Exploratory Data Analysis (EDA)

Note: *This paper presents diagrams, comments, and conclusions of the EDA analysis. If you want to display these diagrams in code, you need to comment/uncomment the indicated lines.*

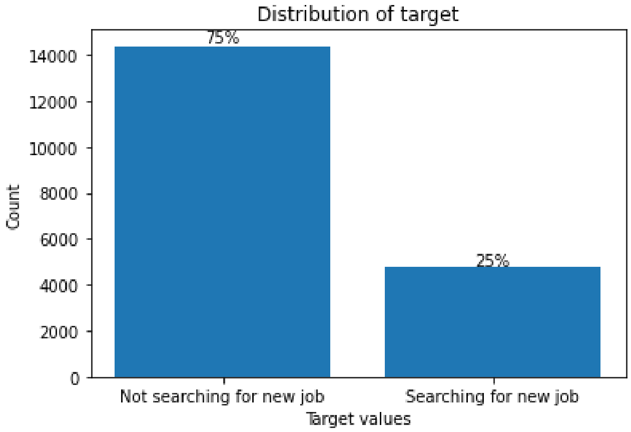

Fig. 1: Distribution of target

From the Fig. 1, we can see there is an imbalance in the classes – 75% candidates are not searching for a new job and 25% candidates searching for one.

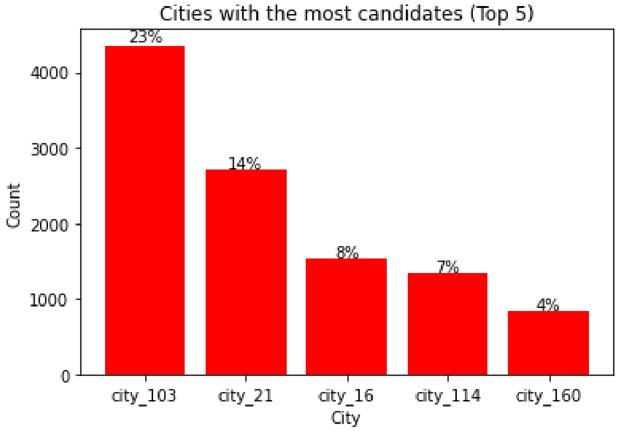

Fig. 2: Top 5 cities with the most candidates

From the Fig. 2, we can see that the top 4 cities represent more than half of the candidates.

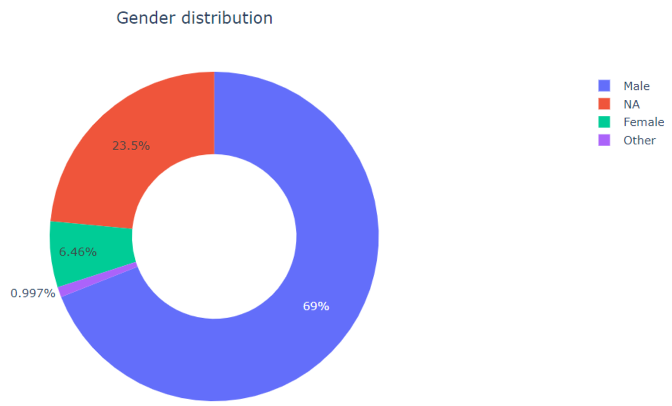

Fig. 3: Gender distribution

In the Fig. 3 it is shown that males make up a majority of participants. There are 23.5% of participants that didn’t share their gender.

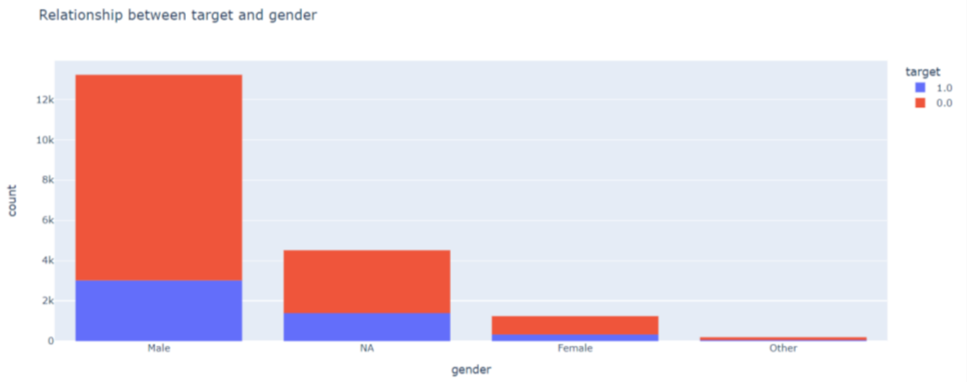

Fig. 4: Relationship between target and gender

The percentage of people looking for a new job is almost the same for all genders. Gender has no effect on changing the proportion of job seekers.

