Kelompok 2 - Data Wizard

7th February 2022

HR Analysis of Job Change

Purwadhika JC Data Science After Hour Bootcamp

Team Members:
Rizky Syamsudin Halim
Ferdiansyah Ashil
Arief Tri Munandar

Hi! Meet Data Wizards!

Rizky Syamsudin Halim

Management Associate - Allianz Indonesia

Arief Tri Munandar

General Affair Expatriate - KORINDO GROUP

Ferdiansyah Ashil

Procurement Staff - Central Bank of Indonesia

Outline

Problem Statement

Data Understanding

EDA

Pre-Processing and Modelling
(Methodology)

Conclusion

Problem Statement

Description of the issue to be adressed or solve.

Problem Statement - Context

A company which is active in Big Data and Data Science wants to hire data scientists among people who successfully pass courses conducted by the company…

The company needs to know, which of these people really wants to work for the company

This will help the company keeps costs and time investment down.

Business Objective

'Predict which candidates really wants to work for the company after training'

Output

Model to Classify who will stay after training/not

Performance Measure

Recall, Precision, and F1 Test

Data Understanding

Knowledge, attribute, or general contents of the data.

Data Understanding

Overview

This dataset contained anonymous information of candidates applying for a training and their various background.

Dataset Statistics	Count	Variable Types	Count
Number of variables	14	Numeric	4
Number of observations	19158	Categorical	10
Missing cells	20778		
Missing cells (%)	7.74 %		
Duplicate rows	0		
Duplicate rows (%)	0.0 %		
Total size in memory	1.87 MB		

Source

This dataset could be found on kaggle.

Column Explanation

No	Feature	Data Type	Null (%)	Description
1	enrolee_id	int64	0 %	Unique Id for enrolee
2	City	object	0 %	City Code
3	City_development_index	float64	0 %	Scaled development index for the city (How developed is the city)
4	gender	object	23.53 %	The gender of the enrolee
5	relevant_experience	object	0 %	Shows if the enrolee have any relevant experience or not
6	enrolled_university	object	2.01 %	The university type of the enrolee
7	education_level	object	2.40 %	The level of the enrolee education
8	major_discipline	object	14.68 %	The discipline of the enrolee education
9	experience	object	0.33 %	The total experience of the enrolee
10	company_size	object	38.67 %	Number of employee in the enrolee current company
11	company_type	object	32.04 %	Type of the current employee
12	last_new_job	object	2.20 %	Difference in years between current and last job
13	training_hours	int64	0 %	Training hours completed by the enrolee
14	target	float64	0 %	Defined if the candidate looking for a job change or not

Exploratory Data Analysis

Summarizing characteristics of Data (e.g: Column relationships, multicollinearity, null values in columns, etc.)

Null Values

First, let's see how complete our dataset is..

Number of Null/Missing Values

enrollee_id	0
city	0
city_development_index	0
gender	4508
relevent_experience	0
enrolled_university	386
education_level	460
major_discipline	2813
experience	65
company_size	5938
company_type	6140
last_new_job	423
training_hours	0
target	0

Percent Missing

	Count Missing	Percent Missing
enrollee_id	0	0.000000
city	0	0.000000
city_development_index	0	0.000000
gender	4508	23.530640
relevent_experience	0	0.000000
enrolled_university	386	2.014824
education_level	460	2.401086
major_discipline	2813	14.683161
experience	65	0.339284
company_size	5938	30.994885
company_type	6140	32.049274
last_new_job	423	2.207955
training_hours	0	0.000000
target	0	0.000000

Insight

We've discovered that 4 columns have a lot of data missing, the rest have a bit of data missing or none at all.

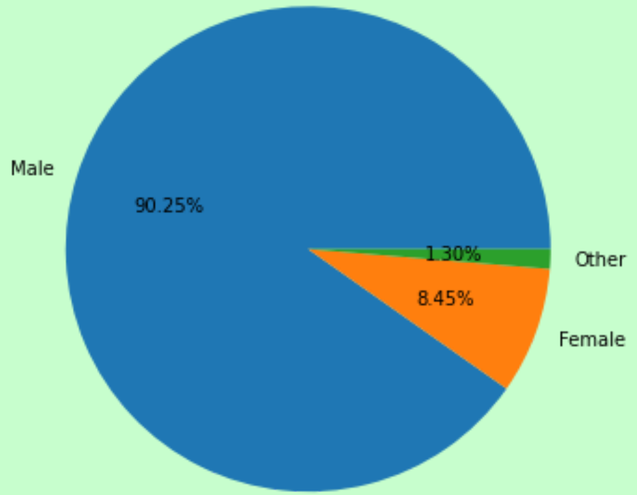
Solution

Drop empty values if it's only missing around 2%, impute with mode or mean if too much is missing.

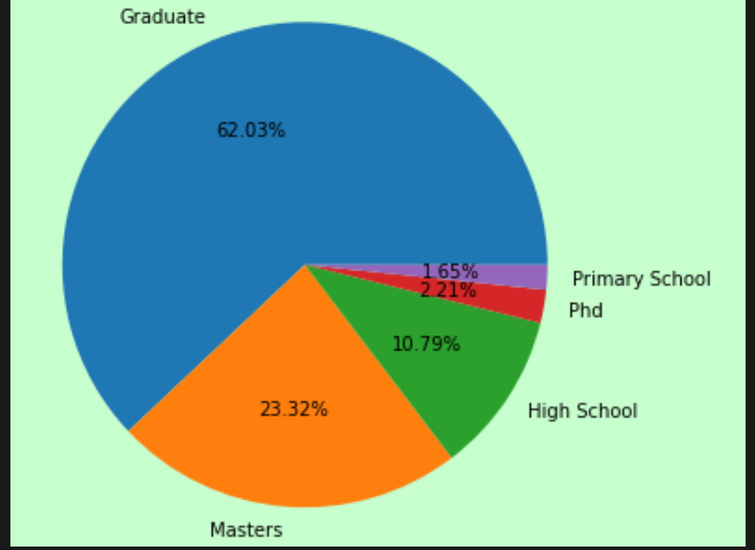
Data Distribution

Next, we will try to see the data distribution on some columns..

