

Pathrise Project

Data Roadmap final Assignment
Reza Saeedisepehr



Outlines

- Introduction to Pathrise company
- Pathrise Project
- Methodology
- Data Collection
- Data Wrangling
- Exploratory data analysis(EDA)
- Performing Machine learning



Pathrise Project

Introduction to Pathrise company

MANIFESTO: We seek to uplift job seekers in their careers and help them fulfill their hopes, ambitions and livelihoods.



Pathrise Project

Pathrise's company as a recruitment agency holds a program which helps job seekers find a job. Actually this project is a combination of a classic classification problem and regression. According to data of people getting involved in Pathrise's program in the past, the project has two main Objectives.

1. Preparing a model to predict whether people would find a job or not ?
2. Preparing a model to predict how long does it take to find a job?



Methodology

Executive Summary

1. Data collection methodology:

1. Data is provided by Pathrise company in excel format.

2. Perform data wrangling

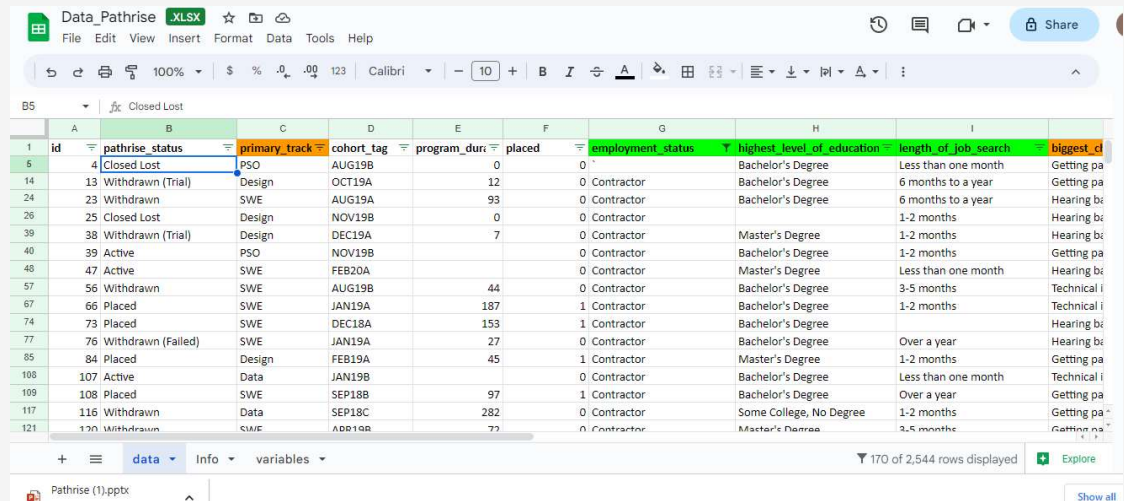
1. Converting categorical data
2. Dealing with missing values
3. Working with outliers

3. Perform exploratory data analysis (EDA) using visualization

4. Perform predictive analysis using classification and regression models

1. Four models are trained and examined by grid search method with different hyper-parameters and eventually the best model with the lowest error is selected to predict whether or not someone would find a job
2. Three regression models are trained and finally the best model with the lowest error is selected to predict how long how long a person would find a job

Data Collection



Pathrise (1).pptx

Tabular data is provided by Pathrise company in excel format

Items	values
Number of column	16
Number of rows	2544
percentage of Numerical columns	31.25%
percentage of Categorical columns	68.75%
Average percentage of missing values	5.46%

Data Wrangling And Main Challenges



Taking an appropriate approach to deal with Categorical data

More than 68% of data is categorical

Choosing suitable methods to solve the missing values issues

Some columns includes more than 24% missing values.

Data preparation approach

Action plan to deal with different columns

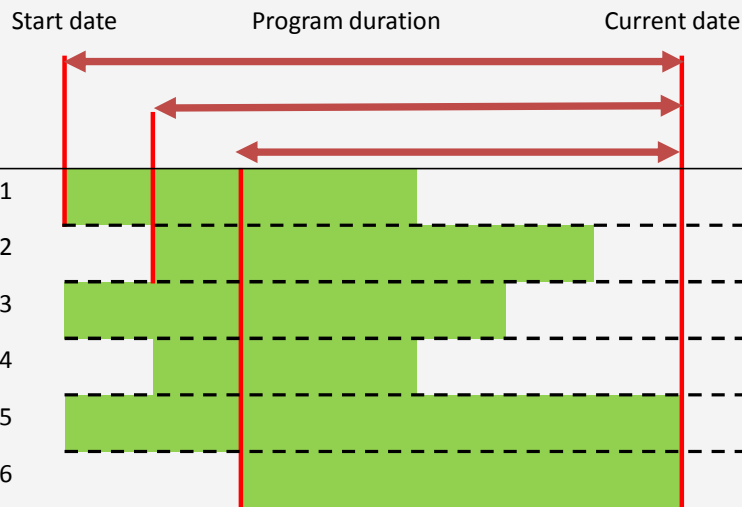
Column Name	Type	Approaches	Percentage of Missing values	Method to deal with missing values
id	Numerical	Remove/Useless	0.00%	-
pathrise_status	Categorical/Nominal	Remove/Data leakage	0.00%	-
primary_track	Categorical/Nominal	Covert to dummy values	0.00%	-
cohort_tag	Categorical/Ordinal	Convert to start date\Remove	0.31%	-
program_duration_days	Numerical	-	24.21%	calculation based on cohort_tag
placed	Numerical	-	0.00%	-
employment_status	Categorical/Ordinal	Replaced by ordinal number/Remove	9.00%	Calculation based on high frequency
highest_level_of_education	Categorical/Ordinal	Replaced by number of year spent on education/Remove	2.28%	Calculation based on high frequency
length_of_job_search	Categorical/Ordinal	Replaced by average value of period/Remove	2.91%	Calculation based on Average
biggest_challenge_in_search	Categorical/Nominal	Covert to dummy values	0.94%	Replaced by No challenge
professional_experience	Categorical/Ordinal	Replaced by average value of period/Remove	8.73%	Calculation based on Average
work_authorization_status	Categorical/Nominal	Covert to dummy values	10.14%	Calculation based on high frequency
number_of_interviews	Numerical	-	8.57%	Calculation based on Average
number_of_applications	Numerical	-	0.00%	-
gender	Categorical/Nominal	Remove/prevention of model bias	19.97%	-
race	Categorical/Nominal	Remove/prevention of model bias	0.71%	-