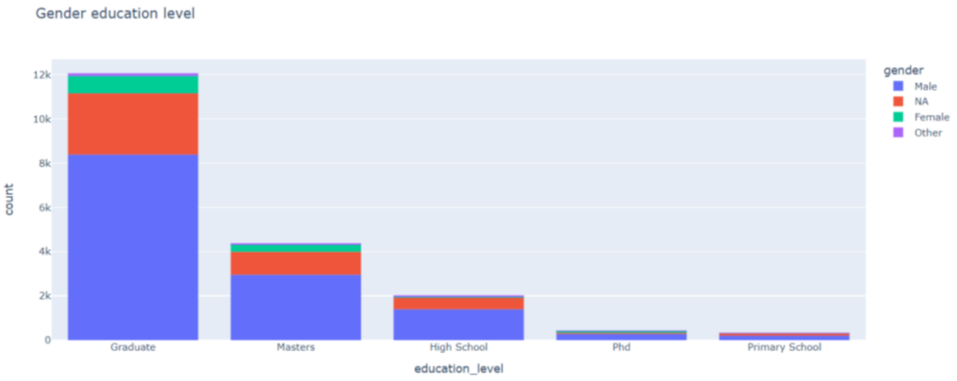

Fig. 5: Gender education level

We can observe that as education level increases, the percentage of females also increases.

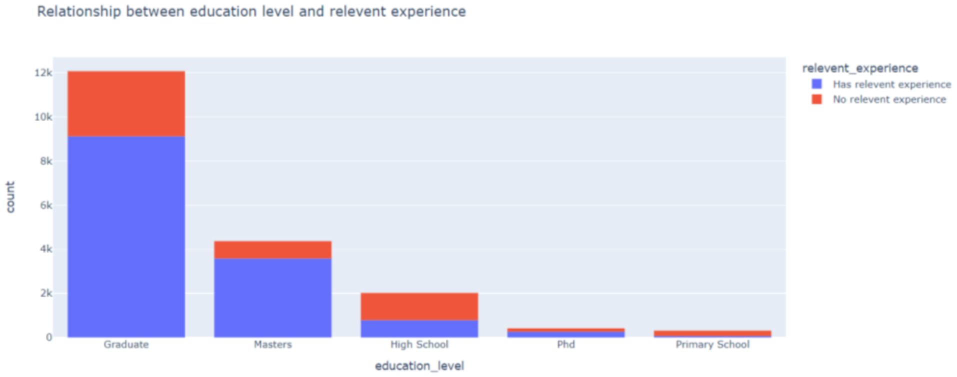

Fig. 6: Relationship between education level and relevant experience

From the Fig. 6, we can see that the higher education level means the greater amount of candidates with relevant experience in the data science field.

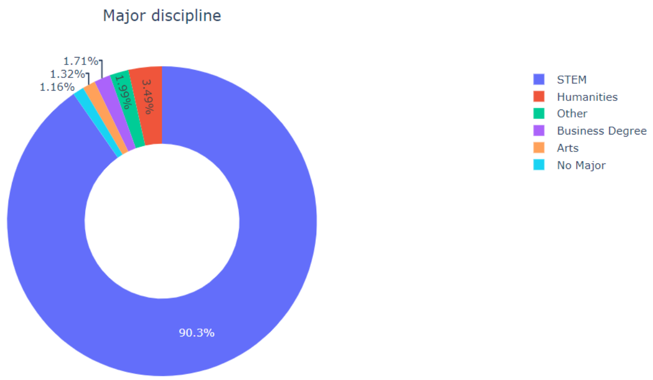

Fig. 7: Major discipline

STEM is the most common discipline for data scientists.