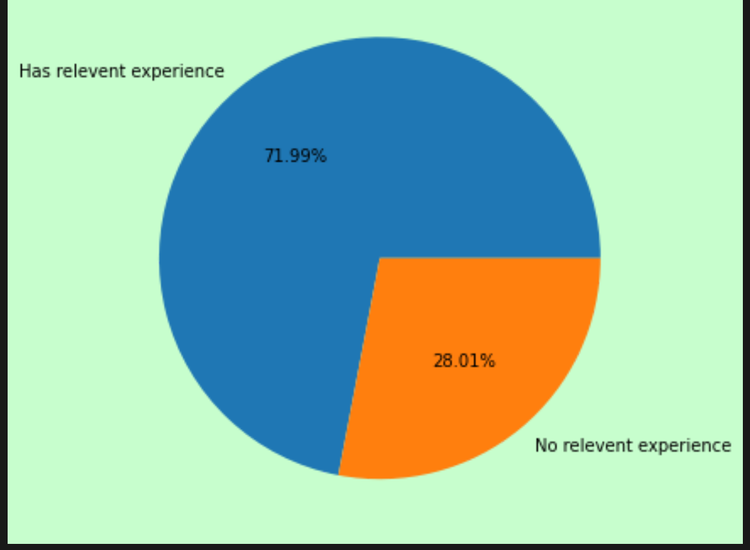
Pie Chart for Gender



Pie Chart for Relevant Exp



Pie Chart for Relevant Exp



Insight

The data is mostly imbalanced, as we see here on the next sets of plot.

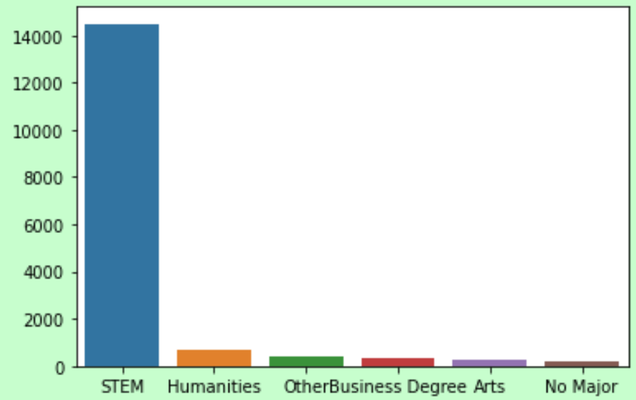
Solution

Using undersampling/oversampling later, using normalized graph to explore columns.

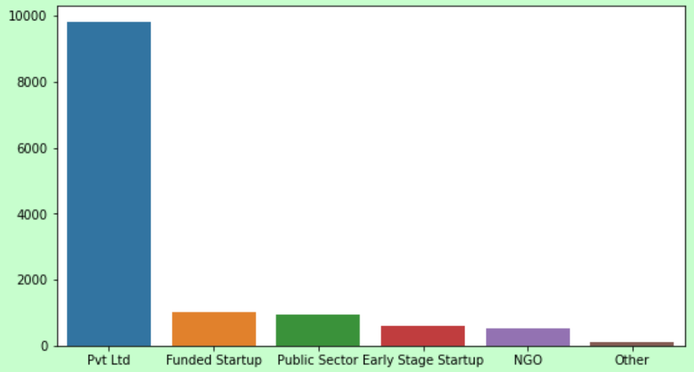
Data Distribution

Continuing on from the previous slides..

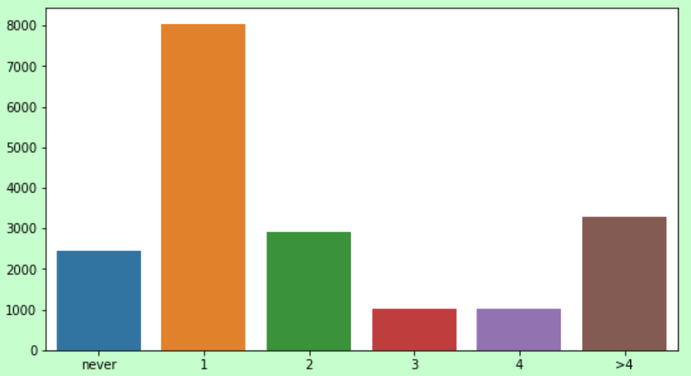
Bar Chart for Majors



Bar Chart for Company Types



Bar Chart for Last New Job



Insight

These second sets of data is also mostly imbalanced.

