International University, HCMC National University
School of Computer Science and Engineering

FINAL REPORT

HR Analytics: Potential trainees to become Data Scientists

Submitted by **GROUP 2**

No.	FULL NAME	STUDENT ID	PHONE NUMBER	Evaluation (%)
1	Tôn Võ Thu Uyên	ITDSIU20135	0961357102	50
2	Lê Ngọc Uyên Phương	ITDSIU20079	0867754712	50

Course name: Data Analysis

Professor: Assoc. Prof. Nguyen Thi Thuy Loan, PhD

December 02, 2022

TABLE OF CONTENTS

Abstract

Data analytics is the process of analyzing a dataset to make judgments about the information it contains. **HR analytics**, which is an application of data analytics, is a data-driven approach to Human Resources Management that enables your company to measure the impact of various HR metrics on overall business performance, then make data-driven decisions [1].

In other words, **HR analytics** is about changing the way of recruiting and retaining talent based on data-driven insights (Isson & Harriot, 2016) [2]. This is done by applying statistical methods to analyze integrated HR, talent management, financial, and operational data. The measurements of the HR function, such as the time to hire, the cost of training per employee, and the duration until promotion, are the focus of HR analytics. One of the most popular concern of HR analytics is to predict the turnover rate of employees in a company.

Turnover, commonly referred to as "attrition" in this project, is the phenomenon through which an employee leaves a company for some reason. An employee who leaves their work within the first year of employment is referred to as a **new hire attrition** [3]. Every organization must cope with turnover, but it is particularly challenging when that turnover comes from a recently hired employee. High turnover rates have negative effects on an organization's knowledge, skills, and investment as well as on employee morale, which can spread quickly. One risk of a high new hire turnover rate is that it frequently has negative consequences that cause your turnover to keep rising [4].

Chapter 1. Introduction

A company specializing in Big Data and Data Science fields is looking for data scientists among those who have completed some of the company's training programs. The training attracts a lot of participants, but not all of them choose to stay with the company. Therefore, the company wants to know which of these candidates really wants to work for them after training or looking for new employment. **HR analytics** can be used to determine the factors that lead a person to quit their current employment in order to enhance the quality of training, course planning, and candidate classification. This helps to cut down on time and money for the business [5].

The major goal of this project is to identify the causes of new hire employee turnover and forecast whether a candidate will work for the company or look for another position. To better understand the features in our dataset, we will perform exploratory data analysis (EDA) and other types of data analysis. Then, machine learning models as kNN, Logistic Regression, and Decision Tree will be applied for prediction.

Chapter 2. Project Overview

2.1. Project Objectives

This project mainly aims to test our understanding of the business analysis process which includes data collecting, exploring, cleaning, preprocessing, analyzing, modeling, and visualization.

It is also a practical example of how Data Analytics is used in the field of Human Resources. After going through all of the steps in the analysis process, this project should be able to:

- Use Exploratory Data Analysis to gain better understanding on the dataset.
- Use statistical methods to calculate some important descriptive statistics.
- Create detailed illustrations of how each variable affects the final decision.
- Use preprocessing methods to convert the raw data with anomalies and outliers to the clean data that can be used for training model.
- Predict the probability of a candidate who will work for the company.
- Interpret model(s) in such a way that illustrates which features affect candidate decision with prediction accuracy.

2.2. Project Scope

- Step 1: Data overview
- Step 2: Exploratory Data Analysis (EDA)
- Step 3: (Preprocessing) Data cleaning
- Step 4: Data analysis
- Step 5: (Preprocessing) Data transformation
- Step 6: Modeling (Predictive analysis)

2.3. Development Tools

In this project, Python is the main programming language used to prepare and process the data on Google Colab (Google Colaboratory). Some of the Python libraries used in this project are:

- Pandas, Numpy
- Matplotlib, Seaborn, Plotly
- Missingno
- Category Encoders
- Scipy stats
- (Sklearn) train_test_split, StandardScaler, PCA
- (Sklearn) LogisticRegression, DecisionTreeClassifier

Chapter 3. Data Overview

3.1. Data collection

Kaggle is a platform for data scientists and machine learning enthusiasts to interact online, find inspiration **with a large number of free datasets**, learn new coding techniques, and see real-world data science applications through competitions. There are over 8 million people registered on Kaggle until 2021. Competitions on Kaggle are one of the sub-platforms that contributed to its popularity. "Kaggle Competitions" is crucial for data scientists, much like HackerRank serves the same purpose for programmers and computer engineers.

The "**HR Analytics: Job Change of Data Scientists" dataset from Kaggle** is where we got the data for our project. The business provides us with 2 datasets, which are the train and the test datasets.

The **train dataset** contains **14 attributes** with data on **19158** training candidates. The information includes the trainees' ids, addresses, genders, experience, levels of education, etc. Different from the test dataset, the train dataset has a "target"

column, which can serves as the label for a model, that stores the decision of candidates after the training program in the form of binary values 0 and 1.
A preview sample of the train dataset is provided below.

	enrollee_id	city	city_development_index	gender	relevant_experience	enrolled_university	education_level
0	8949	city_103	0.920	Male	Has relevant experience	no_enrollment	Graduate
1	29725	city_40	0.776	Male	No relevant experience	no_enrollment	Graduate
2	11561	city_21	0.624	NaN	No relevant experience	Full time course	Graduate
3	33241	city_115	0.789	NaN	No relevant experience	NaN	Graduate
4	666	city_162	0.767	Male	Has relevant experience	no_enrollment	Masters
...	...	...	...	...	...	...	...
19153	7386	city_173	0.878	Male	No relevant experience	no_enrollment	Graduate
19154	31398	city_103	0.920	Male	Has relevant experience	no_enrollment	Graduate
19155	24576	city_103	0.920	Male	Has relevant experience	no_enrollment	Graduate
19156	5756	city_65	0.802	Male	Has relevant experience	no_enrollment	High School
19157	23834	city_67	0.855	NaN	No relevant experience	no_enrollment	Primary School

19158 rows × 14 columns

Figure 3.1a. HR dataset (train)

	major_discipline	experience	company_size	company_type	last_new_job	training_hours	target
0	STEM	>20	NaN	NaN	1	36	1.0
1	STEM	15	50-99	Pvt Ltd	>4	47	0.0
2	STEM	5	NaN	NaN	never	83	0.0
3	Business Degree	<1	NaN	Pvt Ltd	never	52	1.0
4	STEM	>20	50-99	Funded Startup	4	8	0.0
...	...	...	...	...	...	...	...
19153	Humanities	14	NaN	NaN	1	42	1.0
19154	STEM	14	NaN	NaN	4	52	1.0
19155	STEM	>20	50-99	Pvt Ltd	4	44	0.0
19156	NaN	<1	500-999	Pvt Ltd	2	97	0.0
19157	NaN	2	NaN	NaN	1	127	0.0

Figure 3.1b. HR dataset (train) (cont.)

3.2. Attributes Description

The dataset has 14 attributes, but most of them are categorical variables, such as Nominal and Ordinal.

Table 3.1. Attributes description

No.	Attribute	Description	Data type
1	enrollee_id	Unique ID for candidate	int
2	city	City code	string
3	city_ development _index	Development index of the city (scaled)	float
4	gender	Gender of candidate	string
5	relevant_experience	Relevant experience of candidate	string
6	enrolled_university	Type of University course currently enrolled (if any)	string
7	education_level	Education level of candidate	string
8	major_discipline	Education major discipline of candidate	string
9	experience	Candidate total experience (in years)	string
10	company_size	Number of employees in current employer's company	string
11	company_type	Type of current employer	string
12	last_new_job	Difference in years between the previous job and current job	string
13	training_hours	Training hours completed	int
14	target	0 – Not looking for a job change, 1 – Looking for a job change	float

3.3. Dataset description

There is no duplicate values in this dataset. However, half of the columns in the train dataset contain missing values (Table 3.2), so we cannot drop all those values but try to impute them instead.

Table 3.2. Dataset quality and unique

No.	Attribute	Data Type	Count	Missing Value	Unique Value
1	enrollee_id	int	19158	0	19158

2	city	string	19158	0	123
3	city_ development _index	float	19158	0	93
4	gender	string	14650	4508	3
5	relevant_experience	string	19158	0	2
6	enrolled_university	string	18772	386	3
7	education_level	string	18698	460	5
8	major_discipline	string	16345	2813	6
9	experience	string	19093	65	22
10	company_size	string	13220	5938	8
11	company_type	string	13018	6140	6
12	last_new_job	string	18735	423	6
13	training_hours	int	19158	0	241
14	target	float	19158	0	2

Chapter 4. Exploratory data analysis (EDA)

Exploratory data analysis (EDA) provides a visual approach to explain the data, which makes it easier to gain the most understanding of the characteristics and key elements of the data, to discover previously unknown relationships, and to detect outliers and missing numbers. It is frequently the first stage in data analysis, carried out before any additional processing steps like data preprocessing and modeling.

A variety of plots and graphs can be used when conducting a visual analysis of data. Bar plots, histograms, box plots with whiskers, and other visualization techniques can be used to explore one variable (univariate analysis), while scatter plots are used in multivariate analysis [6].

4.1. Numerical data

Table 4.1. Numerical attribute descriptive statistics

	count	mean	std	min	25% (Q1)	50% (Q2)	75% (Q3)	max
enrollee_id	19158.0	16875.36	9616.29	1.00	8554.25	16982.5	25169.75	33380.00

city_ development_index	19158.0	0.83	0.12	0.45	0.74	0.9	0.92	0.95
training_hours	19158.0	65.37	60.06	**1.00**	**23.00**	**47.0**	**88.00**	**336.00**
target	19158.0	0.25	0.43	0.00	0.00	0.0	0.00	1.00

The train dataset has 4 numerical attributes, which fortunately contain no missing values. Since “enrollee_id” is just an ID attribute, its descriptive values bring about no valuable insights, so it needs to be converted into categorical types.