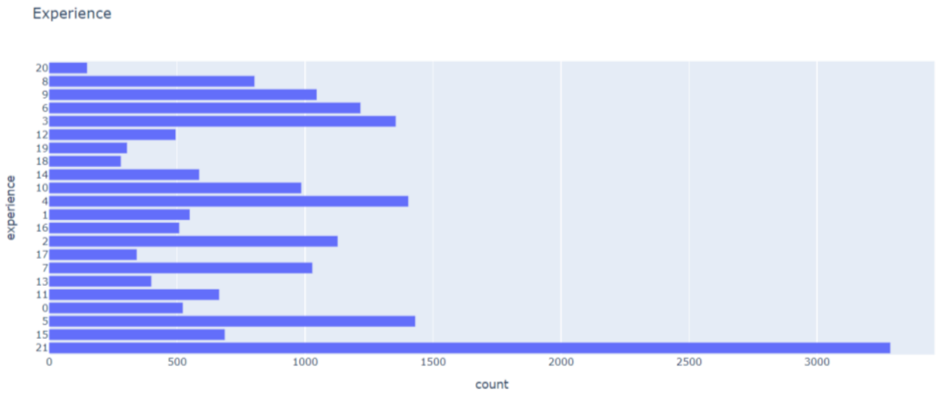

Fig 8: Distribution of experience

Most of the candidates have more than 20 years of experience.

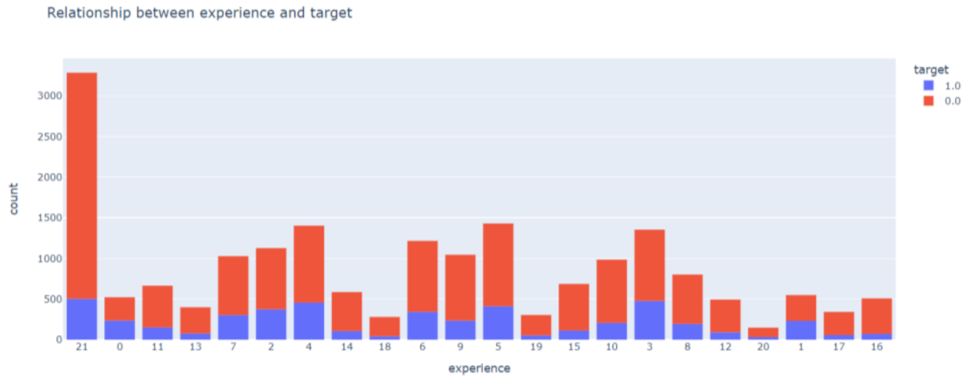

Fig. 9: Relationship between experience and target

The less work experience, the probability that the candidate is searching for a new job is higher.

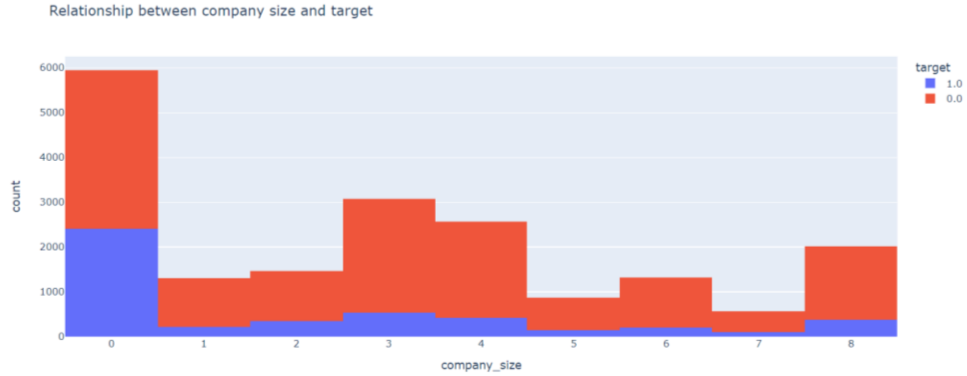

Fig. 10: Relationship between company size and target

If we look at the percentage, the proportion of candidates looking for a new job is consistent across all company sizes.

