Proposal for predicting a candidate looking for a new job

Akbanu Tleubayeva

TOC

- Overview
- Dataset description
- Data visualization
- Filling in missing values
- Data Preprocessing
- Encoding

Overview

1. 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.

2. 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.

Dataset description

	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_discipline	experience	company_size
0	8949	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM	>20	NaN
1	29725	city_40	0.776	Male	No relevent experience	no_enrollment	Graduate	STEM	15	50-99
2	11561	city_21	0.624	NaN	No relevent experience	Full time course	Graduate	STEM	5	NaN
3	33241	city_115	0.789	NaN	No relevent experience	NaN	Graduate	Business Degree	<1	NaN
4	666	city_162	0.767	Male	Has relevent experience	no_enrollment	Masters	STEM	>20	50-99

- **Data sources:**
- The datasets are sourced from the website kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners that allows users to find and publish datasets, explore and build models in a web-based data-science environment.
- The dataset aug_train.csv used in this project was collected in August, 2020. This data set contains **19,158 observations,** 14 columns. i.e 13 features + 1 label.

- 10 categorical variables: city, gender, relevent_experience, enrolled_university, education_level, major_discipline, experience, company_size,last_new_job,company_type

  3 numerical variables: enrollee_id, city_development_index,training_hours

  1 label variable - target

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

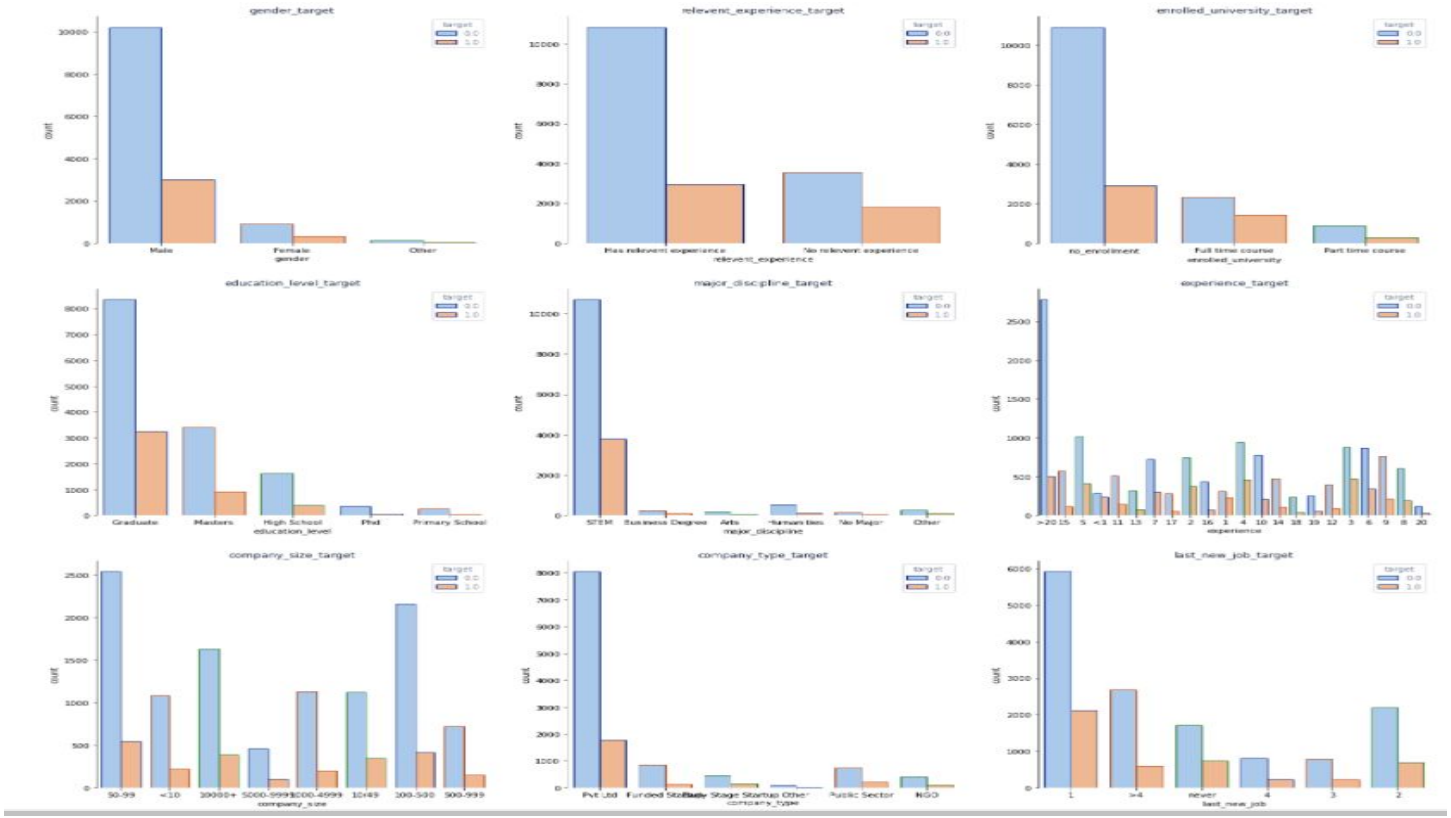
Key Findings

- 8 Columns have missing values.

	null_numm	null_percent
gender	4508	0.240000
enrolled_university	386	0.020000
education_level	460	0.020000
major_discipline	2813	0.150000
experience	65	0.000000
company_size	5938	0.310000
company_type	6140	0.320000