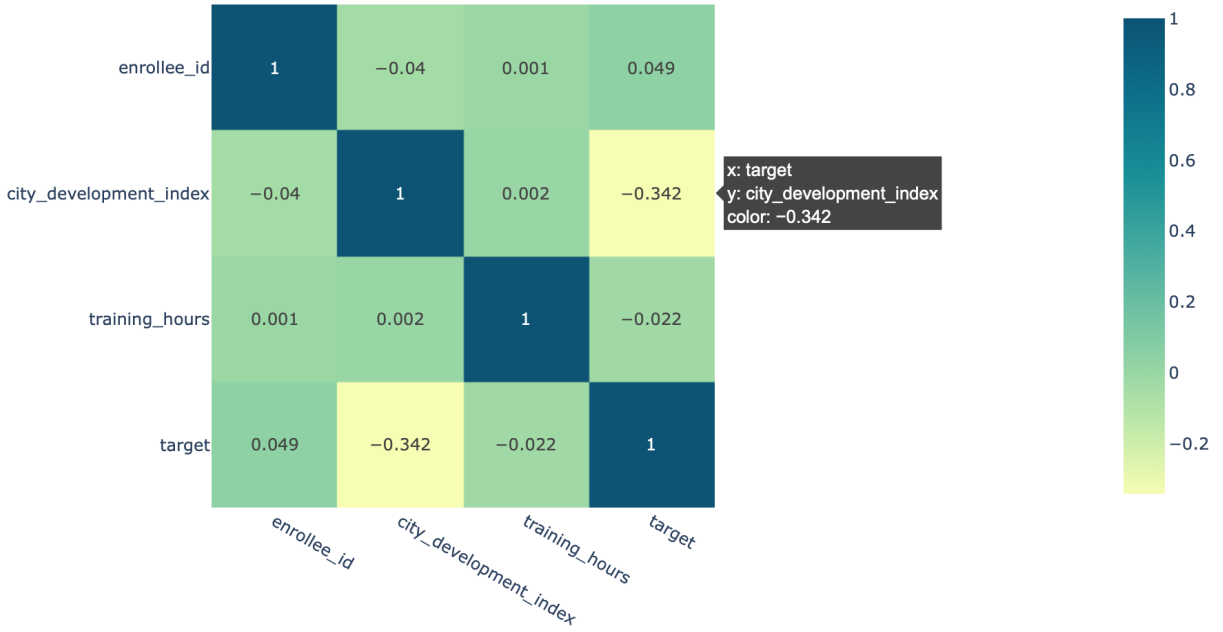


Figure 4.1. Correlation between Numerical attributes

There is a negative correlation (-0.342) between "city_development_index" and "target", which means that when the values of "city_development_index" increase, the values of "target" decrease, and the trainees are more likely to stay with the company. The correlations between other numerical attributes and "target" are close to 0, so they can be considered as no correlation, as in Figure 4.2.

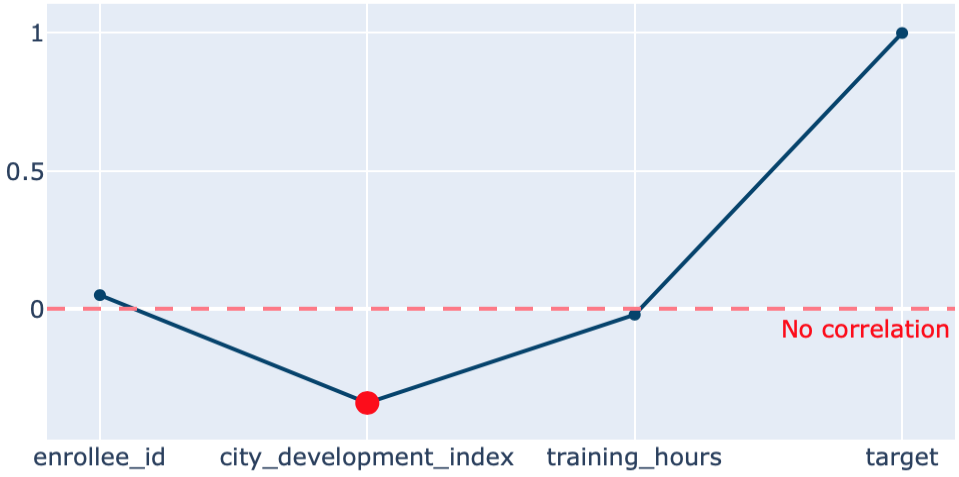


Figure 4.2. Correlation of Numerical attributes with target

Since "enrollee_id" has too many values for candidate IDs, and "target" has too few values as binary numbers 0 and 1, only the distributions of the remaining 2 numerical attributes can be plotted for analysis.

The **City Development Index** was created in 1996 for the Second United Nations Conference on Human Settlements (Habitat II) as a measure of average well-being and access to urban facilities by individuals. It is used to measure urban poverty and urban governance as well as rank cities around the world according to their degree of development. The five subindices that make up the CDI are: infrastructure, waste, health, education, and city product. The CDI, which has values ranging from 0 to 100, is then created by combining those sub-indices.

Table 4.2. City development index formula

Index	**Formula**
Infrastructure **(1)**	(Water connections x 25) + (Sewerage x 25) + (Electricity x 25) + (Telephone x 25)
Waste **(2)**	(Wastewater treated x 50) + (Formal solid waste disposal x 50)
Health **(3)**	[(Life expectancy - 25) x 50] / 60 + [(32 - Child mortality) x 50] / 31.92
Education **(4)**	(Literacy x 25) + (Combined enrollment x 25)
City product **(5)**	[(log City Product - 4.61) x 100] / 5.99
City Development Index (CDI)	[**(1)** + **(2)** + **(3)** + **(4)** + **(5)**] / 5

Figure 4.3 shows that most of the trainees live in cites with "city_development_index" higher than 0.9.

"training_hours" has a large number of outliers on the right side. The distribution shows that most of the candidates have up to 65 hours of training, but there is also someone with over 300 hours, so these can be outliers that should be removed carefully to avoid losing important information.

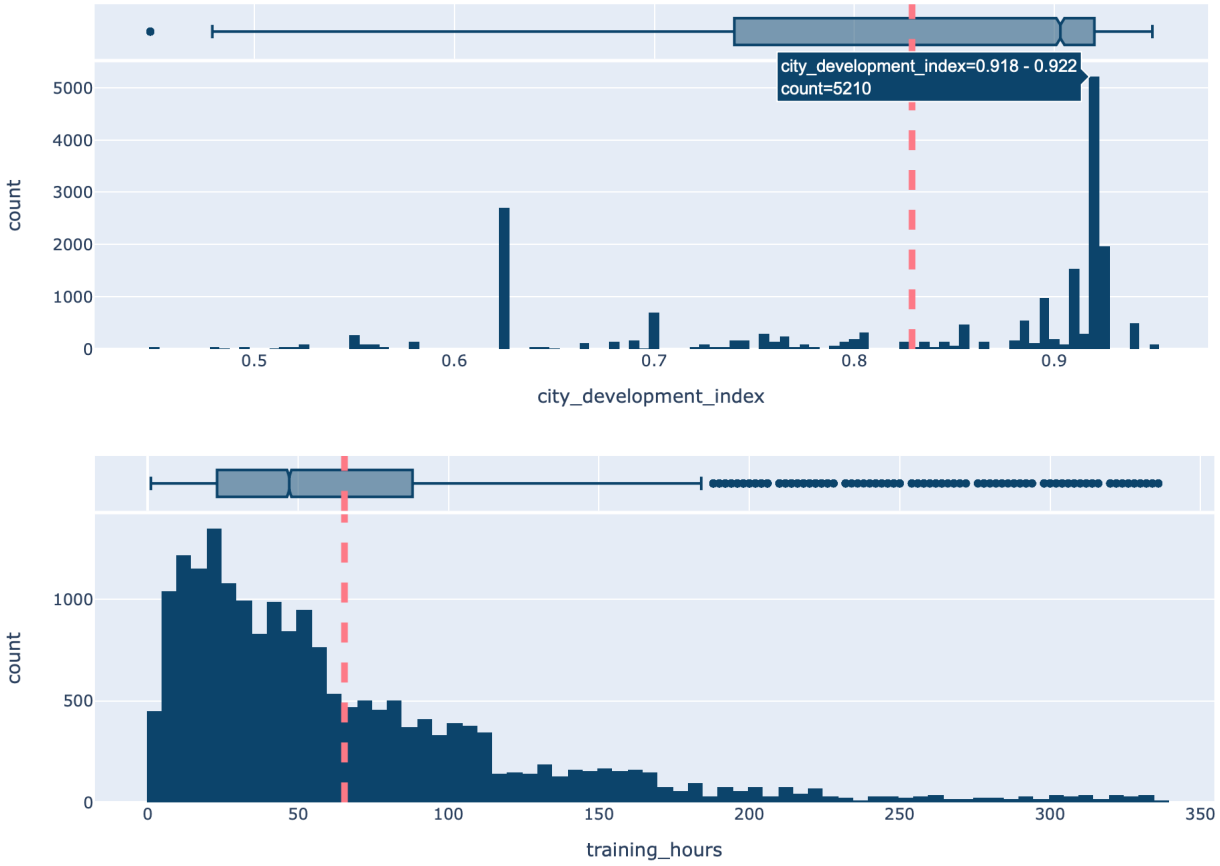


Figure 4.3. Distributions of “city_development_index” & “training_hours”

4.2. Categorical data

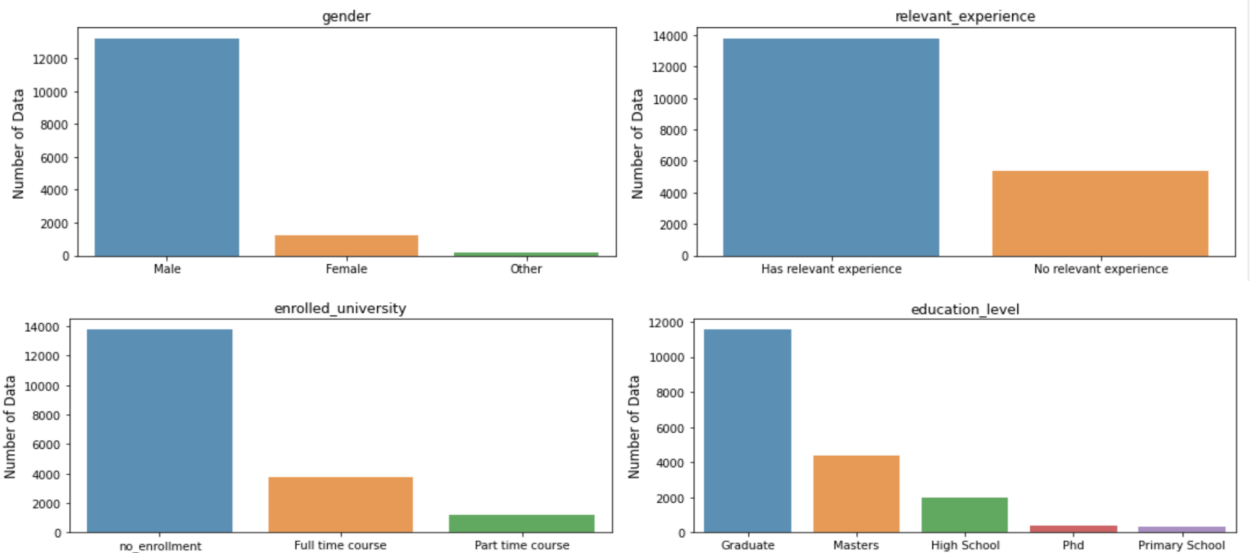
The properties of categorical attributes are summarized using a bar chart, which shows the unique values of each attribute together with their frequency. The "freq" column in Figure 4.4 contains the frequency for the value named as "top," which corresponds to each attribute's most frequent value.

According to the categorical data's descriptive statistics, the majority of trainees are male with degrees in STEM fields. They have more than 20 years of experience, have the relevant experience in data science field, are employed at Pvt Ltd companies, and just took a year off before beginning the training program.

However, some attributes, like “experience”, “company size”, and "last new job", which are expected to be numerical, are categorical due to the special characters ">", "<", and "+" in their values. Thus, those characters should be eliminated.