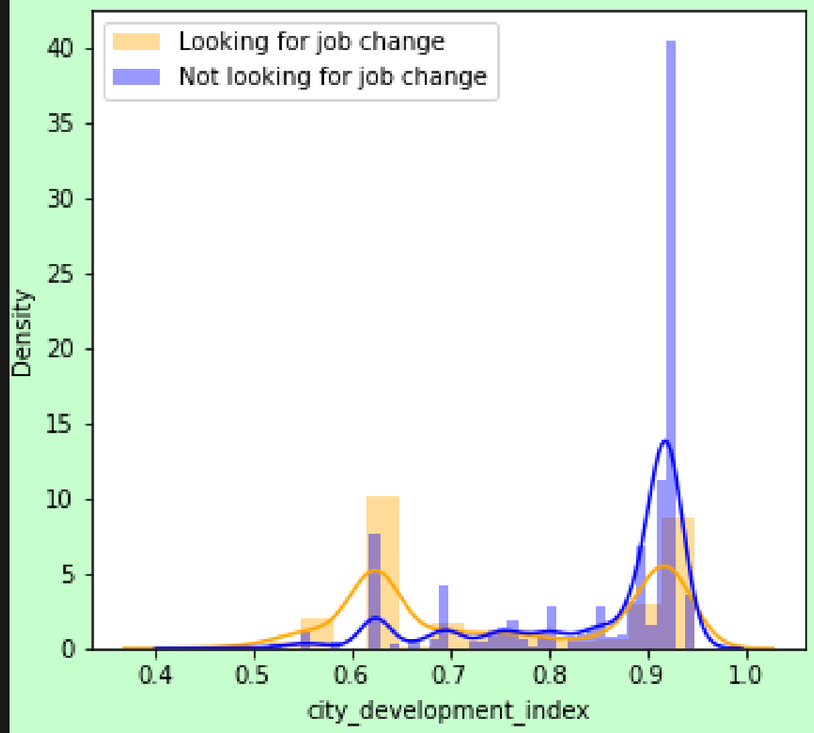
Solution

Using undersampling/oversampling later, using normalized graph to explore columns.

Columns Exploration

Next we will see which columns affects Job Seeking tendency..

City Development Index



Training Hours

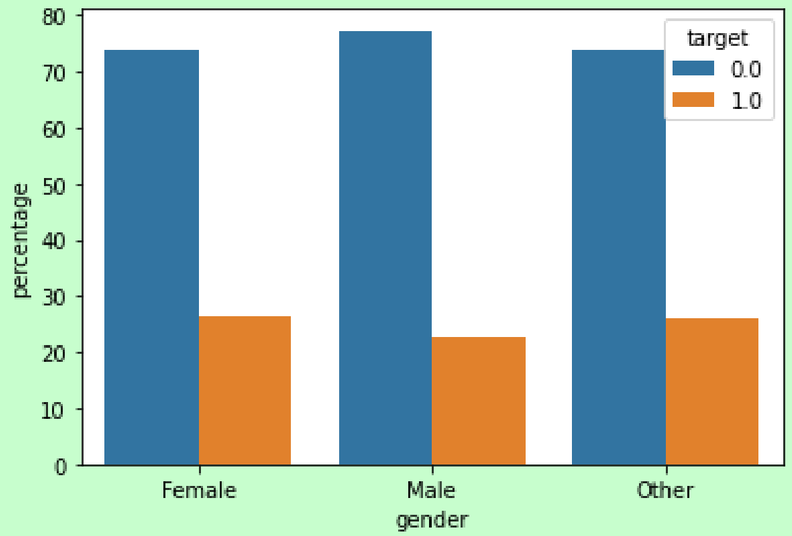
Insight

People from developed cities (e.g: 0.9 index) tend to not look for jobs and people from less developed cities (e.g: 0.6 index), tend to look for jobs more

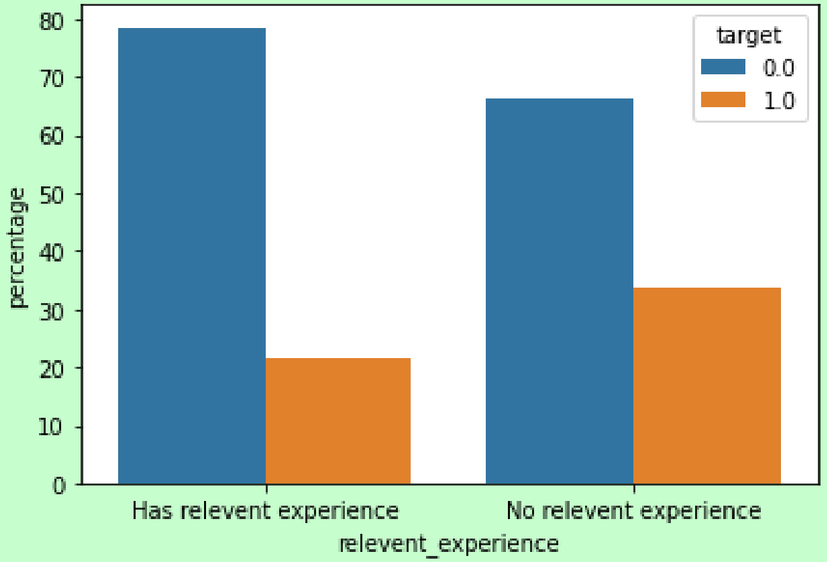
Training hours seems to be the same and does not vary much between people who looks for jobs and people who don't.

Columns Exploration

Gender



Relevent Exp



Insight

Gender does not seems to effect the tendency to look for jobs. We see a relatively same distibution.

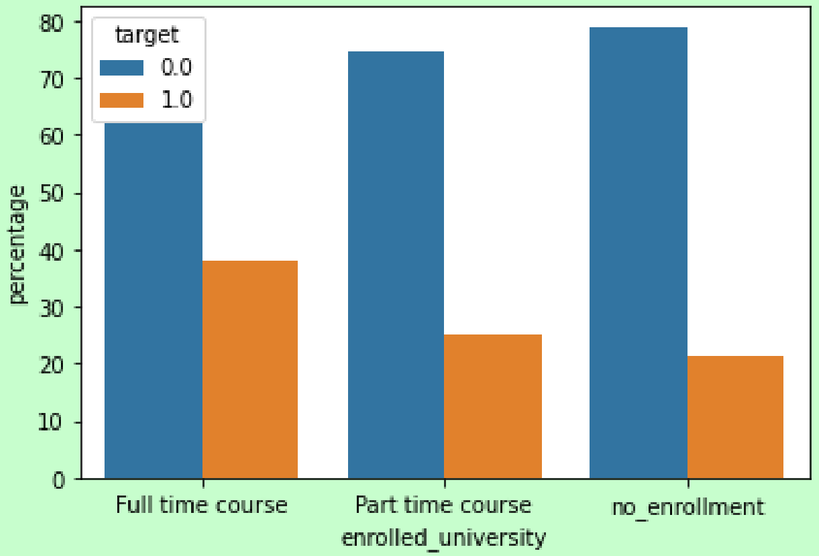
People who have no relevent experience regarding the job seems to have a tendency to look for jobs more.

