      0.90      0.86      4334
           1       0.59      0.43      0.49      1414

    accuracy                           0.78      5748
   macro avg       0.71      0.66      0.68      5748
weighted avg       0.77      0.78      0.77      5748

Accuracy: 78.4794711203897
ROC_AUC: 78.6315678340858
```

Modeling Evaluation - Gradient Boosting

confusion matrix



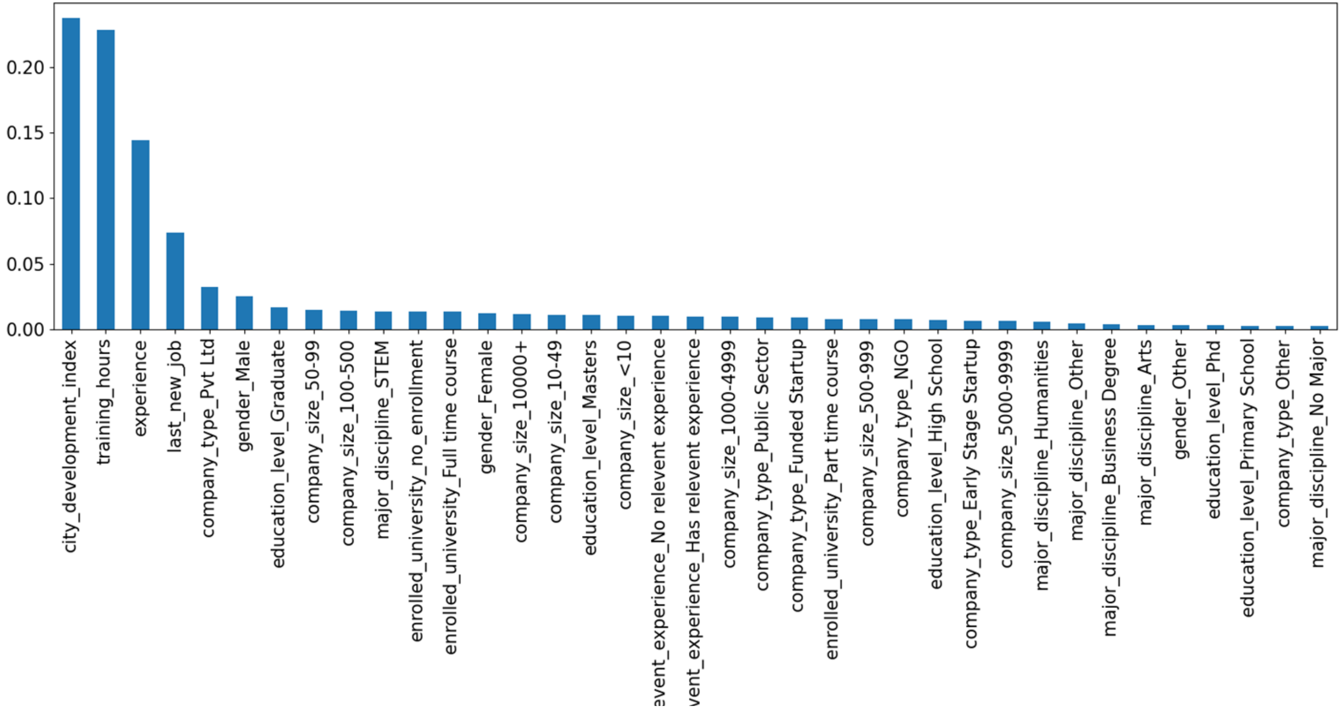
ROC area under curve



report

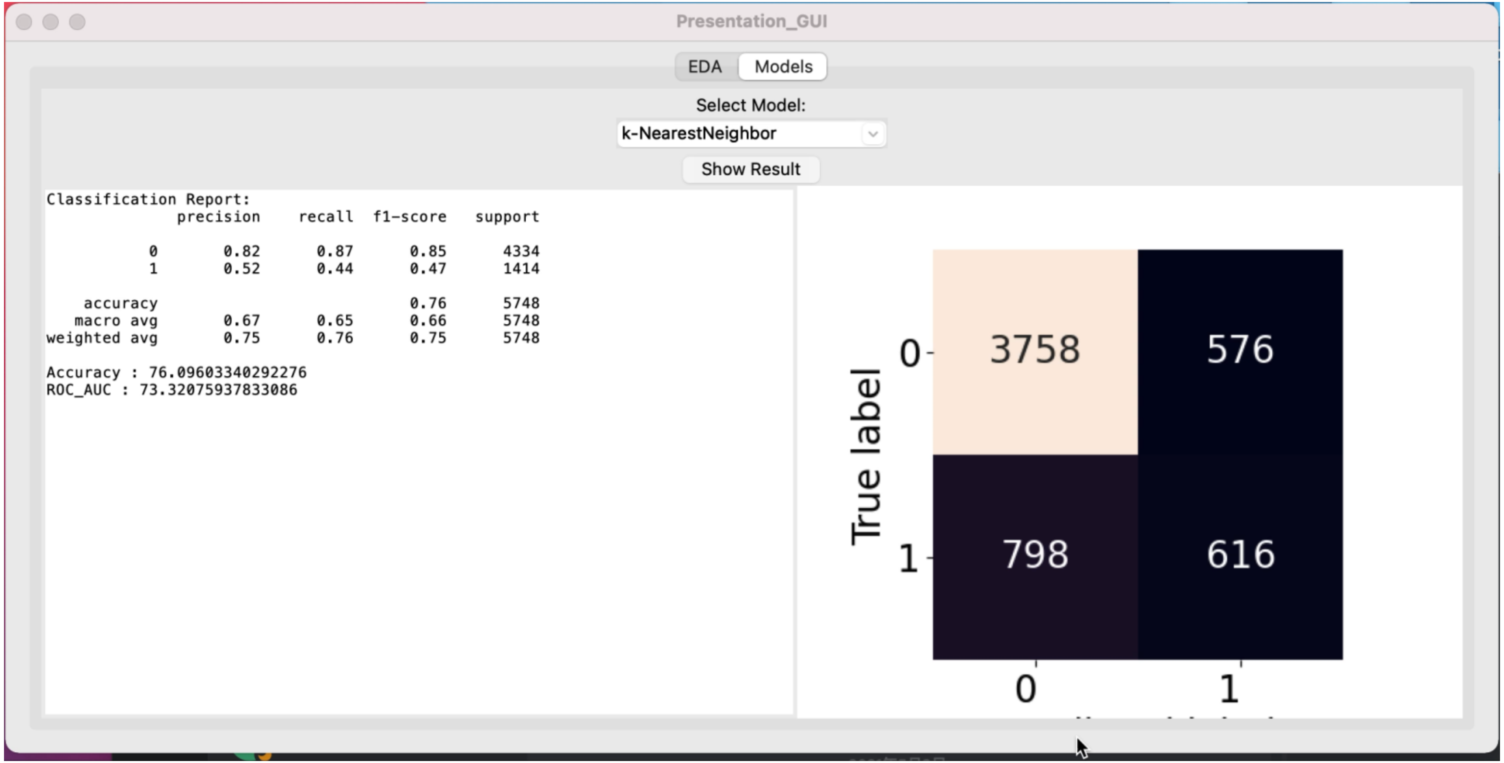
```
              precision    recall  f1-score   support