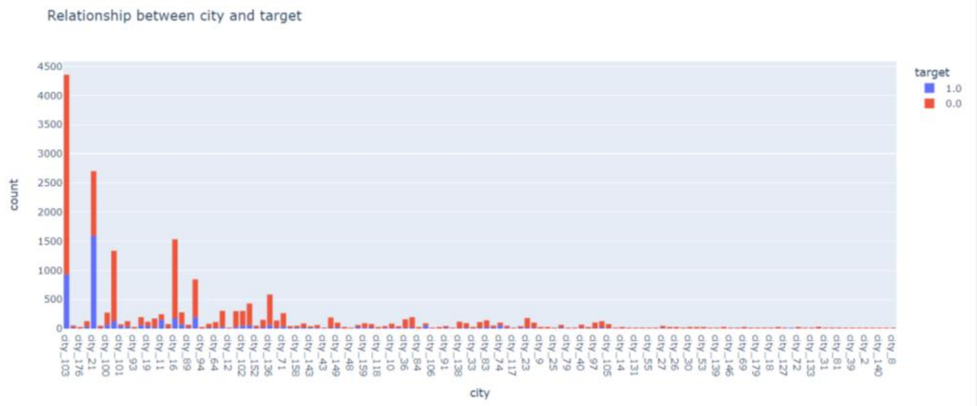

Fig. 11: Relationship between city and target

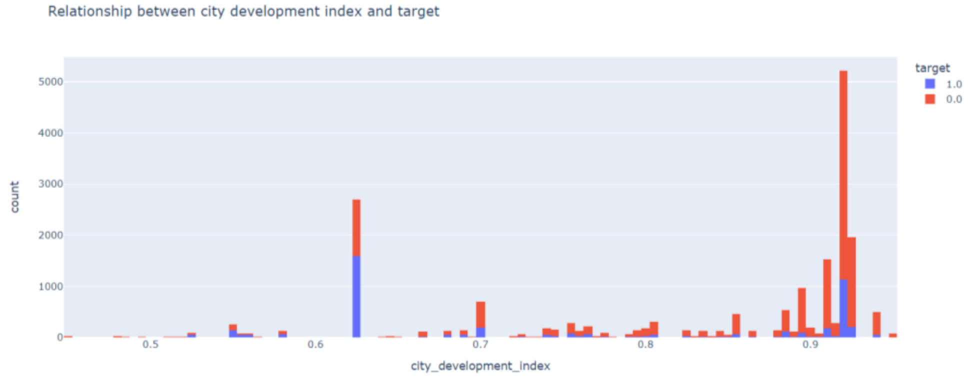

Fig. 12: Relationship between city development index and target

From the Fig. 11 and Fig. 12, we can observe that city and city development index has a very strong relationship and, for that reason, we'll remove the city variable for the predictive model.

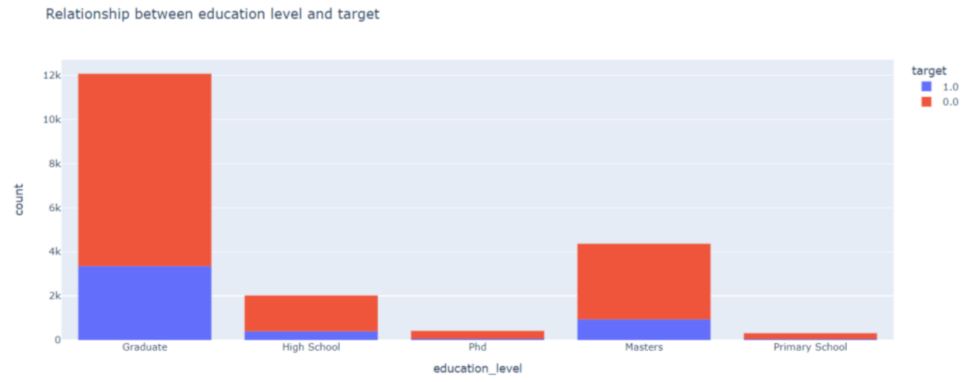

Fig. 13: Relationship between education level and target

The percentage of people looking for a new job is almost the same for all education levels.