Plot Information

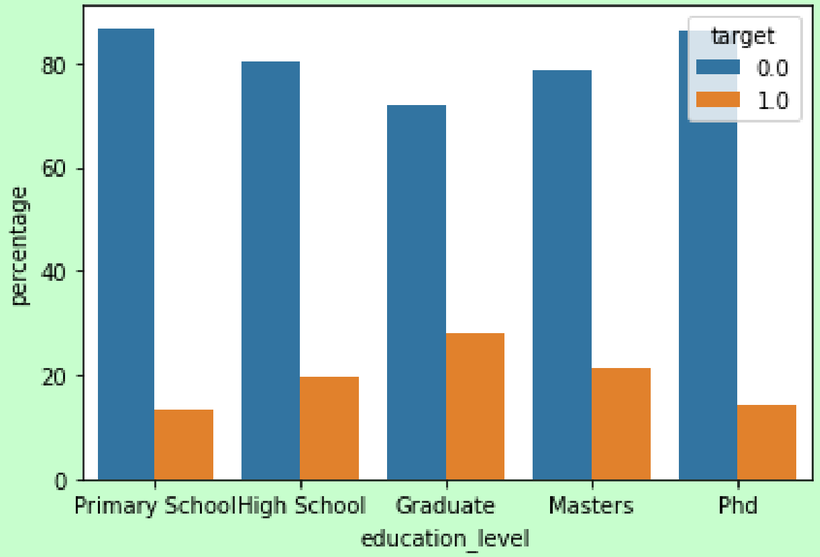
These plots are normalized. Target 0 represents non-job seekers, and target 1 represents job seekers. Other in gender represents null values.

Columns Exploration

Enrolled University



Education Level



Insight

People with longer courses tend to look for jobs more.

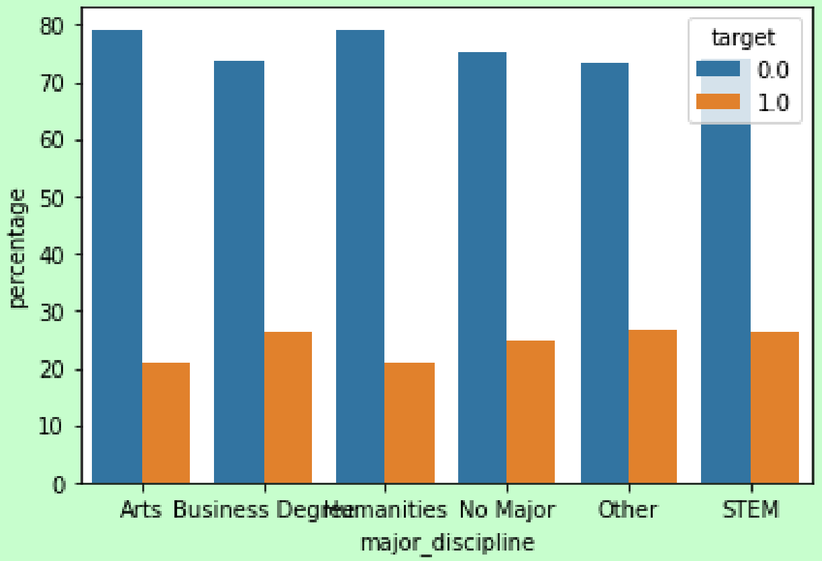
Up to graduate level, the tendency to look for jobs increase, and then decrease after that.

Plot Information

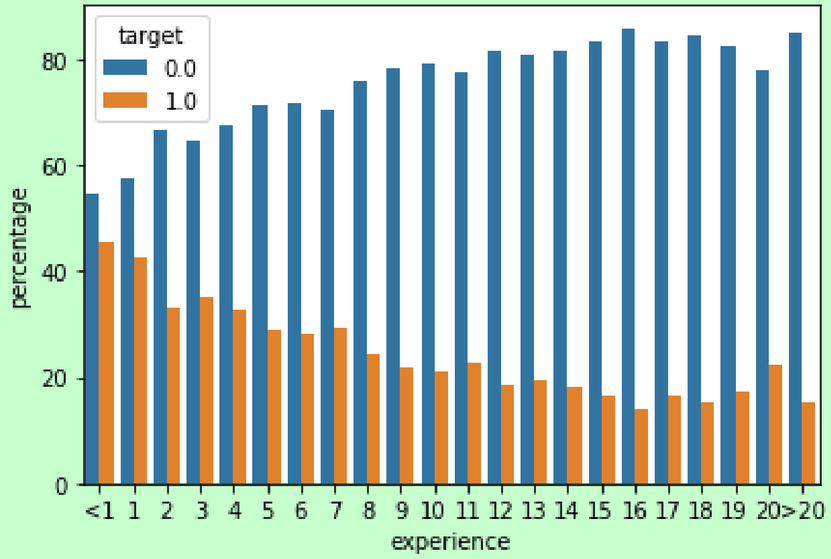
These plots are normalized. Target 0 represents non-job seekers, and target 1 represents job seekers. These graph is ordered (ascending for education level and descending for enrolled university).

Columns Exploration

Major Discipline



Experience



Insight

People with arts or humanities degrees seems less likely to look for jobs.