last_new_job	423	0.020000

1. The company type has the largest missing value, with a missing percentage of 32%.

2. The company size has the largest missing value, with a missing ratio of 31%.


gender_target
relevent_experience_target
enrolled_university_target
education_level_target
major_discipline_target
experience_target
company_size_target
company_type_target
last_new_job_target
target
0.0
1.0
count
Male
Female
Other
gender
Has relevent experience
No relevent experience
relevent_experience
no_enrollment
Full time course
Part time course
enrolled_university
Graduate
Masters
High School
Phd
Primary School
education_level
STEM
Business Degree
Arts
Humanities
No Major
Other
major_discipline
>20
15
5
<1
11
13
7
17
2
16
1
4
10
14
18
19
12
3
6
9
8
20
experience
50-99
<10
10000+
5000-9999
1000-4999
10-49
100-500
500-999
company_size
Pvt Ltd
Funded Startup
Early Stage Startup
Other
Public Sector
NGO
company_type
1
>4
never
4
3
2
last_new_job

1. According to the first picture, we can see that the proportion of men who are not looking for a job change far exceeds the proportion of men who are looking for a job change.

2. In the second picture, we can see that most candidates with relevant experience do not look for job changes in a large proportion.

3. in the types of registered courses, most people are not registered for courses and are not willing to look for job changes.

4. Most of these groups have a high degree of education.

5. The candidate's major is basically STEM. This shows that many people are not changing industries.

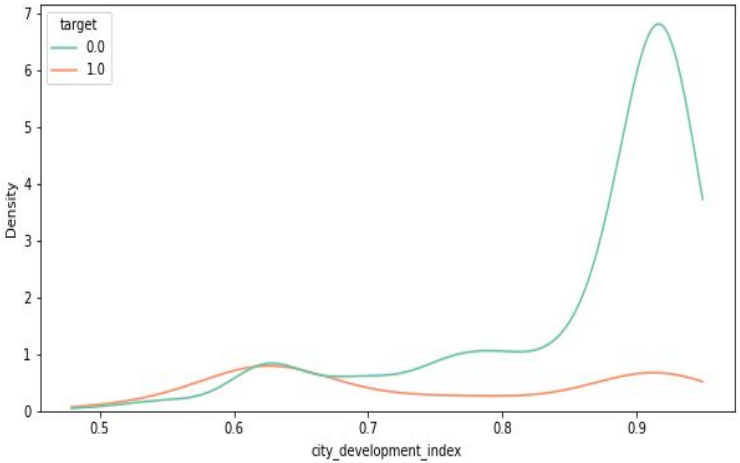
6. In the group that does not change their jobs, many people have more than 20 years of work experience. This can actually explain in disguise that the longer you work, the more you hope you can stabilize. In contrast, those with less work experience will have a significantly higher rate of changing jobs.