Table 4.3. Categorical attribute descriptive statistics

	Description	count	unique	top	freq	Unique values
city	City code	19158	123	city_103	4355	city_103, city_40, city_21, city_115, city_162...
gender	Gender of candidate	14650	3	Male	13221	Male, nan, Female, Other
relevant_ experience	Relevant experience of candidate	19158	2	Has relevant experience	13792	Has relevant experience, No relevant experience
enrolled_ university	Type of University course enrolled (if any)	18772	3	no_ enrollment	13817	no_enrollment, Full time course, nan, Part time
education_ level	Education level of candidate	18698	5	Graduate	11598	Graduate, Masters, High School, nan, Phd, Primary School
major_ discipline	Education major discipline of candidate	16345	6	STEM	14492	STEM, Business Degree, nan, Arts, Humanities, ...
experience	Candidate total experience (in years)	19093	22	>20	3286	**>20**, 15, 5, **<1**, 11, 13, 7, 17, 2, 16, 1, 4, 10...
company_ size	Number of employees in current employer's company	13220	8	50-99	3083	nan, **50-99**, **<10**, **10000+**, **5000-9999**, **1000-4999**,...
company_ type	Type of current employer	13018	6	Pvt Ltd	9817	nan, Pvt Ltd, Funded Startup, Early Stage Startup, ...
last_ new_job	Difference in years between previous and current job	18735	6	1	8040	1, **>4**, **never**, 4, 3, 2, nan



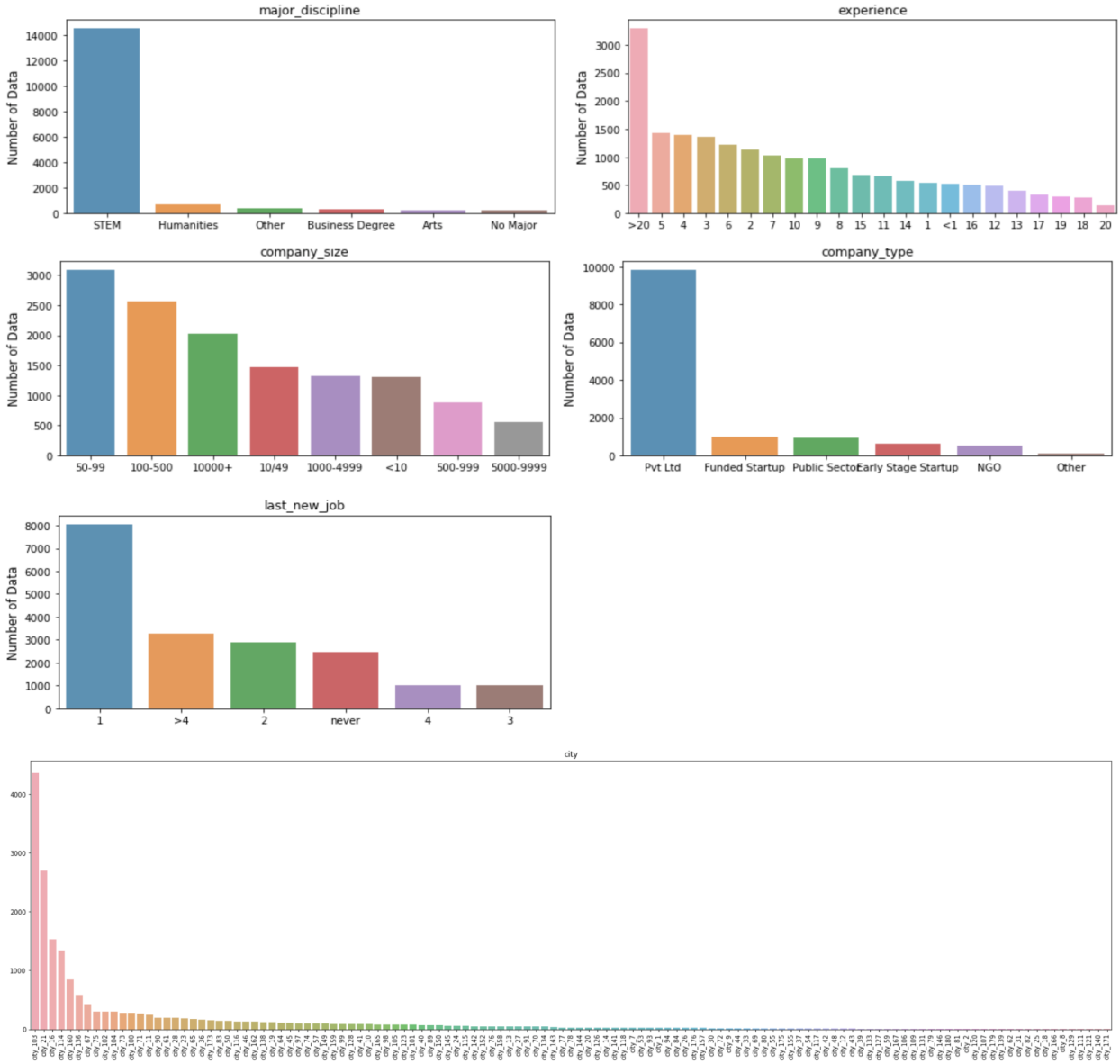


Figure 4.4. Barplots of Categorical attributes

Chapter 5. Pre-processing: Data cleaning

5.1. Dropping unnecessary columns, special characters and correcting data types

The trainee's address (city) attribute has a large number of unique values (Figure 4.4), but it is not really needed in our goal, so it should be ignored and removed from the dataset. Furthermore, since we won't be using “enrollee_id” for any

calculations, it's better to convert it to a string object so that it won't appear whenever we handle numerical numbers.

Table 4.3 and Figure 4.4 show that special characters such as “>” or “<” in some incorrect categorical attributes need to be eliminated so that those attributes can be converted to expected numerical data types. **After the first stage of modification**, the dataset description is shown as beside.

```
RangeIndex: 19158 entries, 0 to 19157
Data columns (total 12 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   enrollee_id             19158 non-null  object
 1   city_development_index  19158 non-null  float64
 2   gender                  14650 non-null  object
 3   relevent_experience     19158 non-null  object
 4   enrolled_university     18772 non-null  object
 5   education_level         18698 non-null  object
 6   major_discipline        16345 non-null  object
 7   experience              19093 non-null  float64
 8   company_type            13018 non-null  object
 9   last_new_job            18735 non-null  float64
 10  training_hours          19158 non-null  int64
 11  target                  19158 non-null  float64
dtypes: float64(4), int64(1), object(7)
```

Figure 5.1. Dataset description (1)

5.2. Handling missing values

The numbers of missing values for all column are shown below:

Table 5.1. Nullity of the dataset

No.	Attribute	Missing Value	No.	Attribute	Missing Value
1	enrollee_id	0	8	major_discipline	2813
2	city	0	9	experience	65
3	city_ development _index	0	10	company_size	5938
4	gender	4508	11	company_type	6140
5	relevent_experience	0	12	last_new_job	423
6	enrolled_university	386	13	training_hours	0
7	education_level	460	14	target	0

It is not reasonable to drop all the missing value as we will lose a large part of our data set. However, for the column that has a small amount of null, for example 65 for "experience", we can simply drop them and deal with other columns instead.

The method that we used to fill up missing values is quite simple. For the columns that fall into categorical type, missing values are replaced with a new class

"Unknown". Whereas in columns that hold numeric information, incomplete data are replaced with the mean of that column.

5.3. Dealing with outliers (numerical variables)

In statistical terms, outliers are observations that lie an abnormal distance from other values of the set. Outliers can highly affect the performance of our machine learning model, causing the model to incorrectly predict the target. Therefore, it is natural to detect and deal with outliers in the cleaning process.

Two common methods for detecting outliers are histogram and boxplot. After defining abnormal values in the data set, we can not simply drop all of them as it will lead to some information loss. A small estimation is conducted to find out to what scale should outliers be ignored.