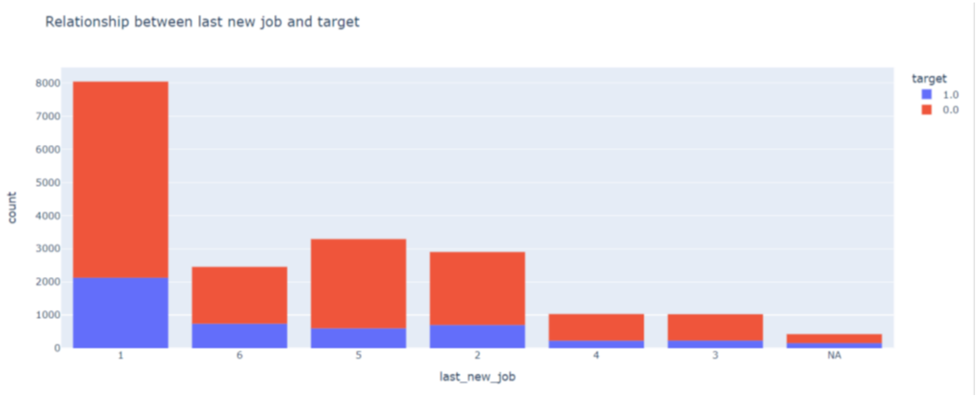

Fig. 14: Relationship between last new job and target

From the Fig. 14, we can observe that we have similar distribution for candidates who are searching for a new job and candidates who are not searching for a new job.

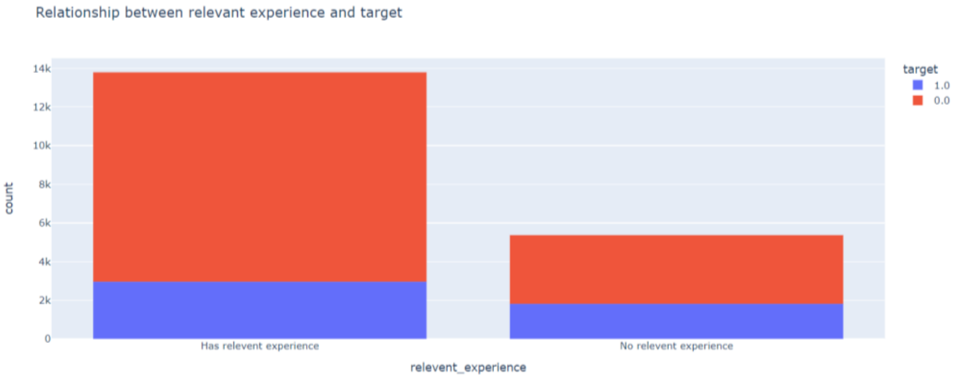

Fig. 15: Relationship between relevant experience and target

The distribution of job seekers is equal in two categories shown on Fig. 15.

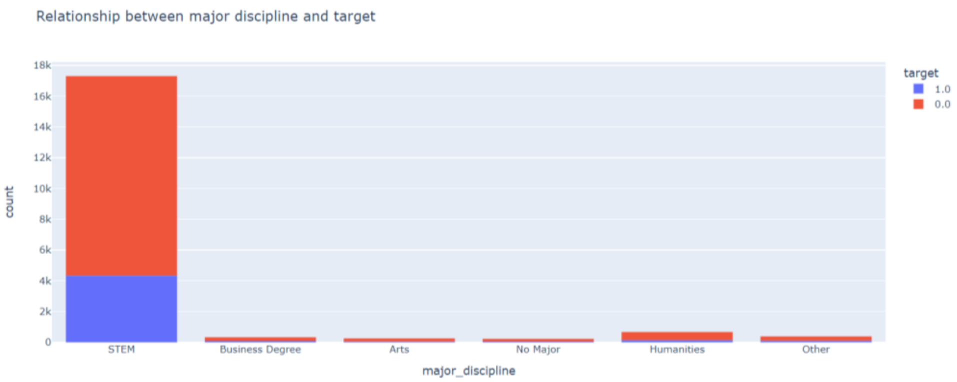

Fig. 16: Relationship between major discipline and target

The distribution of job seekers is equal in all categories shown on Fig. 16.

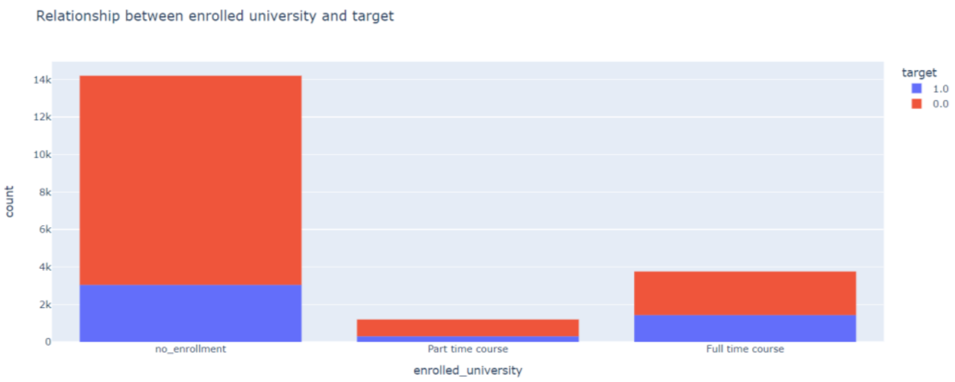

Fig. 17: Relationship between enrolled university and target

From the Fig. 17, we can observe that knowing the enrolled university of the candidate gives us information about whether the person will be looking for a new job.