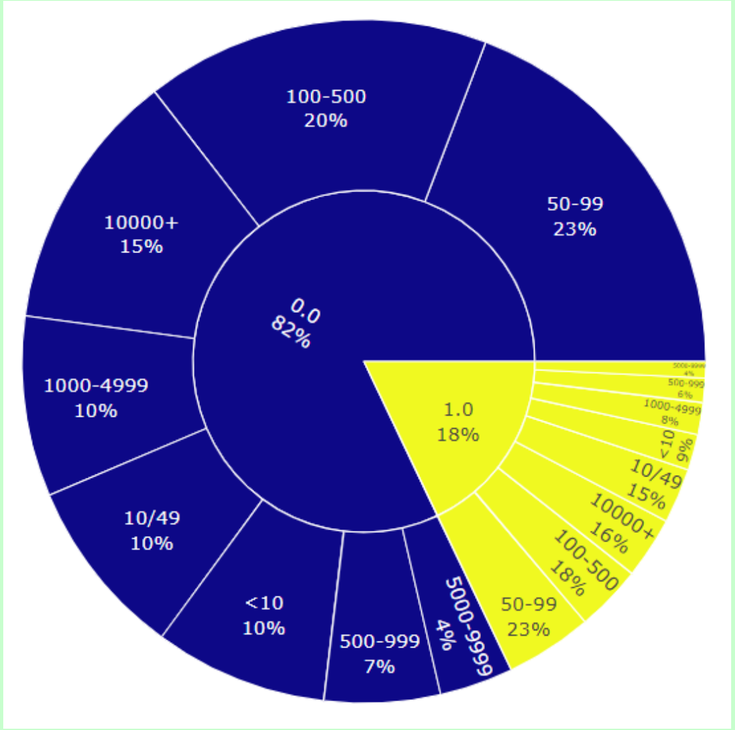
The more experience someone have, the tendency to look for jobs seems to decrease.

Plot Information

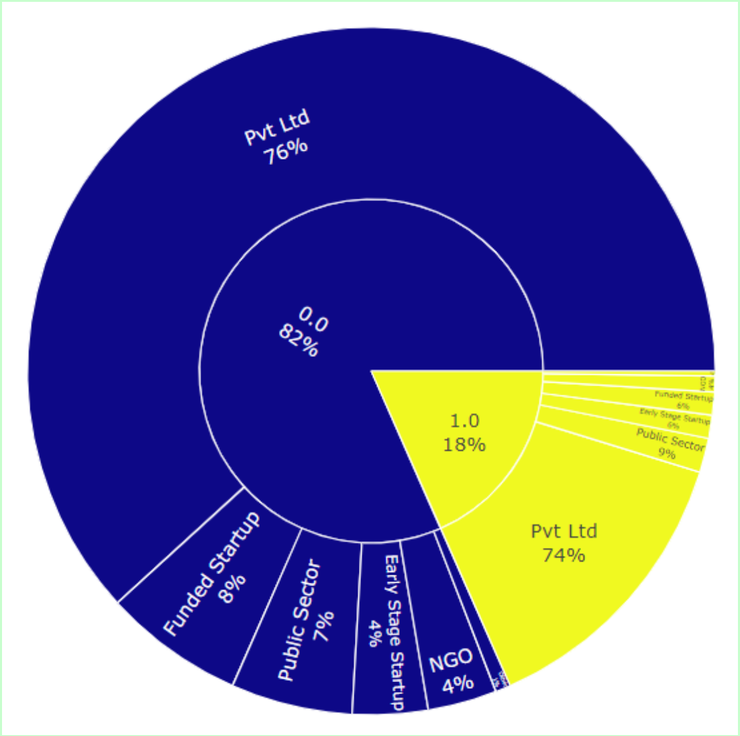
These plots are normalized. Target 0 represents non-job seekers, and target 1 represents job seekers. These graph is ordered ascending for experience.

Columns Exploration

Company Size



Company Type



Insight

The composition of Job seekers and non job seekers seems to be the same for company type. Only differing around 2% for each category.

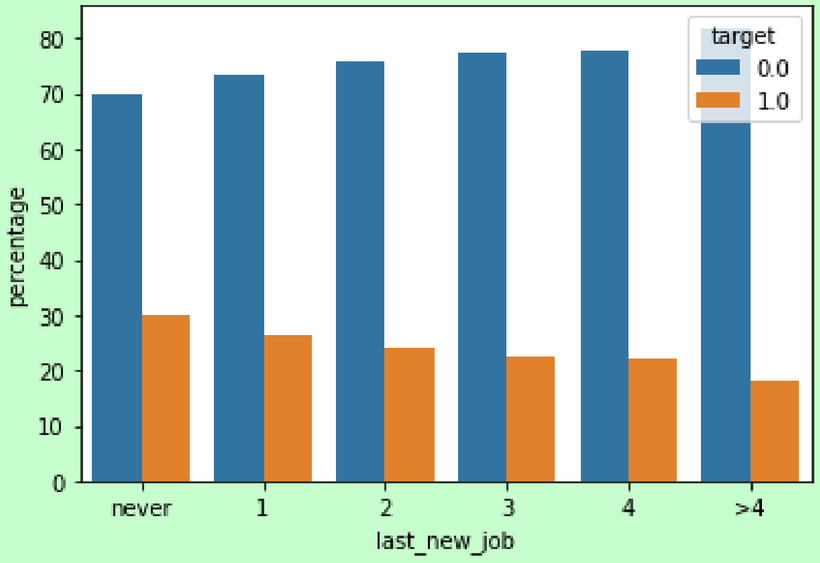
Same case with company size, it seems the composition is pretty similar.

Plot Information

Plots wants to compare the composition of job seekers and non job seekers by company type and company size. 1 represents job seekers and 0 represents non-job seekers.

Columns Exploration and Conclusion

Last New Job



Insight

People who stayed longer at their previous company tend to not look for jobs.

Plot Information

These plots are normalized. Target 0 represents non-job seekers, and target 1 represents job seekers. The graph is in Ascending order

Conclusion

- Data is Imbalanced.
- Gender, company type, training hours, and company size seems to not affect the target column significantly.
- Discovered some of missing values for gender, company type and company size, education level, last new job, experience, and enrollment columns.