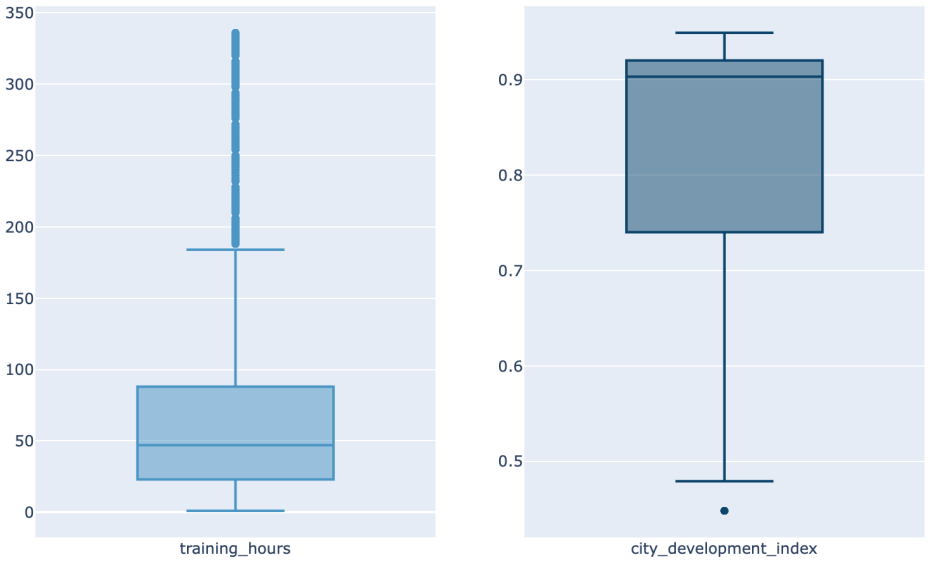


Figure 5.2. Outliers of Numerical attributes

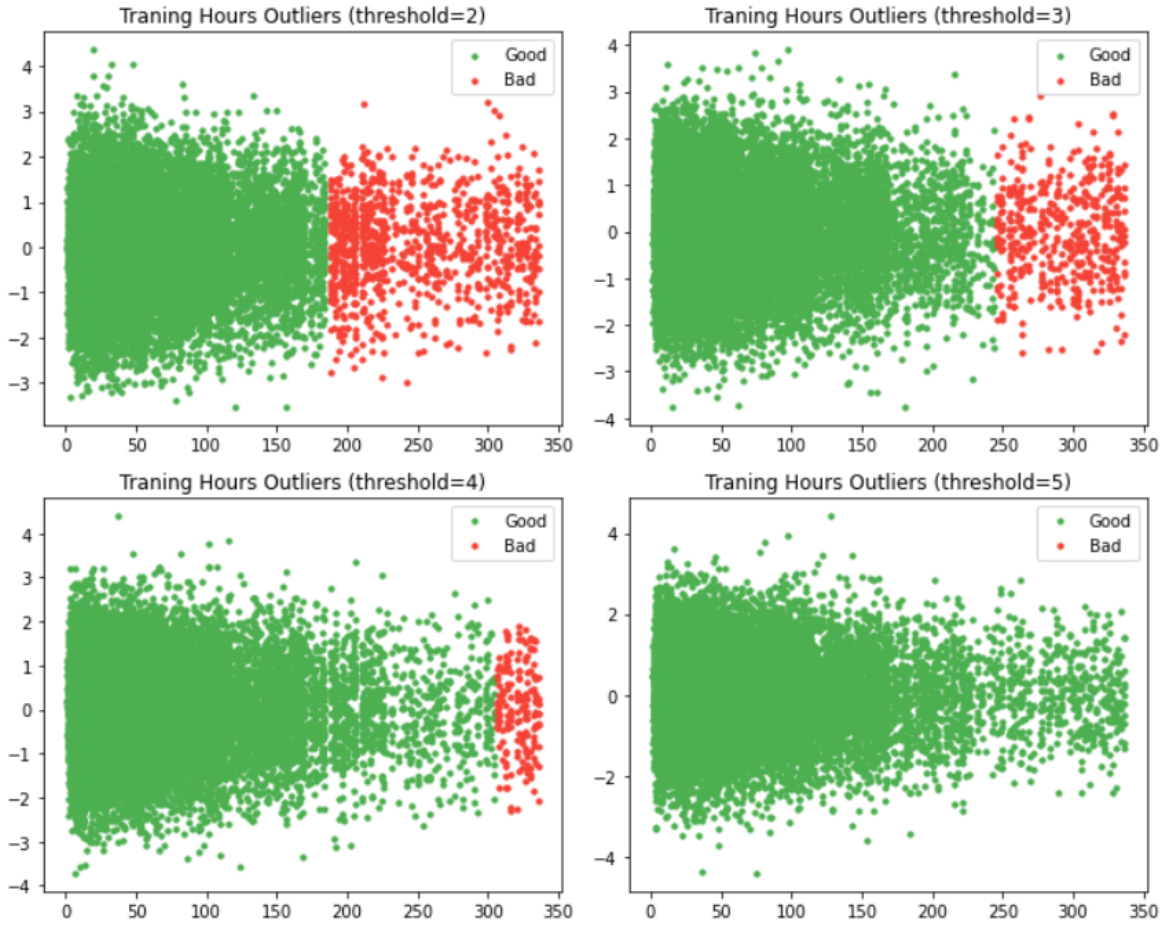


Figure 5.3. Thresholds to remove outliers using z-score

Thus, with a threshold of 3, we can get rid of most of the outliers but still minimize the information loss.

Chapter 6. Descriptive Analysis

Descriptive analysis, which is typically the first kind of data analysis performed on a dataset, is the numerical way to gain insights into the data. It presents a clear picture of what occurred in the past by looking at the historical data and trying to identify specific patterns within the data [6]. The statistics that are summarized in this type of analysis are shown in Univariate analysis part.

6.1. Univariate analysis

Two types of descriptive statistics that should be measured are:

1. Measures of Central Tendency: Mean, Mode, Median.

2. Measures of Dispersion: Min, Max, Range, Quartiles and Inter Quartiles, Variance and Standard Deviation.

	count	mean	mode	std	var	min	25%	50%	75%	max	range	IQR
city_development_index	18644.0	0.829130	0.92	0.123215	0.015182	0.448	0.74	0.903	0.92	0.949	0.501	0.18
experience	18644.0	10.116955	21.00	6.761307	45.715267	0.500	4.00	9.000	16.00	21.000	20.500	12.00
last_new_job	18644.0	2.003915	1.00	1.659969	2.755496	0.000	1.00	1.000	3.00	5.000	5.000	2.00
training_hours	18644.0	59.875724	28.00	49.015650	2402.533983	1.000	23.00	46.000	84.00	244.000	243.000	61.00
target	18644.0	0.249946	0.00	0.432993	0.187483	0.000	0.00	0.000	0.00	1.000	1.000	0.00

Figure 6.1. Numerical attributes descriptive statistics (2)

After some of the outliers in "training_hours" attribute are removed, that attribute has its mean value changing from 65.37 to 59.88, and its standard deviation changing from 60.06 to 49.02, which means that the values are not so far from each other as before. Moreover, the 2 attributes "experience" and "last_new_job" are the new ones appear in numerical attributes after data cleaning.

Those descriptive statistics show that the trainees are those with an average 10 years of experience, 2 years gap from the last job, living in highly-developed cites, and taking about 59 hours of training courses.

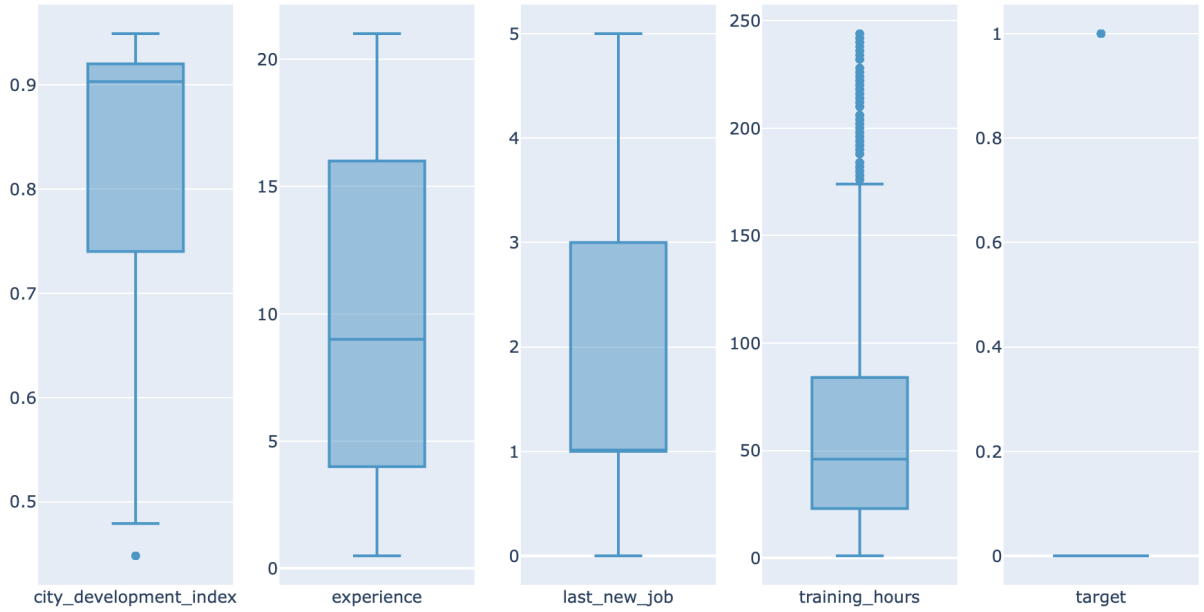


Figure 6.2. Boxplots of Numerical attributes

The statistics of categorical variables consists of the number of their unique values and the frequency of the most frequent ones. The total number of values in each columns has decreased due to the removal of missing values in the "experience" attribute. Furthermore, some categorical attributes have 1 additional unique value named "Unknown" from the imputation step in data cleaning, but it doesn't affect the most frequent value in each column.

	enrollee_id	gender	relevant_experience	enrolled_university	education_level	major_discipline	company_type
count	18644	18644	18644	18644	18644	18644	18644
unique	18644	4	2	4	6	7	7
top	8949	Male	Has relevant experience	no_enrollment	Graduate	STEM	Pvt Ltd
freq	1	12900	13422	13458	11292	14105	9553

Figure 6.3. Categorical attributes descriptive statistics (2)

6.2. Bivariate analysis

6.2.1. Does city development statistics affect the job decision of data scientist trainees?

City development index has a negative correlation with Target, which means that **the trainees have a tendency to stay with the company as the development indices of their cities increase**. When the CDI is in the range [0.6, 0.7], there is a significant change in the percentage of the target decision among candidates.

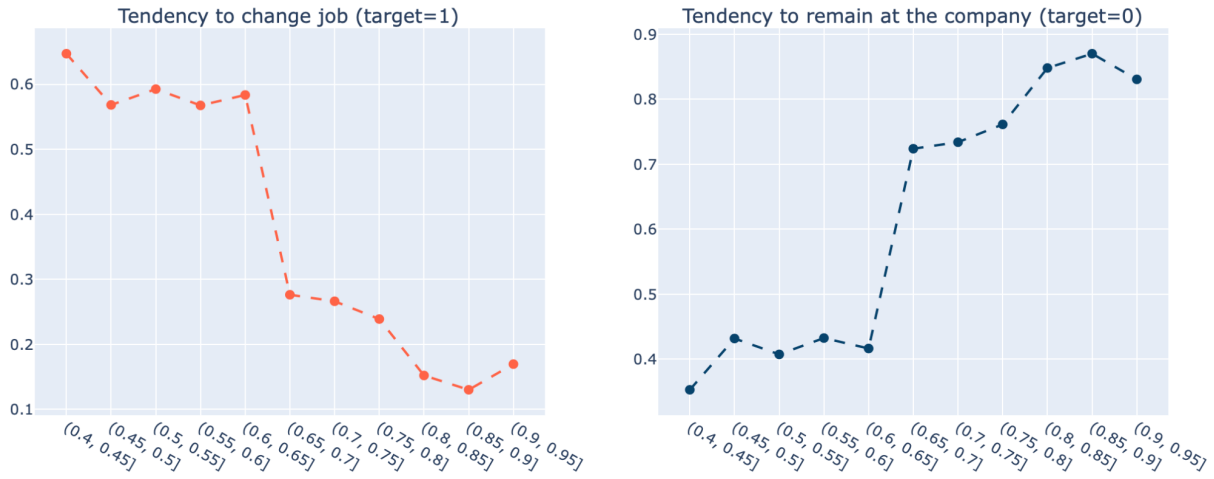


Figure 6.4. City development index vs Target

6.2.2. Which gender prefers to stay with the company after training courses?

The number of male candidates joined in this training program is obviously larger than the other genders, but the turnover rate of them has the smallest value (22.9%). In other words, in comparison with other genders, **male trainees are most likely to stay with the company**.