7. Among the groups that do not plan to change jobs, the number of their employer companies is basically between 50-500.

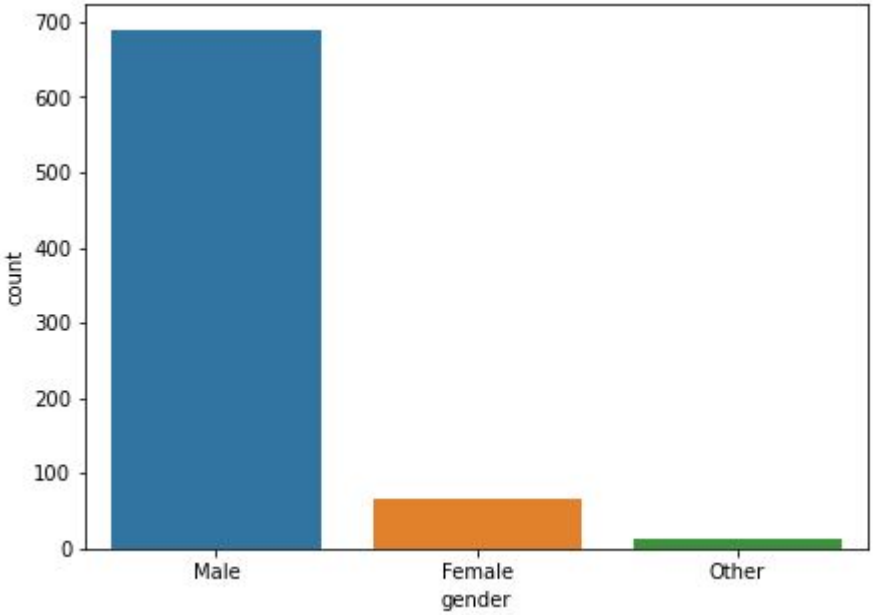
8. Among the groups that do not change their jobs, the type of employer is basically pvt ltd

9. The proportion of unchanged jobs exceeds the proportion of changed jobs, and employees who have just joined the company for about a year are less willing to change jobs.

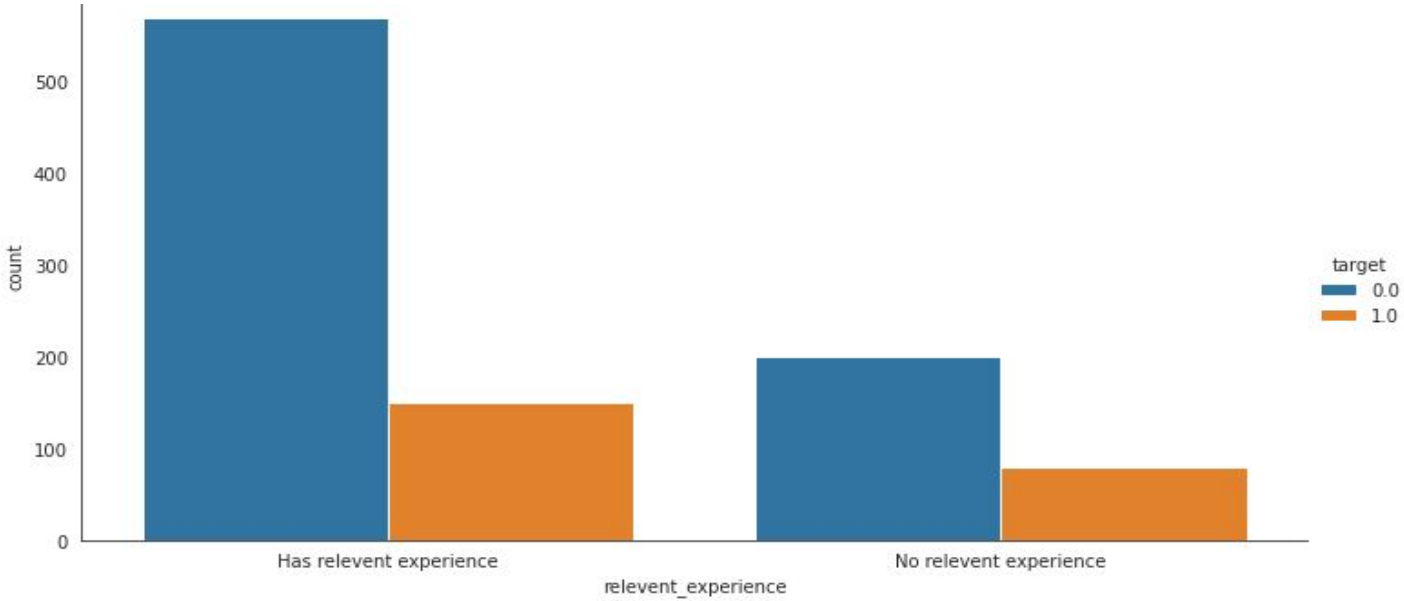
Visualisation and Missing Values Treatment



