Springboard - DSC

Capstone Project II

Detecting Potential Candidates Who are Looking for a New Job

Final Report

Yang Liu Kunz

Nikoloz Skhirtladze

Data Science Career Track

November 2020

Table of Contents

[1 Introduction 3](#_Toc56198811)

[1.1 Objective 3](#_Toc56198812)

[1.2 Significance 3](#_Toc56198813)

[2 Dataset 3](#_Toc56198814)

[2.1 Data description 3](#_Toc56198815)

[2.2 Dataset Characteristic: Class Imbalance and Feature Space Overlap 4](#_Toc56198816)

[3 Package Introduction 6](#_Toc56198817)

[4 Data Wrangling 6](#_Toc56198818)

[5 Exploratory Data Analysis 9](#_Toc56198819)

# 1 Introduction

Company XYZ is a training institute which conducts training for analytics/ data science. They want to expand their business to manpower recruitment (data science only) by connecting their enrollees with their clients who are looking to hire employees working in the same domain. Before that, they want to know among the large number of signups, which of these candidates are looking for a new employment. To understand the factors that lead the enrollees to look for a job change, Company XYZ wants to build a model based on the current credentials/demographic/experience data they collected from the enrollee to predict the probability of them to look for a new job. By identifying the target enrollee (potential candidate) as many as possible from all enrollee registered on the training platform, the recruiting team in company XYZ could take further approach to their targets more efficiently and effectively.

## 

## 1.1 Objective

The objectives of this project are:

* To explore and analyze enrollee data for XYZ training institute
* To identify the key features that lead the enrollee to look for new employment.
* To develop machine learning models that predict the probability of enrollee looking for a new job
* To identify the final model that captures the most target enrollees within the top 20% of the test dataset ordered by their predicted probability

This report is divided into the following sections:

* Section 2: Dataset description
* Section 3: Data cleaning and wrangling process
* Section 4: Data visualization and analysis
* Section 5: Machine learning and model evaluation
* Section 6: Hyper parameters tuning
* Section 7: Future work

The programming codes used for this report can be found [here](https://github.com/yoyo6022/Springboard_Capstone/tree/master/notebooks).

## 1.2 Significance

By thoroughly explore the dataset, we will identify the important features that affect the enrollee’s decision of career change. We will also develop machine learning models that can be used by the recruiting team of company XYZ to filter out the potential candidates from the user data base and approach them with better efficiency and accuracy.

# 2 Dataset

## 2.1 Data description

The datasets are sourced from the website [kaggle](https://www.kaggle.com/), 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 [train.csv](https://www.kaggle.com/aswathrao/hr-analysis?select=train.csv) used in this project, was collected on August 12, 2020. It consists of 18,359 rows and 14 columns, 4 of them are numerical columns and 10 of them are categorical columns. Each row contains credentials/demographic/experience data for each unique enrollee. A description of each of the 14 columns is provided in *Table 1.1*

*Table 1.1 Description of dataset*

|  |  |  |  |
| --- | --- | --- | --- |
| No. | Variable Name | Variable Description | Data Summary |
| 1 | enrollee\_id | unique ID for enrollees | integer, 18359 unique values |
| 2 | city | city code | object, 123 unique values |
| 3 | city\_development\_index | Developement index of the city (scaled) | continuous, 93 unique values |
| 4 | gender | gender | object, 3 unique values |
| 5 | relevent\_experience | relevant experience in analytics, data science | object, 2 unique values |
| 6 | enrolled\_university | type of University course enrolled if any | object, 3 unique values |
| 7 | education\_level | education level | object, 5 unique values |
| 8 | major\_discipline | major discipline | object, 6 unique values |
| 9 | experience | total working experience in years | object, 22 unique values |
| 10 | company\_size | number of employees in current employer's company | object, 8 unique values |
| 11 | company\_type | type of current employer | object, 6 unique values |
| 12 | last\_new\_job | difference in years between previous job and current job | object, 6 unique values |
| 13 | training\_hours | training hours completed | integer, 241 unique values |
| 14 | target | looking for job change or not | integer; 0=not looking for job change, 1=Looking for a job change |

## 2.2 Dataset Characteristic: Class Imbalance and Feature Space Overlap

An inherent characteristic of this dataset is its class imbalance. In the original dataset [train.csv](https://www.kaggle.com/aswathrao/hr-analysis?select=train.csv), there is 15,934 negative classes and 2425 positive classes in column ‘target’.

Value 0 represents the negative class which means the enrollee is not open for career change, we will define this class as ‘non-Target’ enrollee in this study. While value 1 represents the positive class, means the enrollee is open for career change, we will define this class ‘Target’ enrollee in this study, using capitalized ‘Target’ to distinguish it from column ‘target’.