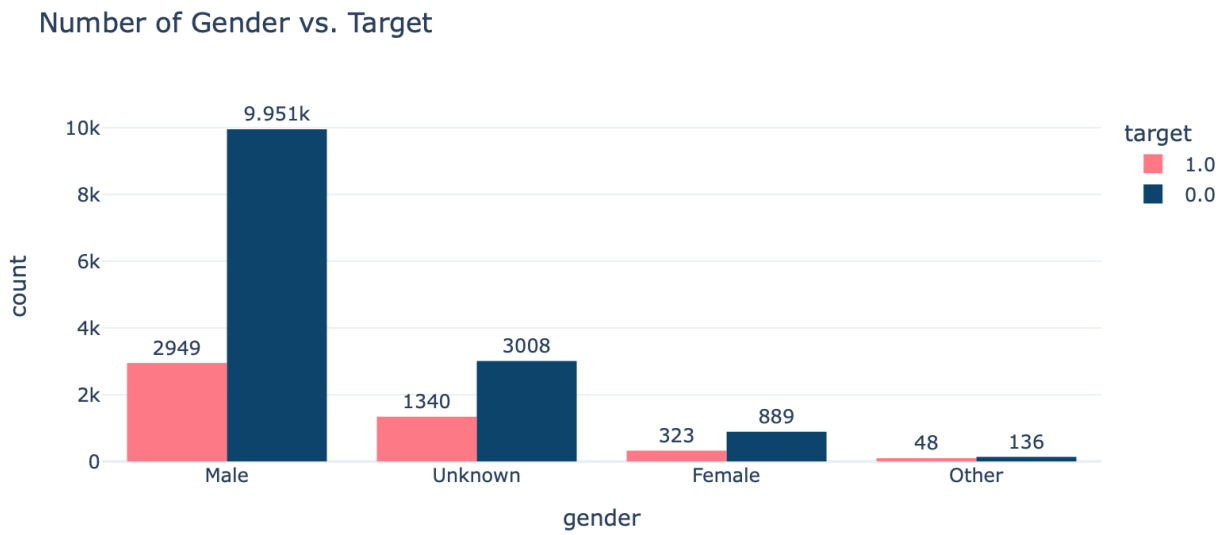


Figure 6.5. Number of Genders in the training program

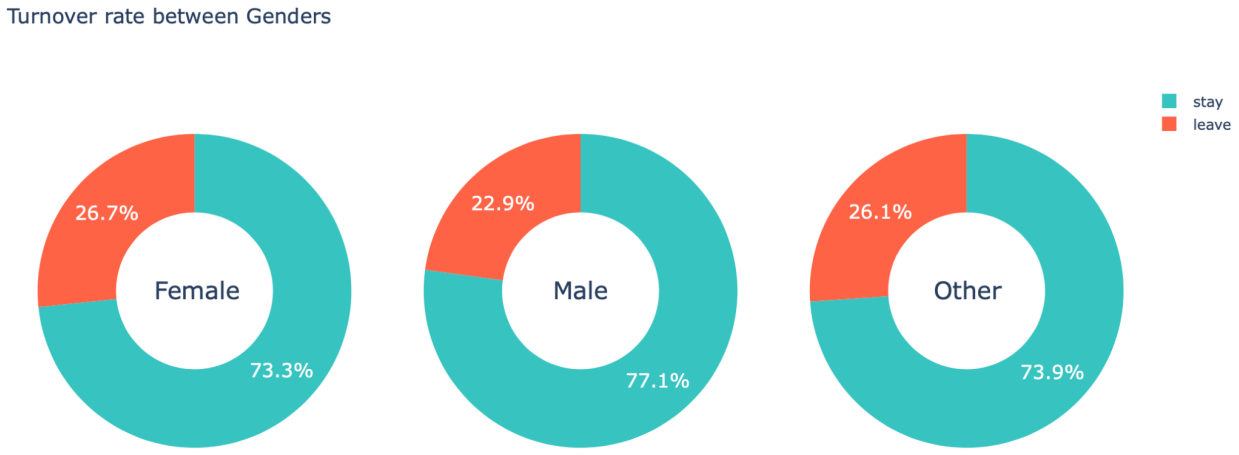


Figure 6.6. Turnover rate between Genders (%)

6.2.3. Is relevant experience in the data science field necessary to join in the program?

Although the majority of trainees have prior relevant experience in the field, those who don't still have an opportunity to enroll in the program. According to this program's data, the higher the expertise, the lower the turnover rate. In other words, **while relevant experience is not required, if the candidates don't have it, they have a greater tendency to leave**.

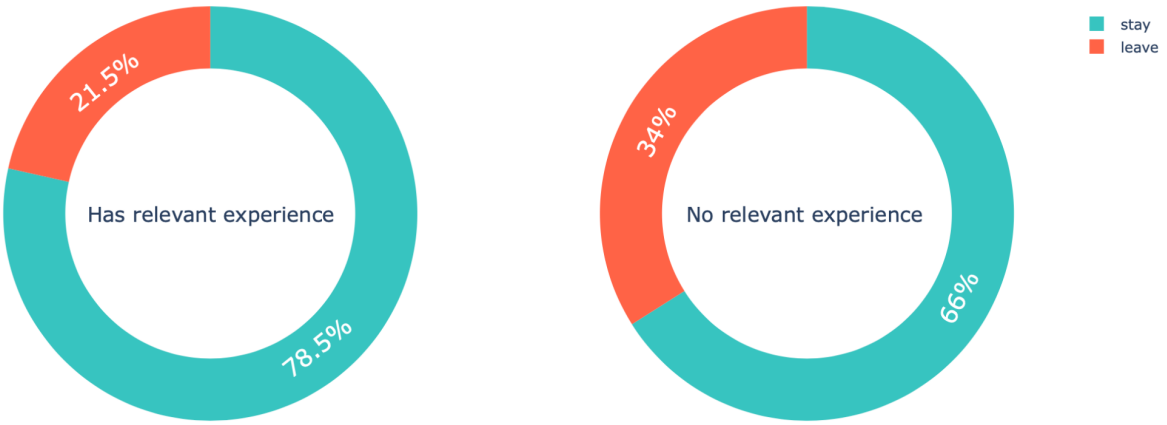


Figure 6.7. Turnover rate vs. Relevant experience (%)

6.2.4. Can full-time students join in the training program?

Candidates enrolling in a full time course have to spend most of their time studying at school, while those who are not joining in any courses at a university have more time to go to work. Therefore, the percentage of staying with the company is higher for those who have finished all their university courses. Furthermore, **the trainees with relevant experience in data science field and without any university courses enrollment are most possible to stay** with the company and work as a data scientist.

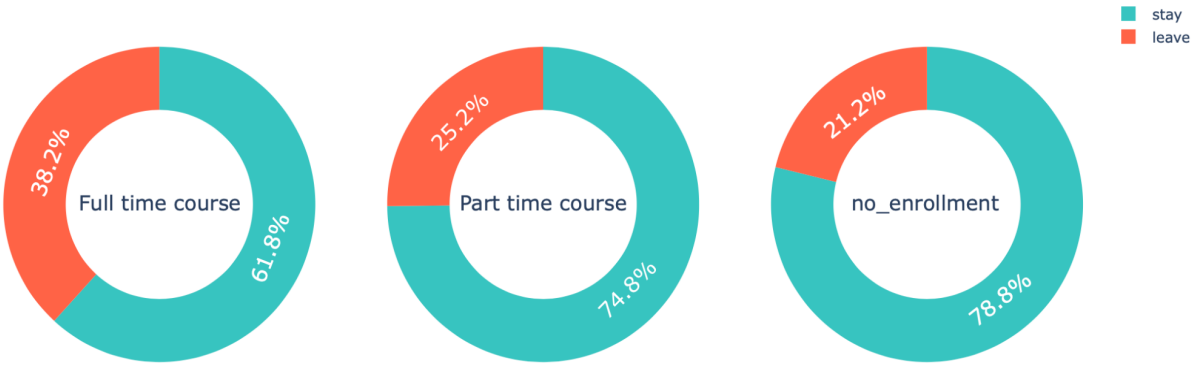


Figure 6.8. Turnover rate vs. University enrolled courses (%)

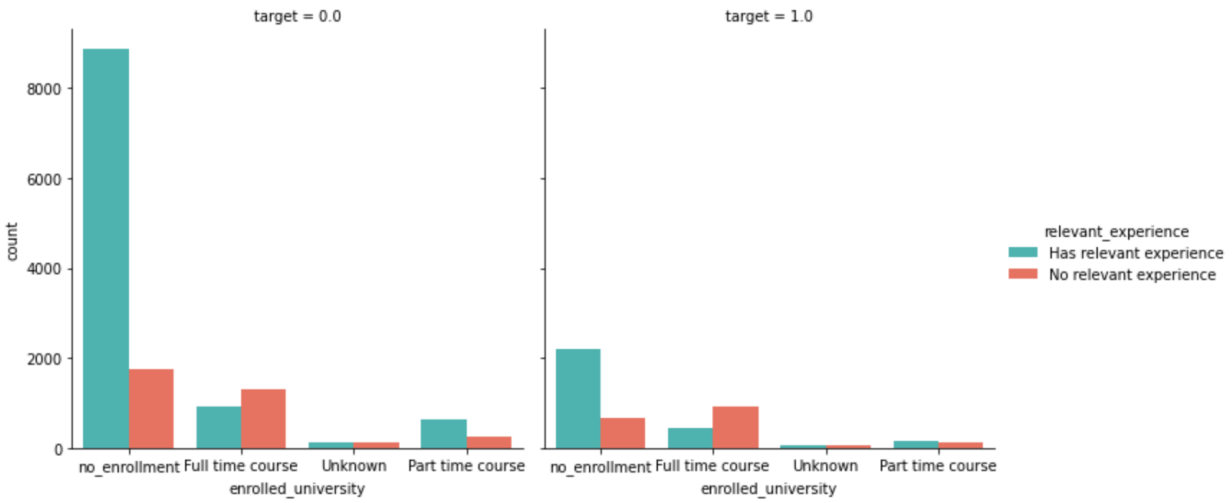


Figure 6.9. Turnover rate according to relevant experience & types of university courses enrolled

6.2.5. Which educational level wants to look for a new job the most?

General college graduates have a higher rate of turnover than those with extremely low or high levels of education. It appears that there is a high proportion of anxiety about making the right career decision among fresher graduates, as they are looking for appropriate employment.

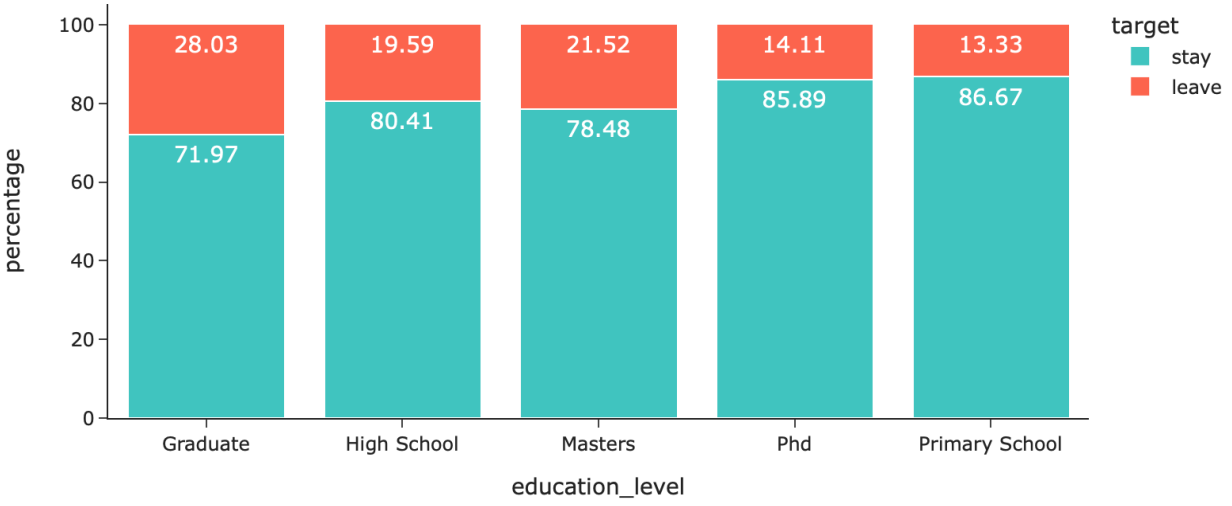


Figure 6.10. Turnover rate vs. Education level (%)

6.2.6. Which company type offers the trainees the best environment to progress their job?

Apart from the "Other" company types that we don't know, early stage start-ups have the highest turnover rate, while funded start-ups have the lowest turnover rate. It indicates that **funded startup companies have much** better communication

between employees, **better working environment**, and more opportunities for the trainees.