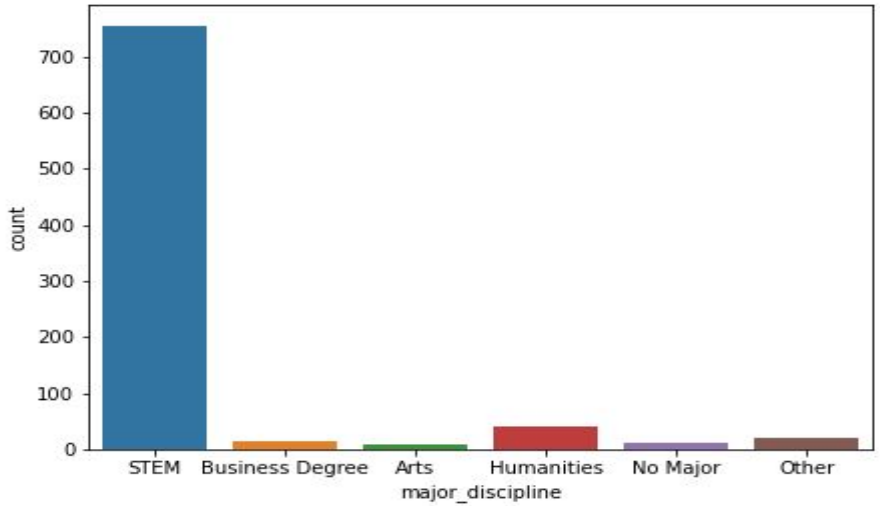
From this graph, we can observe that people who live in developed cities have a lower likelihood of changing occupations.



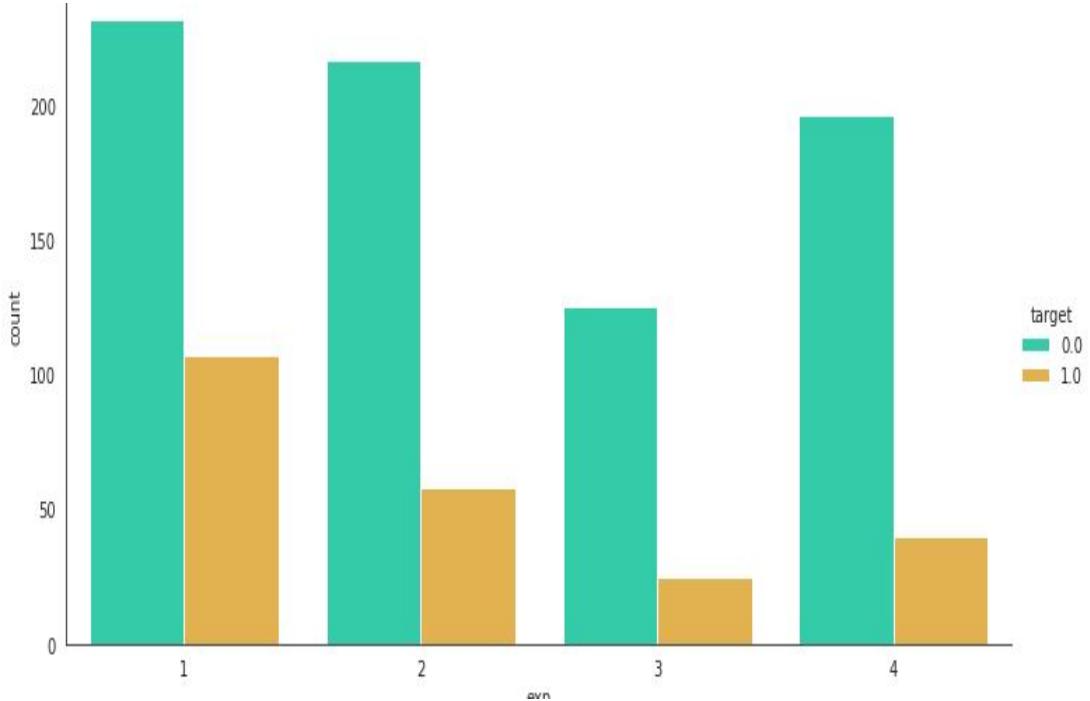
The number of male candidates who applied for the job is the highest.