Statistics and Multicollinearity Check

Lastly, we will check for high correlation or multicollinearity using multiple methods..

Chi-Square

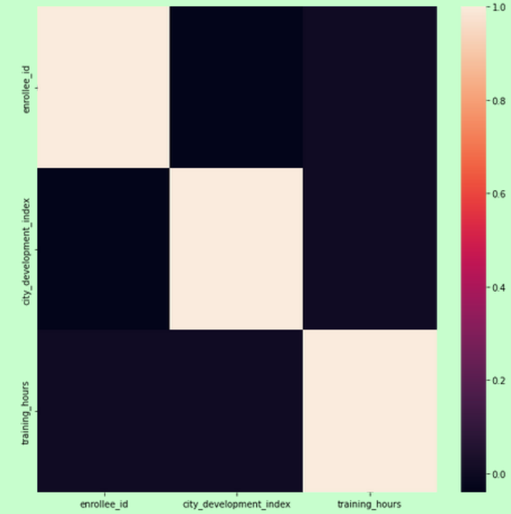
	Feature	p-value
0	city	0.000000e+00
1	relevent_experience	0.000000e+00
2	enrolled_university	0.000000e+00
3	education_level	0.000000e+00
4	experience	0.000000e+00
5	last_new_job	0.000000e+00
6	company_size	1.078000e-07
7	company_type	1.480300e-06
8	gender	1.087715e-02
9	major_discipline	3.205736e-02

F-Statistics

	Numerical_Feature	F-Score	p values
0	city_development_index	2531.716218	0.000
1	enrollee_id	47.004194	0.000
2	training_hours	8.922761	0.003

VIF

	variables	VIF
2	training_hours	2.135706
0	enrollee_id	3.759770
1	city_development_index	4.729460



Correlation Matrix

Insight

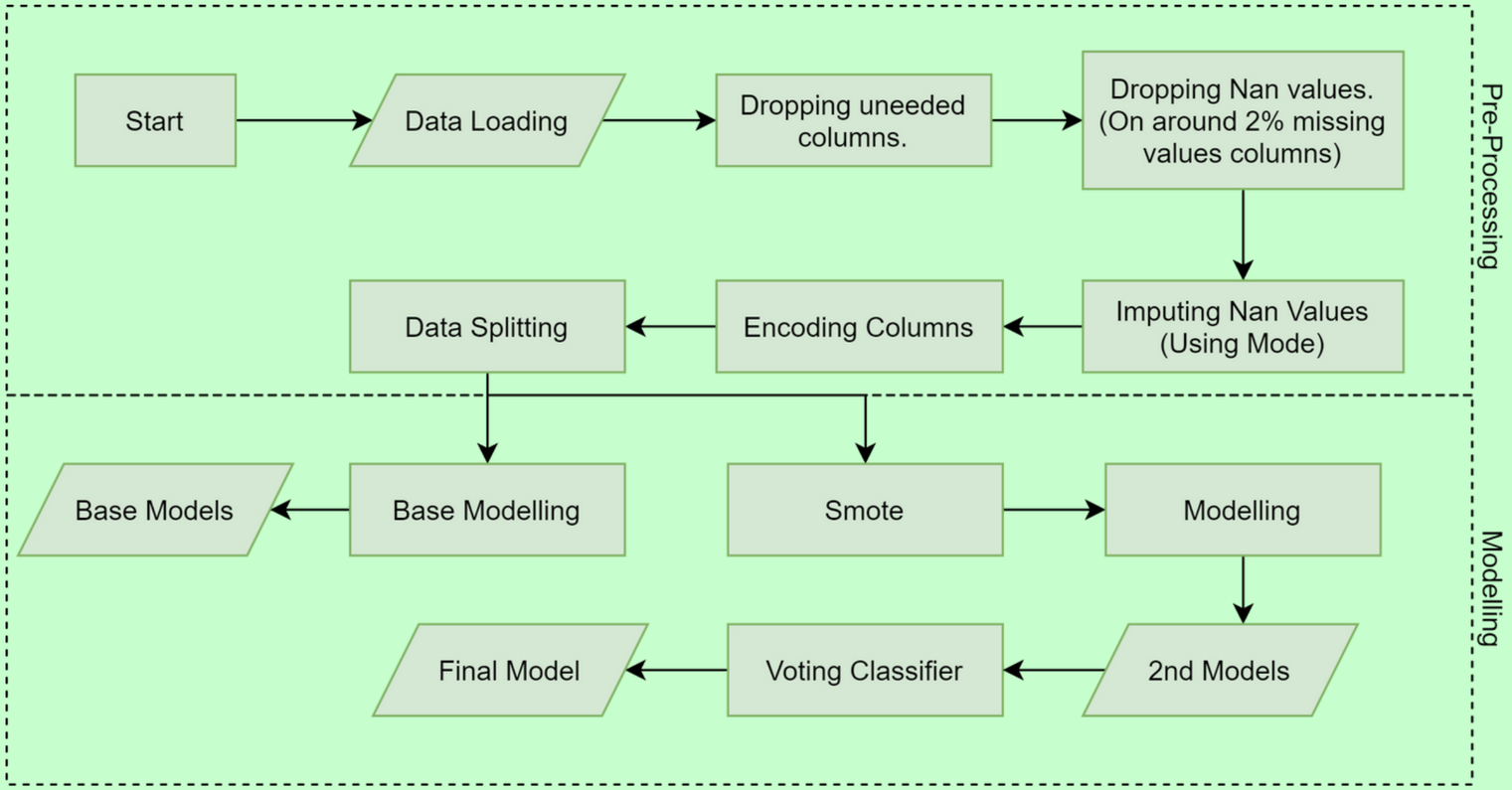
All of the tests shows that there are no big correlation or multicollinearity issue, either for the categorical or numerical data.