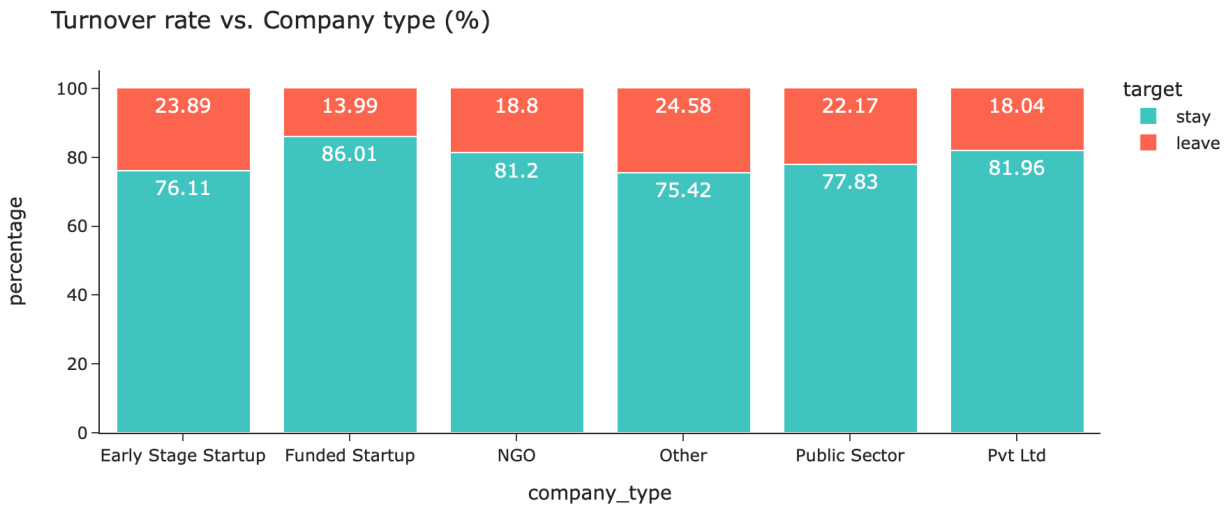


Figure 6.11. Turnover rate vs. Company type (%)

6.2.7. Which random variables in the dataset are the key factors that determine whether a trainee will stay with the company?

After the data is cleaned, it has 2 new numerical attributes which are “experience” and “last_new_job”, both of which have the negative correlation with the final decision of the trainees. Besides, “experience” attribute has positive correlations with “last_new_job” and “city_development_index”.

Figure 6.12. Correlation between Numerical attributes (2)

People with a high total experience years also tend to spend more time before beginning a new job. Besides, the higher the city degree of development, the more the experienced candidates who chose to stay with the company.

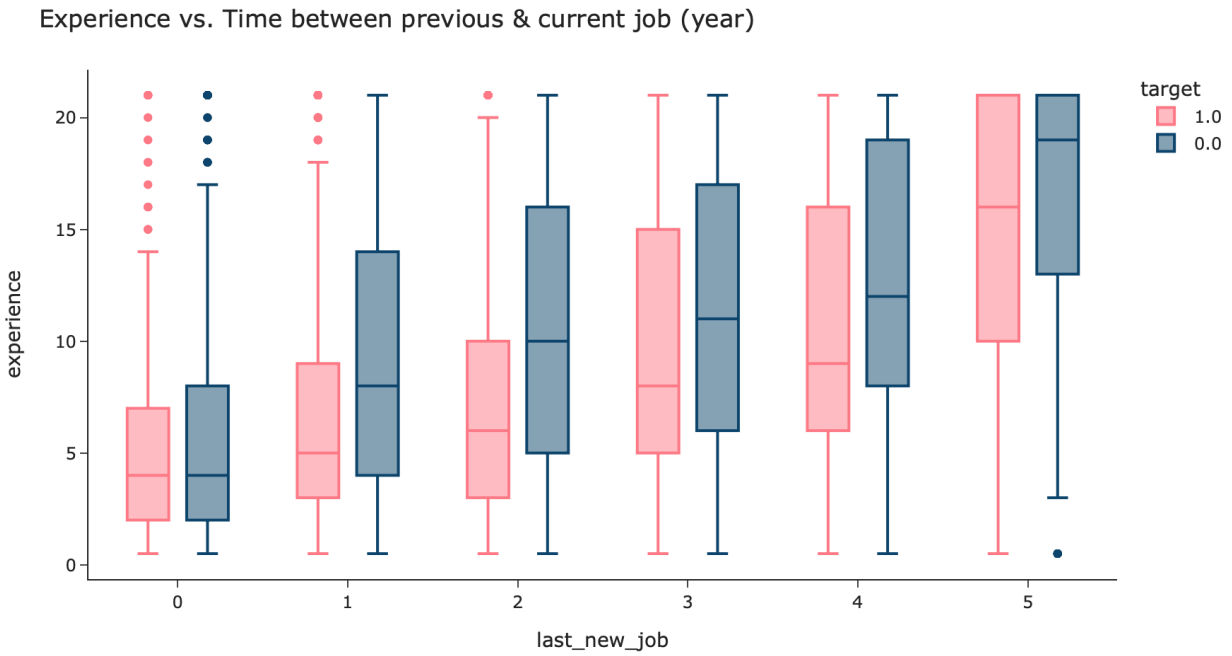


Figure 6.13. Experience vs. Time between previous & current job (year) correlation

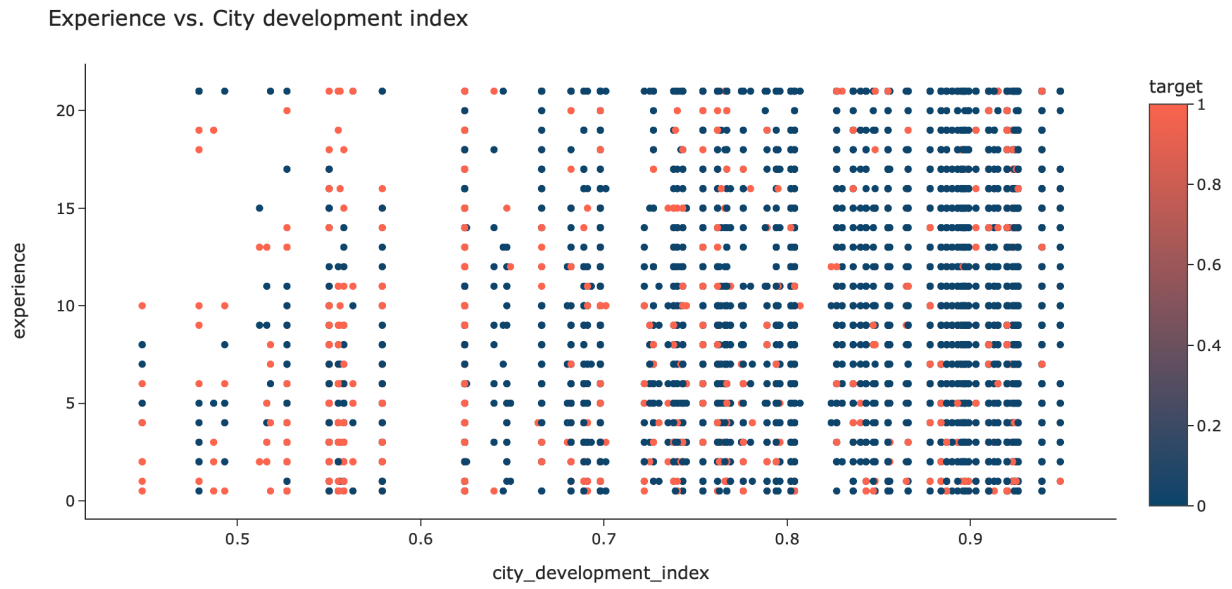


Figure 6.14. Experience vs. City development index

Chapter 7. Pre-processing: Data transformation

7.1. Encoding categorical data

Machine Learning algorithms mostly work with numerical data. Although those data points may have other meanings apart from mathematical ones, they still need to be represented in terms of a number so that the model can recognise and deal with them. Unluckily, in our data set, most of the columns fall in the categorical type, which contain strings. Therefore, we need to transform them into numerical terms before the training model process could be done.

Categorical data are divided into 2 small types: nominal and ordinal. Unlike ordinal columns, nominal ones do not have the order between the data points. For example, if we encode "Male" and "Female" by 1 and 0, the algorithm is likely to see 1 as larger than 0, which means Male is greater than Female. To prevent the model from that kind of misunderstanding, different approaches have to be used to deal with nominal and ordinal data.

For attributes that hold a defined order for data points, Ordinal encoding is straightforward. All data points are encoded as a numeric value according to their order. Data with higher significance are represented by bigger numbers. Whereas for nominal columns such as "gender" or "company type", our method is Dummies encoder, which requires creating new columns, marking the presence of data value as 1 and its absence as 0. In that way, we can make sure that the model understands exactly the meaning of data even though they are in numeric terms.

	enrollee_id	city_development_index	gender_Male	gender_Unknown	gender_Female	gender_Other	relevent_experience	enrolled_university	education_level	major_discipline_STEM	...
0	8949	0.920	1	0	0	0	1	1	3	1	...
1	29725	0.776	1	0	0	0	0	1	3	1	...
2	11561	0.624	0	1	0	0	0	3	3	1	...
3	33241	0.789	0	1	0	0	0	0	3	0	...
4	666	0.767	1	0	0	0	1	1	4	1	...

5 rows × 27 columns

Figure 7.1. Encoded train dataset

7.2. Features Scaling

Different variables with different ranges can highly affect the performance of the model, especially for KNN as it is a distance-based algorithm. To overcome this issue, all of the variables are scaled into only one standard. There are a variety of methods to scale a data set. Among them, **Standard Scaler** has been chosen as it removes the mean and scales all features to unit variance.