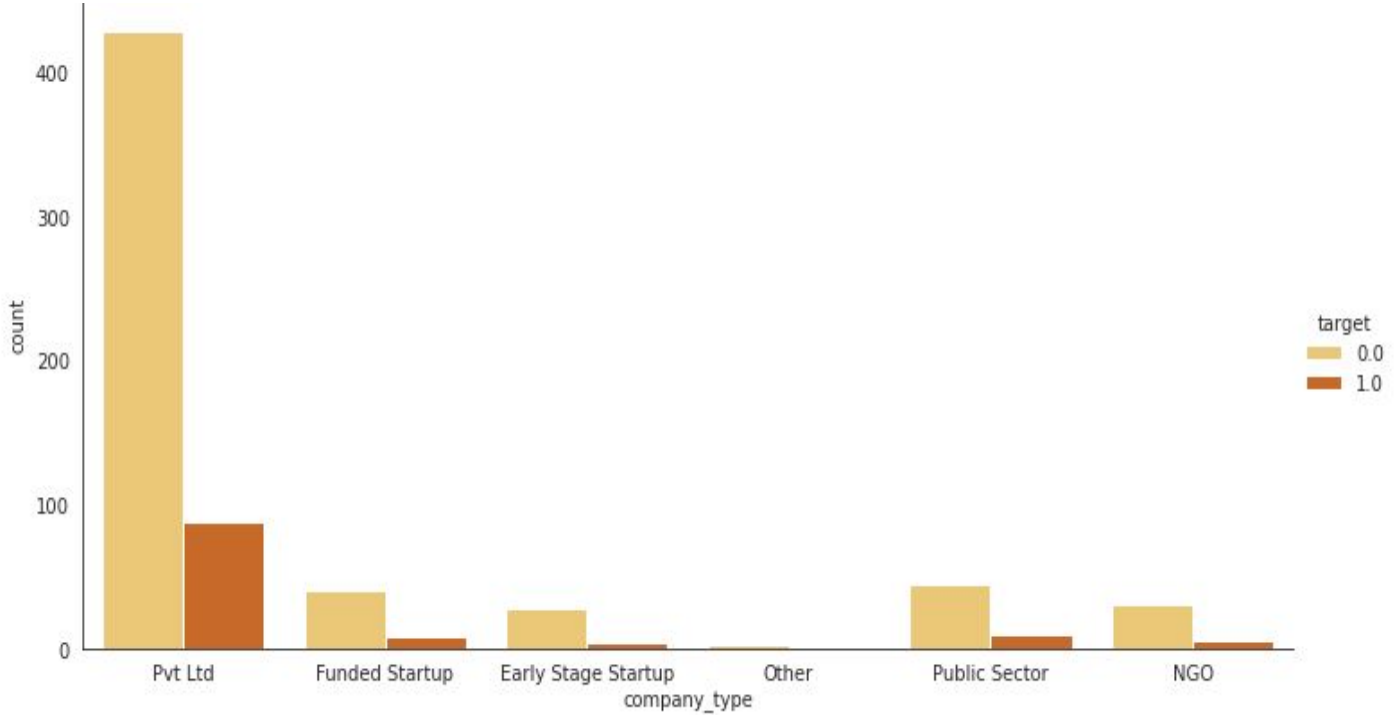
Here, from this chart, we can analyse that candidates who are having relevant experience in the concerned field are less likely to switch their job in comparison to the others who have no relevant experience.