Pre-Processing and Modelling

Summarizing characteristics of Data (e.g: Column relationships, multicollinearity, null values in columns, etc.)

Pre-Processing and Modelling Steps.

Here's an Overview of our Pre-processing and Modelling steps.



Dropped Columns

Genders, enrolee id, city, training hours, company type, company size.

Encoder Used

One hot for relevent experience and Major. Ordinal for education level, last new job, experience , and enroll university.

Models Used

Decision Tree, Random Forest, XGBClassifier, Logistic Regression, AdaBoostClassifier.

Final Look of Pre-Processed Data.

	Count Missing	Percent Missing
city_development_index	0	0.0
relevent_experience	0	0.0
enrolled_university	0	0.0
education_level	0	0.0
major_discipline	0	0.0
experience	0	0.0
last_new_job	0	0.0
target	0	0.0

Missing Values

As you can see, our data no longer have missing values after dropping and imputing nan values. We use mode for imputing because the missing values are Categorical data, which means we cannot use mean in imputing them.

	x0_Has relevent experience	x1_Arts	x1_Business Degree	x1_Humanities	x1_Other	x1_STEM	education_level	enrolled_university	experience	last_new_job	city_development_index
0	1.0	0.0	0.0	0.0	0.0	1.0	4.0	1.0	10.0	6.0	0.843
1	1.0	0.0	0.0	0.0	0.0	1.0	3.0	1.0	19.0	3.0	0.896
2	1.0	0.0	0.0	0.0	0.0	1.0	3.0	1.0	12.0	2.0	0.920
3	1.0	0.0	0.0	0.0	0.0	1.0	3.0	1.0	12.0	6.0	0.730
4	1.0	0.0	0.0	0.0	1.0	0.0	3.0	1.0	7.0	3.0	0.527

One-Hot & Dropped Columns.

We use one-hot encoder when the data does not have a certain 'hierarchy', we also need to drop one encoded column to prevent multicollinearity.

Ordinal Encoded Columns

The columns which have a certain hierarchy or order we will encode with ordinal encoder.

Base Modelling

The main metric used here will be Precision, because false positive can be costly to the company.

Then after pre-processing and splitting (To gather training and testing dataset), we test some modelling methods and here's the result:

Decision Tree

Precision: 0.440
Recall: 0.375
F1 Score: 0.405

Random Forest

Precision: 0.488
Recall: 0.383
F1 Score: 0.429

Logistic Regression

Precision: 0.594
Recall: 0.240
F1 Score: 0.342

XGBClassifier

Precision: 0.603
Recall: 0.424
F1 Score: 0.493

AdaBoost

Precision: 0.599
Recall: 0.338
F1 Score: 0.432

Conclusion

XGBClassifier gives the best performance, although still lacking. We will try handling imbalance next then model them again.

Handling Imbalance

We will try to use a method called SMOTE to oversample and handle imbalance.

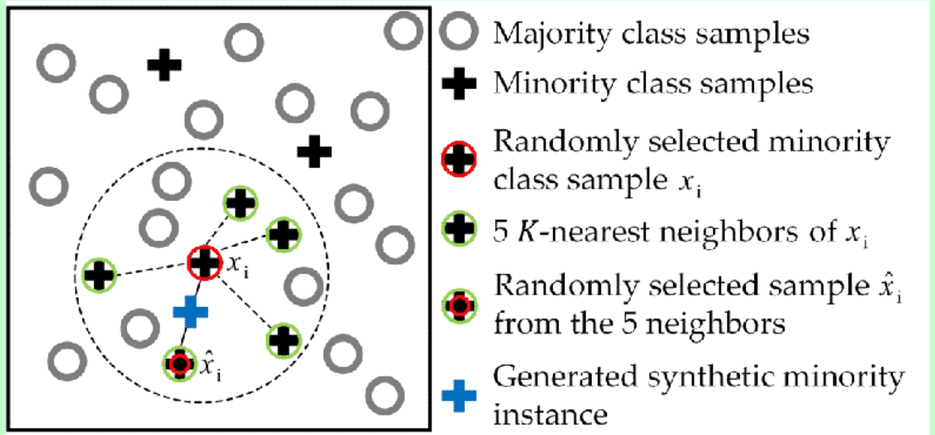
What is SMOTE?

SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling

Why?

Because under-sampling is usually not preferred due to the possible loss of important data.

SMOTE Illustration



Before and After SMOTE

```
y_train_val.value_counts()
```

```
0.0    10874
1.0     3537
Name: target, dtype: int64
```

```
y_over.value_counts()
```

```
1.0    10874
0.0    10874
Name: target, dtype: int64
```

Modelling After SMOTE

The main metric used here will be Precision, because false positive can be costly to the company.

After oversampling we tried to model the result again and get these results:

Decision Tree

Precision: 0.796
Recall: 0.718
F1 Score: 0.755

Random Forest

Precision: 0.786
Recall: 0.729
F1 Score: 0.756

Logistic Regression

Precision: 0.704
Recall: 0.632
F1 Score: 0.666

XGBClassifier

Precision: 0.784
Recall: 0.715
F1 Score: 0.748

AdaBoost

Precision: 0.754
Recall: 0.651
F1 Score: 0.699

Conclusion

Decision tree, random forest, and XGBClassifier is able to perform the best after oversampling. Next we will use these 3 in a Voting Classifier.

Voting Classifier

Next, We will use a Voting Classifier to combine the best performing models into 1 Classifier..