At the end of the preprocessing process, the final csv files are exported. In the next part of the project, the modeling, will use these files for the sake of simplification and data integrity.

	count	mean	std	min	25%	50%	75%	max
city_development_index	18644.000000	0.000000	1.000027	-3.093295	-0.723392	0.599534	0.737508	0.972875
gender_Male	18644.000000	-0.000000	1.000027	-1.498607	-1.498607	0.667287	0.667287	0.667287
gender_Unknown	18644.000000	0.000000	1.000027	-0.551490	-0.551490	-0.551490	-0.551490	1.813270
gender_Female	18644.000000	0.000000	1.000027	-0.263680	-0.263680	-0.263680	-0.263680	3.792471
gender_Other	18644.000000	0.000000	1.000027	-0.099837	-0.099837	-0.099837	-0.099837	10.016291
relevant_experience	18644.000000	-0.000000	1.000027	-1.603209	-1.603209	0.623749	0.623749	0.623749
enrolled_university	18644.000000	-0.000000	1.000027	-1.742069	-0.527169	-0.527169	0.687731	1.902631
education_level	18644.000000	-0.000000	1.000027	-3.684082	-0.075043	-0.075043	-0.075043	2.330983
major_discipline_STEM	18644.000000	0.000000	1.000027	-1.762814	0.567275	0.567275	0.567275	0.567275
major_discipline_Business Degree	18644.000000	0.000000	1.000027	-0.131939	-0.131939	-0.131939	-0.131939	7.579257
major_discipline_Unknown	18644.000000	0.000000	1.000027	-0.414094	-0.414094	-0.414094	-0.414094	2.414913
major_discipline_Arts	18644.000000	0.000000	1.000027	-0.116346	-0.116346	-0.116346	-0.116346	8.595086
major_discipline_Humanities	18644.000000	-0.000000	1.000027	-0.190666	-0.190666	-0.190666	-0.190666	5.244773
major_discipline_No Major	18644.000000	0.000000	1.000027	-0.108771	-0.108771	-0.108771	-0.108771	9.193636
major_discipline_Other	18644.000000	-0.000000	1.000027	-0.142293	-0.142293	-0.142293	-0.142293	7.027744
experience	18644.000000	-0.000000	1.000027	-1.422390	-0.904724	-0.165202	0.870128	1.609650
company_type_Unknown	18644.000000	0.000000	1.000027	-0.686918	-0.686918	-0.686918	1.455778	1.455778
company_type_Pvt Ltd	18644.000000	-0.000000	1.000027	-1.025095	-1.025095	0.975519	0.975519	0.975519
company_type_Funded Startup	18644.000000	-0.000000	1.000027	-0.234526	-0.234526	-0.234526	-0.234526	4.263927
company_type_Early Stage Startup	18644.000000	0.000000	1.000027	-0.180142	-0.180142	-0.180142	-0.180142	5.551189
company_type_Other	18644.000000	-0.000000	1.000027	-0.079809	-0.079809	-0.079809	-0.079809	12.529964
company_type_Public Sector	18644.000000	-0.000000	1.000027	-0.230166	-0.230166	-0.230166	-0.230166	4.344690
company_type_NGO	18644.000000	0.000000	1.000027	-0.166004	-0.166004	-0.166004	-0.166004	6.023952
last_new_job	18644.000000	0.000000	1.000027	-1.207233	-0.604796	-0.604796	0.600078	1.804953

Figure 7.2. Scaled train dataset

Chapter 8. Modeling (Predictive analysis)

8.1. K - Nearest Neighbors

When it comes to Machine Learning, the first algorithm that comes to use is the KNN (K - Nearest Neighbors) as it is one of the simplest algorithms to learn. KNN is a supervised machine learning method that can be used for classification and regression problems, which is exactly our purpose in this project.

The most important parameter that affects the accuracy of this algorithm is the number of n_neighbors. To make sure this parameter is chosen wisely, we have

conducted an investigation of how accuracy and error vary with different n_neighbors.

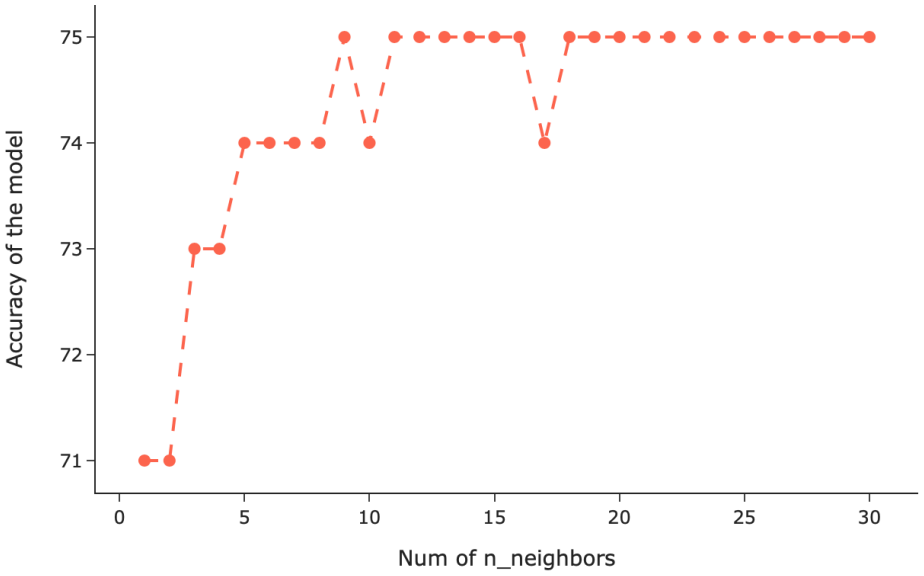


Figure 8.1. KNN accuracy vs. Number of neighbors

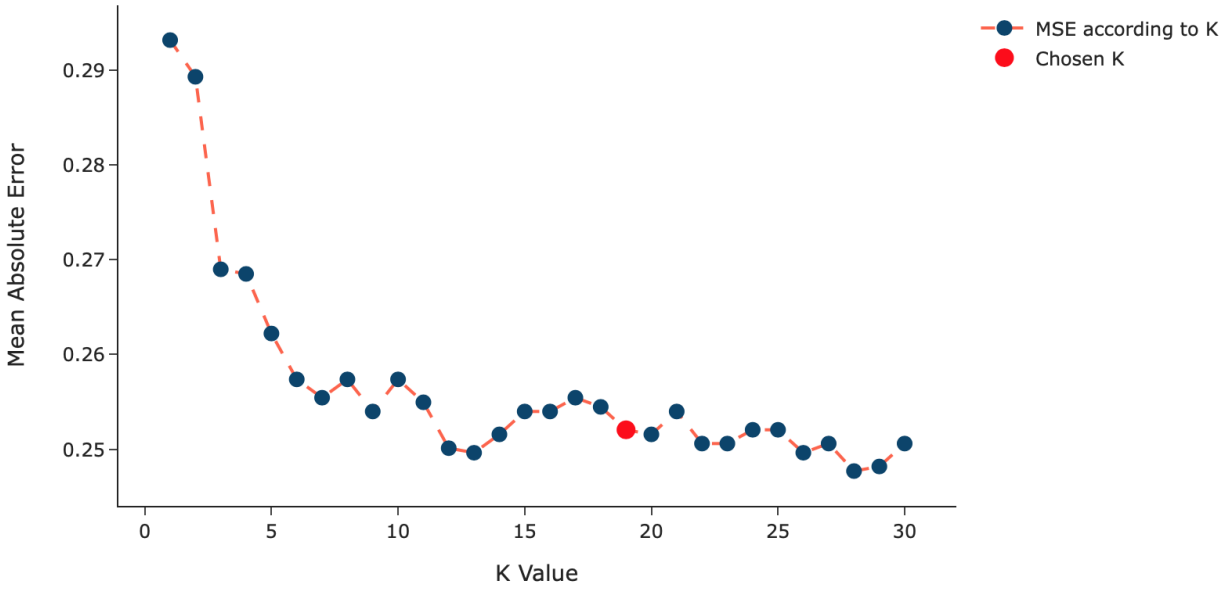


Figure 8.2. KNN Mean absolute error (MAE) vs. Number of neighbors

The visualization indicates that **K = 19** gives the highest performance. Therefore, we will accept this number to train the dataset and make predictions. After K is chosen, it is necessary to look further into the model's performance by evaluating the accuracy and classification report. Whereas K = 20 gives the accuracy score of 75%, Python library sklearn.metrics provides a method to conduct the classification report which is shown below.

	precision	recall	f1-score	support
0.0	0.80	0.87	0.83	1504
1.0	0.55	0.42	0.48	563
accuracy			0.75	2067
macro avg	0.67	0.65	0.65	2067
weighted avg	0.73	0.75	0.74	2067

Figure 8.3. KNN Classification report

8.2. Logistic Regression

Logistic Regression is a supervised machine learning algorithm that can be used to model the probability of a certain class or event. It is used when the data is linearly separated and the outcome is binary. Unlike KNN, Logistic Regression is mostly used for binary classification problems, when the labels are just “Yes” and “No” [8]. Compared to KNN, Logistic Regression is much faster. Moreover, it allows us to dig in the coefficients of the variables to see which of them affects the model more deeply.

There are some parameters which are defined in the model implementation in order to optimize the outcomes.

- random_state: this parameter makes sure we do not receive a different value every time we run the model.
- solver = "liblinear": stands for Library for Large Linear Classification, is an algorithm to be used in the optimization problem.
- penalty = “L2”: this parameter adds a penalty to the coefficients to prevent the model from overfitting or underfitting. Ridge regularization or L2 means we are taking squares of the weights.
- max_iter: refers to the maximum number of iterations taken for the solvers to converge. In this case it is set to 5000.

There is a slight difference between the classification reports of the 2 models. However, the accuracy that Logistic Regression gives, which is 75%, is still the same.

	precision	recall	f1-score	support
0.0	0.77	0.93	0.84	1504
1.0	0.58	0.27	0.37	563
accuracy			0.75	2067
macro avg	0.68	0.60	0.60	2067
weighted avg	0.72	0.75	0.71	2067

Figure 8.4. Logistic Regression Classification report

Also, as mentioned above, the model gives us a list of coefficients of every variable. In summary, despite having the same accuracy as KNN, LR provides us with way more benefits and takes less time to run.

	Coefficient		Coefficient
city_development_index	-0.706075	major_discipline_No Major	0.040736
gender_Male	-0.019762	major_discipline_Other	0.038703
gender_Unknown	-0.001907	experience	-0.160546
gender_Female	0.031517	company_type_Unknown	0.332778
gender_Other	0.021854	company_type_Pvt Ltd	-0.211757
relevent_experience	-0.118097	company_type_Funded Startup	-0.141612
enrolled_university	0.103132	company_type_Early Stage Startup	-0.052010
education_level	-0.128343	company_type_Other	0.010360
major_discipline_STEM	0.163610	company_type_Public Sector	-0.002314
major_discipline_Business Degree	0.074524	company_type_NGO	-0.057090
major_discipline_Unknown	-0.311564	last_new_job	0.101062
major_discipline_Arts	0.055694	training_hours	-0.036360
major_discipline_Humanities	0.076537		

Figure 8.5. Logistic coefficient

8.3. Perform Logistic Regression and Decision Tree Classifier with PCA

8.3.1 Dimensionality Reduction (PCA)

Our dataset contains so many random variables and it takes time for the model to handle all of them. In real life, it takes even more time and money to predict the target for the dataset whose dimensionality is big like this. One feasible solution for

this problem is that the dataset's scale has to be reduced to an acceptable volume. PCA, stands for Principal Components Analysis, is an algorithm for such a task.

Specifically, PCA is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large data set of variables into a smaller one that still contains most of the information from the original one [9]. This project's data set consists of more than 20 variables. Our purpose is to get a smaller set with a reasonable number of principal components from which the model can still give high performance.