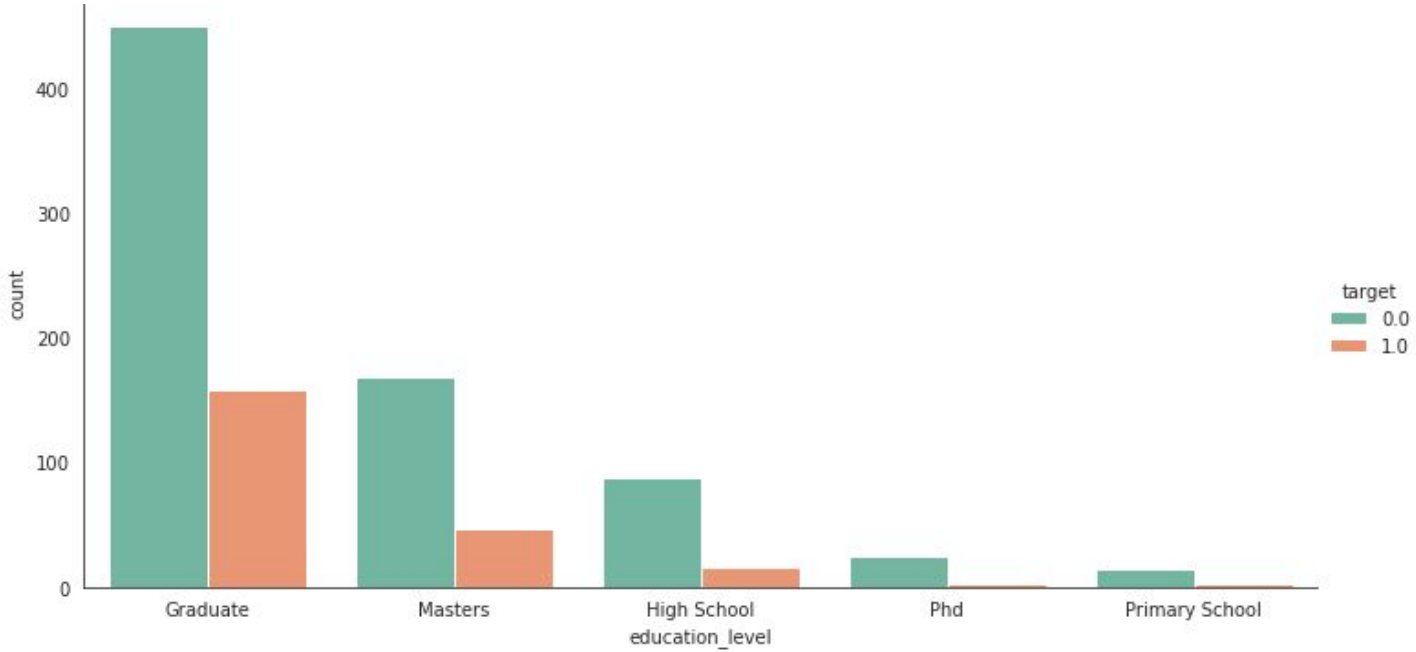
Majority of the candidates are from STEM.