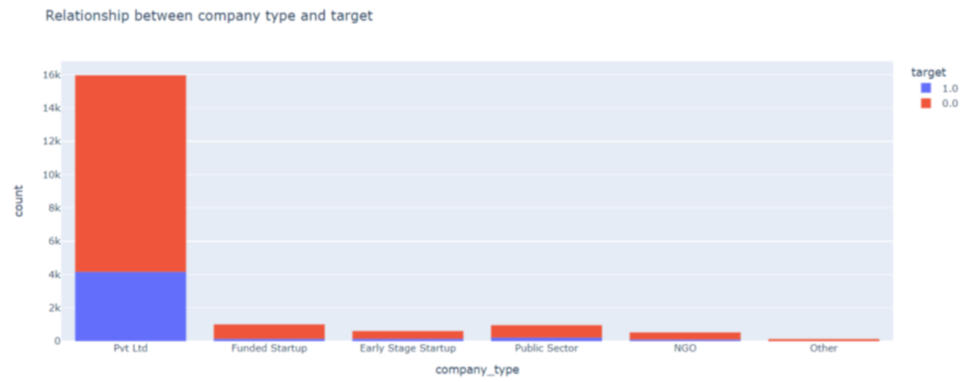

Fig. 18: Relationship between company type and target

From the Fig. 18, we observe that knowing the company type of the candidate giving us information about whether the person will be looking for a new job.

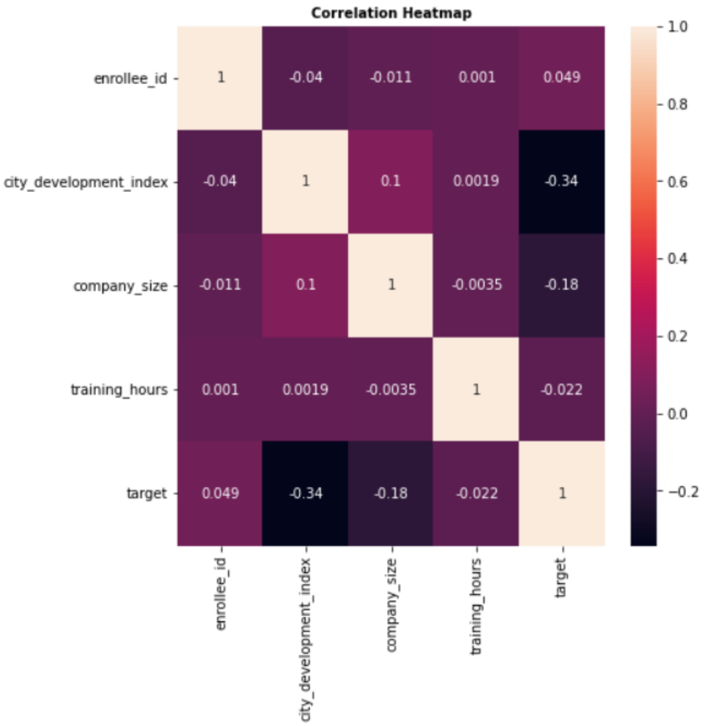

Fig. 19: Correlation Heatmap

From the Correlation Heatmap (Fig. 19), we can observe that the correlation is stronger between the city development index and target. Also, we can see that correlation between target and training hours also exist, but the abs value of this correlation is lower.

Comment on (un)balanced data and possible effects

From the distribution of target (Fig. 1), we can see that our dataset is unbalanced. This means that a dataset is biased towards a class in the dataset. If the dataset is biased towards one class, an algorithm trained on the same data will be biased towards the same class. We need to train some of the algorithms that can handle this type of unbalanced data better, e.g. logistic regression, SVM, decision trees, bagging.