8.3.2 Logistic Regression with PCA

In order to choose the best number of components, the same algorithm has to be run several times with different n_components. The algorithm that is chosen is Logistic Regression, later there will be a comparison between the model's performance before and after applying PCA. To investigate how the model's effectiveness varies with n_components, we have two line charts visualizing those relationships.

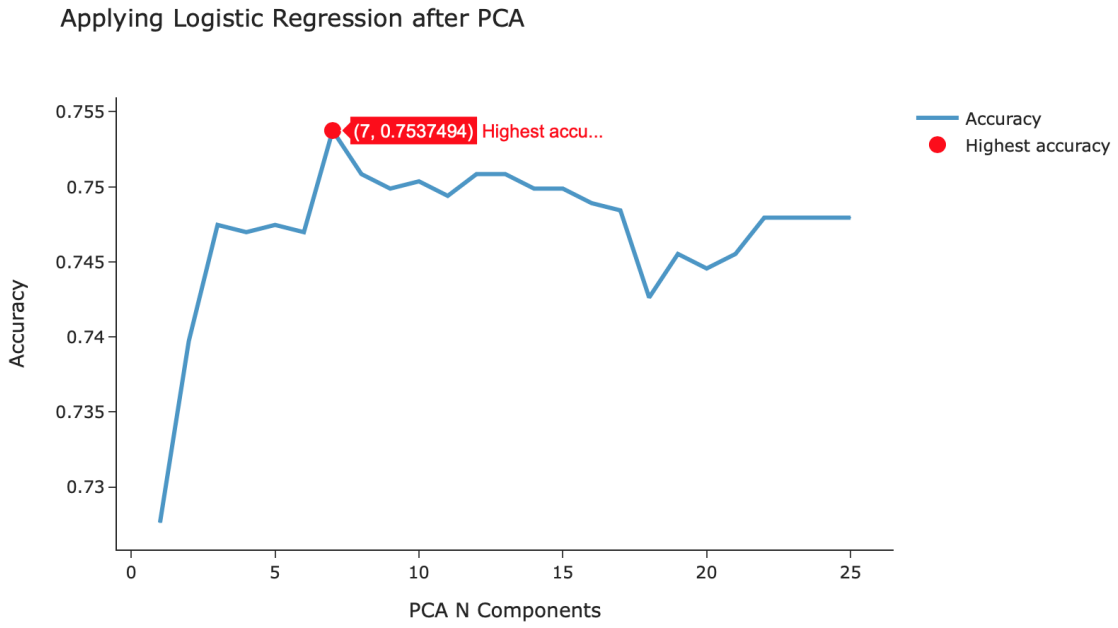


Figure 8.6. PCA accuracy vs. Number of components

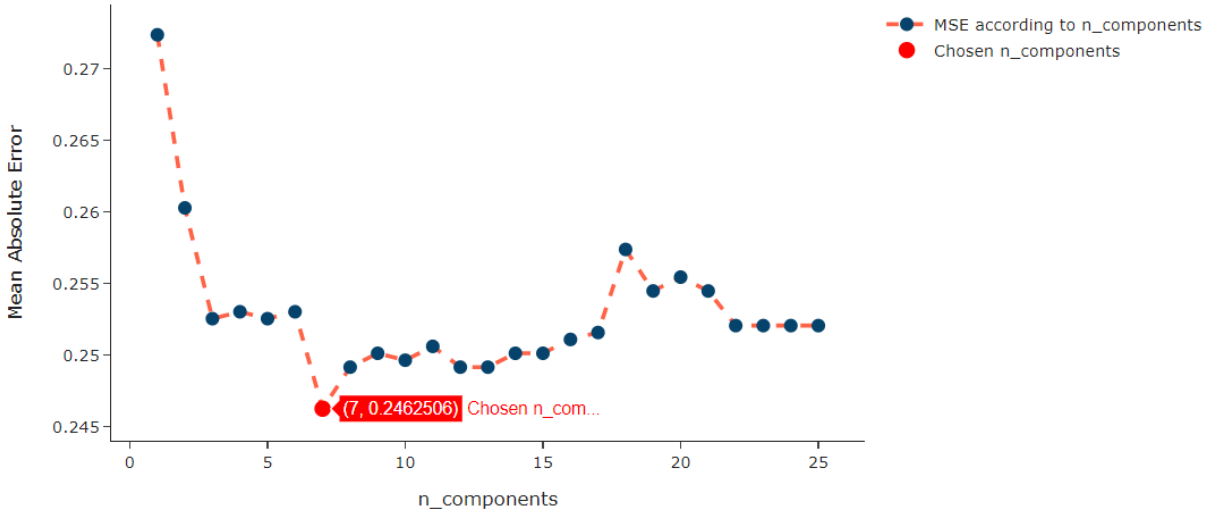


Figure 8.7. PCA Mean absolute error (MAE) vs. Number of components

Two line graphs illustrate that **n_components = 7** gives the highest score and lowest Mean Absolute Error. After deciding which n_component is the best for this model, we have to rerun the model and compare that with its before-PCA version. Two crucial questions are: "What is the accuracy of the model?" and "How long does it take the model to handle?".

```
Accuracy of Logistic Regression: 75.0 %
This took 0.08 seconds
Accuracy of Logistic Regression (with PCA: n_components = 7): 75.0 %
This took 0.03 seconds
```

Figure 8.8. Runtime comparison of Logistic Regression model before and after PCA

Obviously, Logistic Regression still provides the same accuracy. However the size of the data set is optimized and the runtime is approximately 2 times faster.

8.3. Decision Tree Classifier

Decision Tree falls under the type of supervised machine learning algorithm. It can be used for both classification and regression problems. Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented by the internal node of the tree [10].

To make an instance of the Tree model, we have set the max_depth, which means limit to stop further splitting of tree nodes, to be 5. The visualization of the tree indicates that **the most significant attribute to determine the label is *city_developement_index***. Other attributes' influence on the output decreases as we go down the tree.

Chapter 9. Conclusion

9.1. Summary of findings

The project has answered all of the questions proposed from the first time we received the data. Interaction visualization has been made to clarify the findings and insights. Besides, 3 different types of model with the accuracy of above 70% have been built to make predictions based on train and test set. Generally, every requirement is satisfied. The findings of this project are summarized below.

- The majority of trainees are male with degrees in STEM fields. They have more than 20 years of experience, have the relevant experience in data science field, are employed at Pvt Ltd companies, and just took a year off before beginning the training program.
- The trainees are those with an average 10 years of experience, 2 years gap from the last job, living in highly-developed cites, and taking about 59 hours of training courses.
- The trainees have a tendency to stay with the company as the development indices of their cities increase.
- In comparison with other genders, male trainees are most likely to stay with the company.
- Although relevant experience is not required, if the candidates don't have it, they have a greater tendency to leave.
- The trainees with relevant experience in data science field and without any university courses enrollment are most possible to stay with the company and work as a data scientist.
- General college graduates have a higher rate of turnover than those with extremely low or high levels of education.
- Funded startup companies offer much better communication between employees, better working environment, and more opportunities for the trainees.
- "city_development_index", "experience" and "last_new_job", with negative correlation with the final decision of the trainees, are the key features that affect the final decision of the trainees.

9.2. Limitations

Generally, every goal that we set on this project is achieved. However, there are some limitations that can be improved in the future.

- Standard error of the models is still high which can be optimized if we have more time and effort for study.
- This project involves the use of many models. However, we were not able to find the best one for this dataset.

9.3. Future plans

Some improvements can be made if this project is developed in the future:

- Apply more types of model to bring the best performance
- Optimize all of the already-used methods, modify their parameters to enhance the exactness.
- Try preprocessing the data and running the model on other platform apart from Google Colab, such as Tableau, or try involving other distributed computing frameworks (PySpark) to lower the run time.
- Use the models to predict real-life data.

Chapter 10. Reference list

[1] Lalwani, P. (2021, March 11). *What Is HR Analytics? Definition, Importance, Key Metrics, Data Requirements, and Implementation |*. Spiceworks. https://www.spiceworks.com/hr/hr-analytics/articles/what-is-hr-analytics/

[2] Roos, H. (2021, July 18). *Human Resource analytics — Can we predict Employee Turnover with caret in R?* Towards Data Science. https://towardsdatascience.com/human-resource-analytics-can-we-predict-employee-turnover-with-caret-in-r-3d871217e708

[3] *How to Do a New Hire Turnover Calculation*. (n.d.). Resources. https://resources.skillwork.com/new-hire-turnover-calculation

[4] Isom, M. (n.d.). *How to Calculate and Reduce Your New Hire Turnover Rate*. Cangrade. https://www.cangrade.com/blog/talent-management/how-to-calculate-and-reduce-your-new-hire-turnover-rate/

[5] *HR Analytics: Job Change of Data Scientists*. (n.d.). Kaggle. https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists

[6] Pedamkar, P. (n.d.). *Types of Data Analysis | Four Popular Data Analysis Methodologies*. eduCBA. https://www.educba.com/types-of-data-analysis/

[7] City Development Index for Armenian Cities. (n.d.). Encyclopedia.pub. https://encyclopedia.pub/item/revision/f96a983dc8d8507299201433f8c20a23

[8] Varghese, D. (n.d.). *Comparative Study on Classic Machine learning Algorithms | by Danny Varghese*. Towards Data Science. https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222

[9] Jaadi, Z. (2022, August 8). *Principal Component Analysis (PCA) Explained*. Built In. https://builtin.com/data-science/step-step-explanation-principal-component-analysis

[10] *Decision Tree Introduction with example*. (2022, November 10). GeeksforGeeks. https://www.geeksforgeeks.org/decision-tree-introduction-example/

A. List of Figures and Tables

B. Project schedule

Week no.	Date	Task name	Note
2	05/09/2022 - 11/09/2022	- Select a **field** to choose the topic (Medical, Education, Business, etc.) - Search for usable datasets of each fields	
3	12/09/2022 - 18/09/2022	- **Collect data**: Find an appropriate dataset - Decide the **topic** for the project - Develop the project **questions** (Initial thoughts)	
4	19/09/2022 - 25/09/2022	- **Explain** the topic information - Make a project implementation **plan**	
5	26/09/2022 - 02/10/2022	- **PROJECT PROPOSAL**: The first ideas for the project - **Explore data**	
6	03/10/2022 - 09/10/2022	- **Pre-processing**: Data cleaning	
7	10/10/2022 - 16/10/2022	- **Pre-processing** - **Data analysis**: Identify the type of analysis performed	
8	17/10/2022 - 23/10/2022	- **Data analysis** - Apply statistical methods to calculate some descriptive statistics	
9	24/10/2022 - 30/10/2022	- **Data analysis** - **Pre-processing**: Data transformation & reduction - Extract features	
10	31/10/2022 - 06/11/2022	- Build model - **Train model**	
11	07/11/2022 - 13/11/2022	**- Test model** - Evaluate model	
12	14/11/2022 - 20/11/2022	**- Data visualization:** Briefly explain the key take-away based on the data visualization chart	
13	21/11/2022 - 27/11/2022	**PROJECT FINAL REPORT**	
14	28/11/2022 - 02/12/2022	**The final project submission** Files + Presentation	