A breakdown of non-Target enrollee and Target enrollee is provided in *Table 2* and *Figure 1*.

*Table 1.2 Number of non-target and target enrollee*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Label | Counts | Percentage |
| 0 | non-Target | 15934 | 86.79 |
| 1 | Target | 2425 | 13.21 |

Chart, pie chart

Description automatically generated

*Figure 1.1 Pie graph showing the ratio of target and non-target enrollee*

We can see there is an approximately 6:1 ratio of non-Target enrollee to Target enrollee. And this causes the baseline models to be overly insensitive when predict Target enrollee. Most of the models misclassify almost all the positive class (Target enrollee) in the test set (the baseline models’ recall on Target enrollee is very low). Therefore we will address this issue in the further section (Section 5) with oversampling technique.

Another issue of this dataset is high degree of feature overlap between classes. Many Target enrollee have feature values that are very similar to feature values of non-Target enrollee. This could bring difficulty for the models to predict Target enrollee with high precision.

Graphical user interface

Description automatically generated

*Figure 1.2 All variables (except ‘enrollee\_id’) plotted against variable ‘city*’. *Target enrollee are plotted in orange, non-Target enrollee are plotted in blue.*

By plotting different combination of variables onto two dimensional spaces, we can observe a lack of distinction between Target enrollee and non-Target enrollee. In *figure1.2*, the 2 classes can be seen to have almost indistinguishable distributions across various variables.

# 3 Package Introduction

# 4 Data Wrangling

Please find the general information of the original dataset from *Table 4.1,* sorted by the count of unique valuein each variable in descending order.

It contains columns of original variable index, variable name, counts of unique values, percentage of unique value, percentage of missing value and data type for each variable.

*Table 4.1*

| Index | Variable Name | Counts | Unique  Value Percentage | Missing Value Percentage | Data Type |
| --- | --- | --- | --- | --- | --- |
| 0 | enrollee\_id | 18359 | 100.00 | 0.00 | int64 |
| 12 | training\_hours | 241 | 1.31 | 0.00 | int64 |
| 1 | city | 123 | 0.67 | 0.00 | object |
| 2 | city\_development\_index | 93 | 0.51 | 0.00 | float64 |
| 8 | experience | 22 | 0.12 | 0.32 | object |
| 9 | company\_size | 8 | 0.04 | 26.03 | object |
| 7 | major\_discipline | 6 | 0.03 | 15.46 | object |
| 10 | company\_type | 6 | 0.03 | 27.45 | object |
| 11 | last\_new\_job | 6 | 0.03 | 2.00 | object |
| 6 | education\_level | 5 | 0.03 | 2.49 | object |
| 3 | gender | 3 | 0.02 | 22.32 | object |
| 5 | enrolled\_university | 3 | 0.02 | 1.86 | object |
| 4 | relevent\_experience | 2 | 0.01 | 0.00 | object |
| 13 | target | 2 | 0.01 | 0.00 | int64 |

As we can see from *Table 4.1*, 10 out of 14 variables are categorical data, 8 out of 14 variables have missing values, and the missing value ratios in variable ***company\_type, company\_size, gender*** are all higher than 20%.

After examining all the variables, we decided to group variables 6 groups based on the their features (we will exclude 'enrollee\_id' and 'target' from the grouping).

Then we chose the optimal data processing method for each group based on their variables’ characteristics. The process we took for each group will be find in *Table 4.2.*

*Tabel 4.2*

| Group | Variable Feature | Variable Names | Missing Value Percentage | Process |
| --- | --- | --- | --- | --- |
| Group 1 | experience related variables | relevent\_experience | 0 | data type correction |
| experience | 0.32 | data type correction, missing value imputation (fill with mode) |
| last\_new\_job | 2 | data type correction, missing value imputation (fill with mode) |
| Group 2 | employer related variables | company\_type, | 27.45 | data grouping, missing value imputation (fill with KNN imputation) \*KNN imputation will automatically convert the variable’s data type to float |
| company\_size | 26.03 | data grouping, missing value imputation (fill with KNN imputation) |
| Group 3 | education related variables | education\_level, | 2.49 | data grouping, data type correction, missing value imputation (fill with 0) |
| major\_discipline | 15.46 | data grouping, data type correction, missing value imputation (fill with 0 or mode) \*when education\_level equals to 0 (None) or 1 (high school or primary school)  fill missing value with 0, then fill the rest with 0 |
| enrolled\_university | 1.86 | fill with mode |
| Group 4 | gender | gender | 22.32 | data type correction, missing value imputation (fill with mode) |
| Group 5 | city related variables | city | - | data grouping, data type correction (dummy encoding) |
| city\_development\_index | - | - |
| Group 6 | training related variables | training\_hours | - | - |

In the cases that the variable has a missing value percentage higher than 25%, we chose KNN imputation over filling the missing value with 0 or with mode value to avoid introducing a certain type of bias to the dataset.