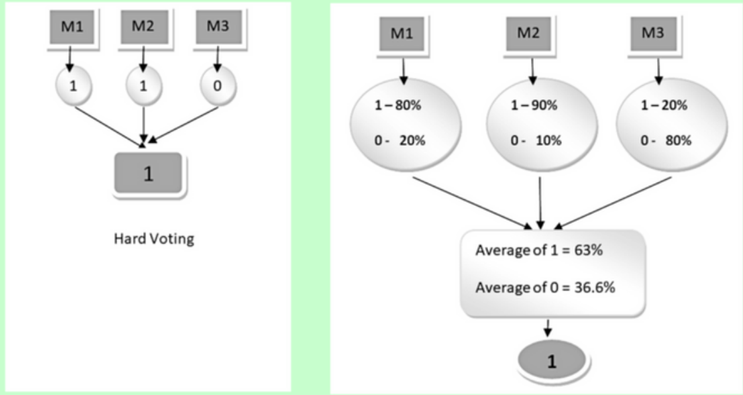
What is VC?

A VC will aggregate the findings of each base estimator. The aggregating criteria can be combined decision of voting for each estimator output.

Why?

VC may be able to help our model performance by resolving an error tendency of one model using voting.

VC Illustration



Final Result After VC is implemented

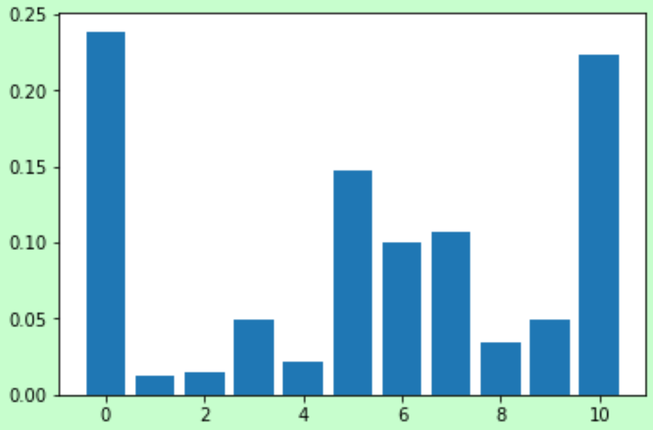
Precision: 0.797
Recall: 0.744
F1 Score: 0.770

Note: We used the 'soft' voting method for this VC.

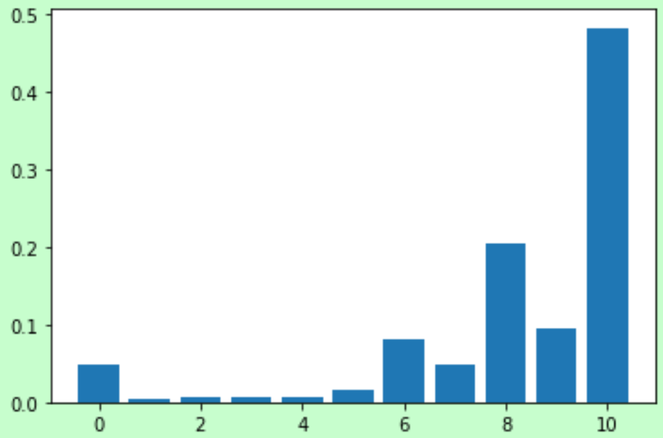
Discovering which Criteria is most Relevant

We will try to see which columns affects the predictions the most..

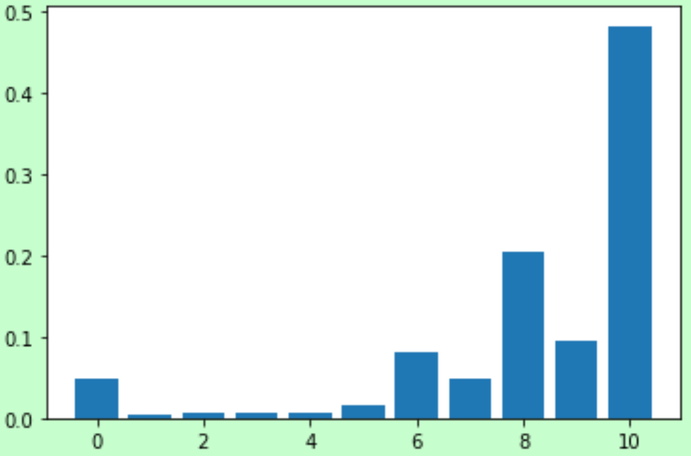
XGB Classifier



Random Forest



Decision Tree



Columns (Ordered from 0 to 10)

x0_Has relevent experience	x1_Arts	x1_Business Degree	x1_Humanities	x1_Other	x1_STEM	education_level	enrolled_university	experience	last_new_job	city_development_index

Insight

Most important features includes city_development_index, experience, last new job, and education level. XGBoost perceive relevent experience as extremely important. Recruiter could look out for these criteria to screen early.

Conclusion and Recommendation

Summary of findings and modelling results and recommended next steps.

Conclusion

Here's the final result and main discovery we learned in this project:

In order to predict job-seekeing tendency, we made some base models, we found that some models performed better than others, and to better the results we used Voting Classifier to produce a model.

Final Model Result

Precision: 0.797
Recall: 0.744
F1 Score: 0.770

Important Features

Next we also discovered some of the features that greatly affects the will to move or not for people. It includes:

1. City Development Index (How developed are the cities they are from).
2. Experience (Total working experience).
3. Education Level.
4. Last job length.
5. Having relevent experience or not.

We hope this discovery also can help the HRs to decide and pre-screen candidates faster.

Recommendation

Next there are some recommendation we would like to give regarding this model and its future deployment or use...

1. Keep improving and retraining the model as more data is getting gathered and business problem is evolving. We cannot assume that the same insights and pattern we gathered here will always stay the same, so this model will need to be retrained in the future (with different insights and approach if needed).
2. Gather more possibly affecting factors, such as current salaries (which may be one of the biggest decision affecting factors for jobs).
3. Tune the model used. One of the step we haven't done in this project is tuning, and it can be done to improve the model performance even more.

.

Thank you for your attention!

- Data Wizard Team