           0       0.87      0.87      0.87      4334
           1       0.60      0.60      0.60      1414

    accuracy                           0.80      5748
   macro avg       0.73      0.73      0.73      5748
weighted avg       0.80      0.80      0.80      5748

Accuracy: 80.28879610299235
ROC_AUC :  80.1568010318073
```

Features Importance



- The top 5 important features--- city_development_index, training_hours, experience, last_new_job and company_type.

GUI


Presentation_GUI
EDA
Models
Select Model:
k-NearestNeighbor
Show Result
Classification Report:
precision recall f1-score support
0 0.82 0.87 0.85 4334
1 0.52 0.44 0.47 1414
accuracy 0.76 5748
macro avg 0.67 0.65 0.66 5748
weighted avg 0.75 0.76 0.75 5748
Accuracy : 76.09603340292276
ROC_AUC : 73.32075937833086
True label
0
1
3758
576
798
616
0
1

Summary and Discussion

- Exploratory data analysis shows that city_development_index, experience, last_new_job , but not gender and training_hours, play important roles in the job change. However, feature importance shows that training_hours is the second important variable.
- Based on the comparison of accuracy, f1 score and ROC_AUC, Gradient Boosting shows the best performance to predict job change.
- Gradient Boosting is a great ML algorithm that handles categorical features and missing values.

THANKS

Q&A