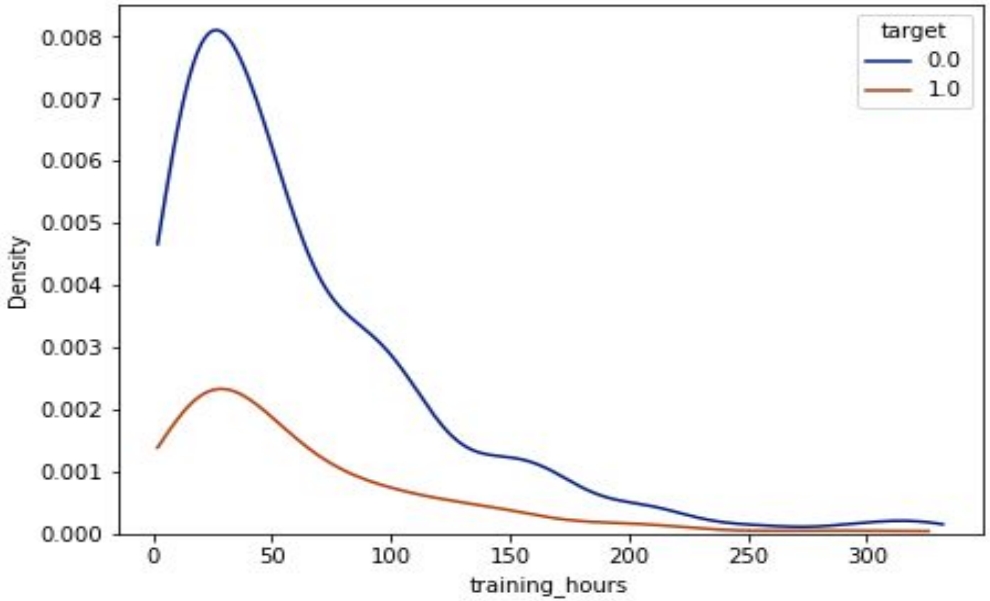
According to the analysis, those who are having less experience are more vulnerable to switch their job, rather than those who are having more experience.



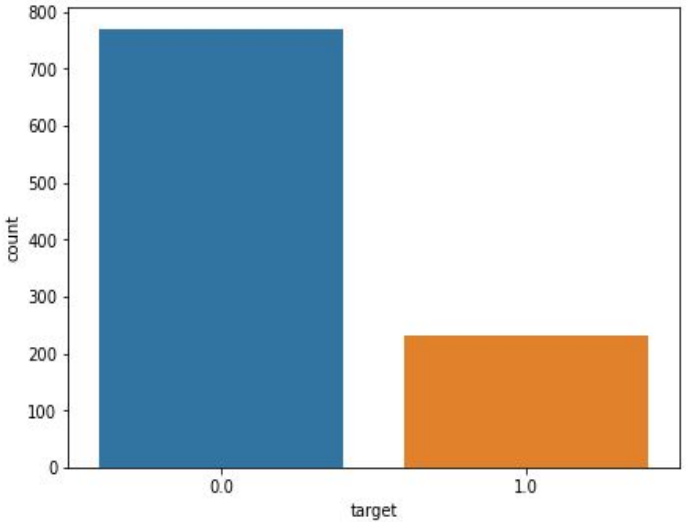
People who are working in private companies have lesser possibility of changing their job.



Most of the candidates are having graduation as their highest qualification followed by masters, and very few candidates are having Phd as their qualification.



From this graph, we can observe that employees with less training hours had a decreased likelihood of changing employment.



Here it is clearly visible that the data is highly imbalanced. In order to train the model properly, we need to balance and for the same we use SMOTE (Synthetic Minority Oversampling Technique) to balance the data.

Modeling and Analysis

Model Selection

We studied the performance of 5 classification models: Logistic Regression, Gaussian Naïve Bayes, k-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, BernoulliNB.

Final Model Selection and Hyperparameters Tuning

	Models	Accuracy Score
3	SVC	0.835
4	RandomForestClassifier	0.830
5	KNeighborsClassifier	0.805
0	LogisticRegression	0.800
2	BernoulliNB	0.775
1	GaussianNB	0.280

- Random Forest Classifier and SCV though comparatively time consuming, were the most accurate with the least number of misclassifications followed by Logistic Regression and KNeighborsClassifier.