The cleaned data contains 18,359 rows and 23 columns. You can find more details about the cleaning process from this jupyter notebook [Detecting Potential Candidate\_Part1(01.01-05.02.06).ipynb](https://github.com/yoyo6022/Springboard_Capstone/blob/master/notebooks/01.01-05.02.06.ipynb).

# 5 Exploratory Data Analysis

The statistic including mean, standard deviation, minimum values, maximum values and percentile for each variable were summarized in *Table 5.1*

*Table 5.1*

| variable name | count | mean | std | min | 25% | 50% | 75% | max |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| city\_development\_index | 18359.0 | 0.847140 | 0.110189 | 0.448 | 0.796 | 0.91 | 0.92 | 0.949 |
| gender | 18359.0 | 1.085299 | 0.314034 | 1.000 | 1.000 | 1.00 | 1.00 | 3.000 |
| relevent\_experience | 18359.0 | 0.740563 | 0.438338 | 0.000 | 0.000 | 1.00 | 1.00 | 1.000 |
| enrolled\_university | 18359.0 | 0.237377 | 0.425487 | 0.000 | 0.000 | 0.00 | 0.00 | 1.000 |
| education\_level | 18359.0 | 2.082194 | 0.693807 | 0.000 | 2.000 | 2.00 | 3.00 | 3.000 |
| major\_discipline | 18359.0 | 1.585326 | 0.755994 | 0.000 | 1.000 | 2.00 | 2.00 | 2.000 |
| experience | 18359.0 | 10.647040 | 6.769905 | 0.000 | 5.000 | 9.00 | 16.00 | 21.000 |
| company\_size | 18359.0 | 1.895419 | 0.730158 | 1.000 | 1.000 | 2.00 | 2.00 | 3.000 |
| company\_type | 18359.0 | 2.700147 | 0.621645 | 1.000 | 3.000 | 3.00 | 3.00 | 3.000 |
| last\_new\_job | 18359.0 | 2.044338 | 1.680945 | 0.000 | 1.000 | 1.00 | 3.00 | 5.000 |
| training\_hours | 18359.0 | 65.899014 | 60.885300 | 1.000 | 23.000 | 47.00 | 89.00 | 336.000 |
| target | 18359.0 | 0.132088 | 0.338595 | 0.000 | 0.000 | 0.00 | 0.00 | 1.000 |

The overall distribution of numerical variables was visualized in *Figure 5.1*.Unique values and their counts breakdown by non-Target and Target enrollee for each categorical variable were visualized in *Figure 5.2*

A screenshot of text

Description automatically generated

*Figure 5.1 distributions of numerical variables*

*please note that this was visualized after data grouping and data type correction for categorical variables*

Graphical user interface, website

Description automatically generated

*Figure 5.2 Barplots of categorical variables break down by non-Target and Target enrollee*

*please note that this was visualized after data grouping*

You may have noticed that variable ‘experience’ is included in *Figure 5.2*, as we consider *Figure 5.3* (boxplot) and *Table 5.2* to be a better option to visualize this variable.

*Table 5.2*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | count | mean | std | min | 25% | 50% | 75% | max |
| non-Target | 15934.0 | 10.872725 | 6.763342 | 0.0 | 5.0 | 10.0 | 17.0 | 21.0 |
| Target | 2425.0 | 9.164124 | 6.625679 | 0.0 | 4.0 | 7.0 | 14.0 | 21.0 |

Chart, box and whisker chart

Description automatically generated

*Figure 5.3 boxplot for variable target and experience*

We can see that non-Target enrollee compares to Target enrollee, has slightly longer average working years and wider range of working experience in years.

By plotting the distribution of two classes in *Figure 5.4,* we noticed that regardless of non-Target enrollee or Target enrollee, there are 2 peaks of the distribution located at city\_development\_index 0.92 (city\_103 and city\_160) and 0.62 (city\_21).

Out of 2425 Target enrollee in our dataset, 1170 of them were from city\_103, city\_160 and city\_21. That shares almost 50% of the total number.

Chart, histogram

Description automatically generated

*Figure 5.4 city\_development\_index distribution breakdown by non-target (0) and Target enrollee(1)*

Chart, bar chart

Description automatically generated

*Figure 5.5 city barplot breakdown by non-target and Target enrollee*

After plotting heatmap *Figure 5.6*, we ca see no significant correlation coefficient was found among variables.

Chart, scatter chart

Description automatically generated

*Figure* 5.6 Heatmap of correlation coefficients between variables