1. Problem Statement

A company wants to hire data scientists among people who successfully pass the courses they conduct. There is a large number of people who sign up for their training and they want to know which of these candidates want to work for the company or are seeking new employment opportunities. Their objective is to reduce the cost and time spent on the recruitment process by predicting whether the candidate is looking for a job change. This is a binary classification problem, and we are interested in classifying: 1 – if the candidate is looking for a job change or 0 – if the candidate is not looking for a job change

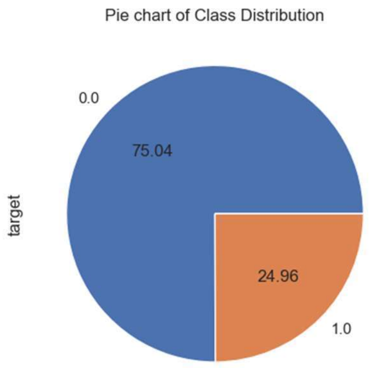
2. Dataset Description

This analysis will focus on several factors that could influence a person to seek new employment opportunities in data science. Specifically, our model and observations will depend on the following variables:

- **city**: The city code of where the candidate is from
- **city_development_index**: Development index of the city (scaled)
- **relevant_experience**: If the person has or does not have relevant experience
- **enrolled_university**: Is the person studying part time, full time or not enrolled in any courses
- **education_level**: The level of education the person has achieved
- **major_discipline**: Education discipline of the candidate
- **experience**: The candidate's total experience in years from '<1' to '>20'
- **last_new_job**: Difference in years between their previous job and current job from 'never' to '>4'
- **training_hours**: The number of training hours they completed

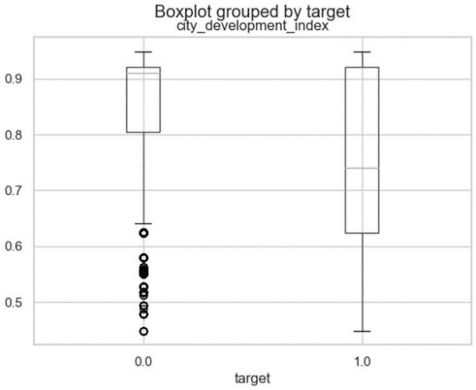
The original dataset also included variables such as enrollee_id, gender, company_size, and company_type. We dropped enrollee_id from our dataset because it does not provide any predictive power in classifying the target variable. Gender was not considered in our analysis because we did not want to introduce potential gender bias since there is an imbalance in male to female candidates. Finally, company_size and company_type had over 30% missing values so we disregarded them from our analysis. After preprocessing and handling missing values, we were left with 18,643 observations from 19,158.

3. Dataset Analysis and Observations

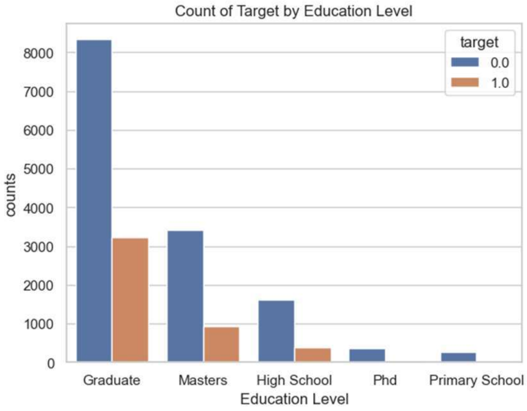


To analyze the dataset, it is important to understand the distribution of our target variable.

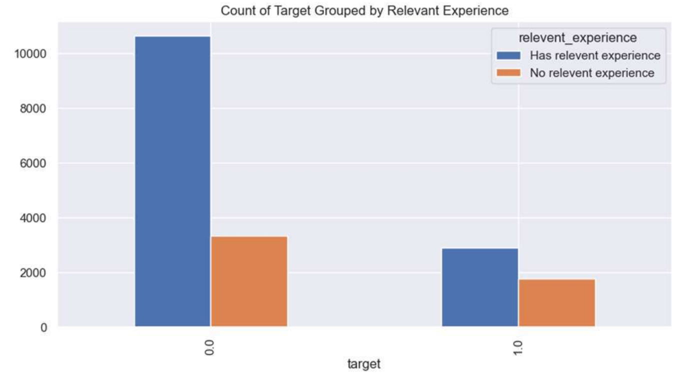
We can see that approximately **75%** of the candidates are not looking for a job change. We will need to be mindful when building our model of this imbalance, because we do not want a model that is biased towards the majority class and deny potential candidates an opportunity at the company.



The boxplot on the right shows the distribution of **city_development_index** grouped by the target variable. Among those who are not interested in a new job, the distribution is heavily skewed and the majority of them seem to be from cities that are more developed. This may be due to more available opportunities other than the one offered at the company of interest in comparison to those from lesser developed cities.



The proportion of candidates seeking new job opportunities is similar among people with high school, graduate, and masters level education. There are almost no candidates with a PhD or primary school level education who are interested in the job. This may be because those who only have primary school education feel underqualified for the opportunity, whereas those with a PhD feel they are overqualified.

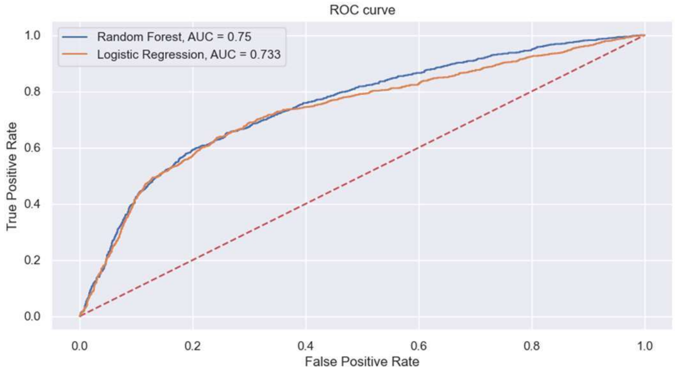


Finally, the figure on the right displays the distribution of the target variable grouped by whether or not a candidate has relevant experience. It appears that there is a noticeably greater proportion of people who are interested in the job that do not have relevant experience compared to those who are not interested. This may be an indication that the opportunities available for this company are great for those at an entry level.

4. Proposed Analytical/Prediction Model

The majority of our features are categorical. Only city_development_index and training_hours are continuous variables. From our categorical features, education_level, experience, and last_new_job are ordinal. For this binary classification problem, we will train a Random Forest Classifier and Logistic Regression model. Random Search Cross Validation will be used for hyperparameter tuning. Since the target variable is imbalanced, prediction accuracy is a misleading metric to evaluate model performances. Our results will be compared using precision, recall, F1 score and AUC/ROC. We will use weights to discourage our model from being biased towards the majority class.

5. Results and Discussions



Metrics (Weighted Average)	
Random Forest	**Logistic Regression**
Accuracy: 0.75	**Accuracy:** 0.72
Precision: 0.75	**Precision:** 0.71
Recall: 0.75	**Recall:** 0.72
F1 Score: 0.75	**F1 Score:** 0.71

The Random Forest Classifier was better across all metrics, but the ROC curve shows both models performed similarly. When fitting our model, we included weights to avoid too many false negatives, but our primary objective is not to maximize recall. We need to keep in mind that the company does not want to waste resources on people who will not move forward with the job process, so eliminating candidates at this step of their HR analysis is not detrimental to their objective. Candidates with a higher probability of being interested should be prioritized overall.