Pathrise project

How to calculate program duration for missing values

Time line



*S stands for current student

Finding start program date

1. Cohort tag: each cohort starts on the first (A) and the third week (B) of the month. For instance, FEB20A/FEB20B cohort starts on the first/third Monday of February 2020.

2. Define a "get_mondy" function to convert Cohort tag data to date format

Current date assumption

1. Program duration day: show many days a fellow was in the program, **N/A for current fellows**

2. The most recent date according to the cohort tag column is assumed as the **current date**

Calculating Program duration for current student

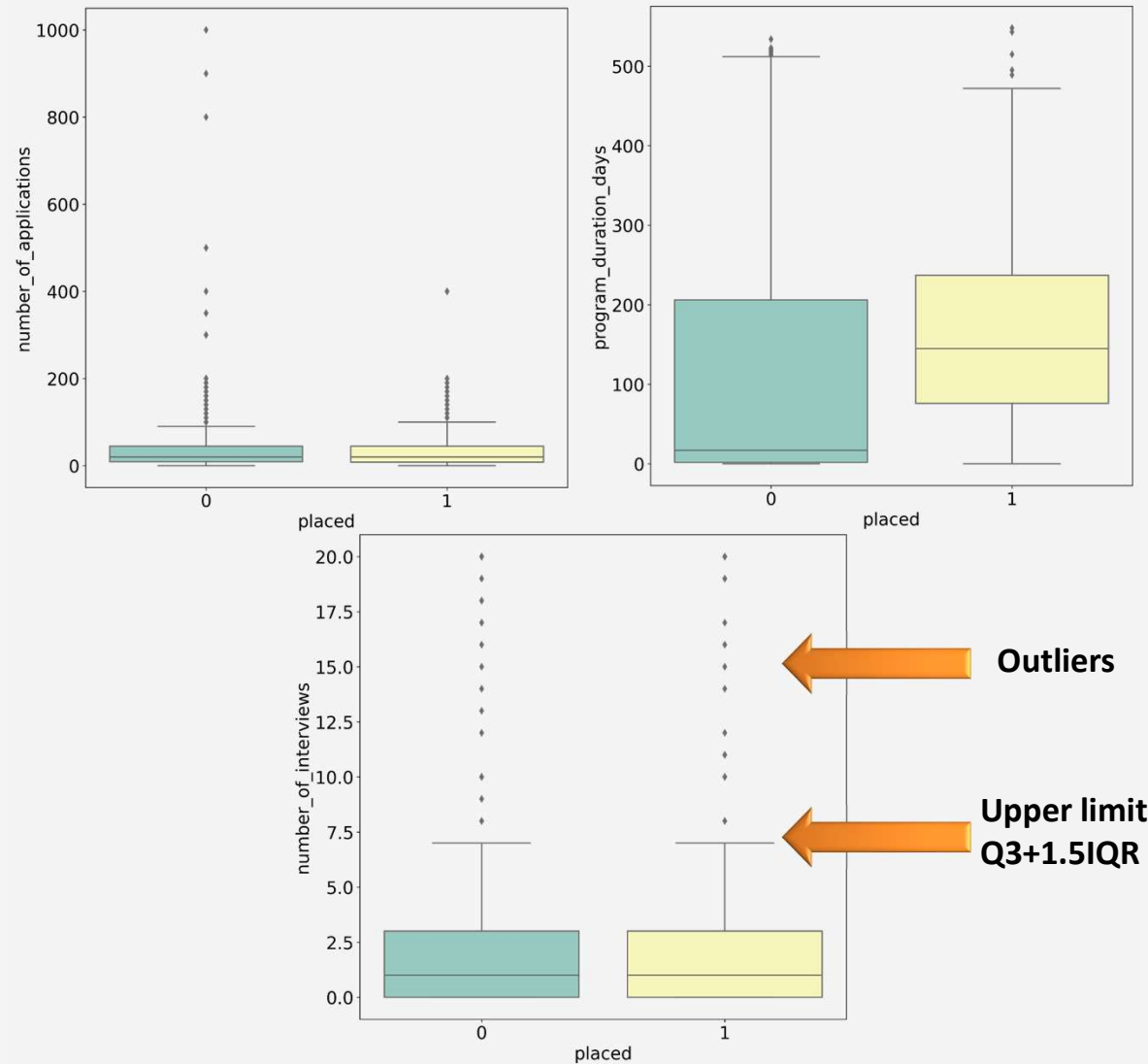
Difference between start program date and current date is considered as program duration days for **current student**



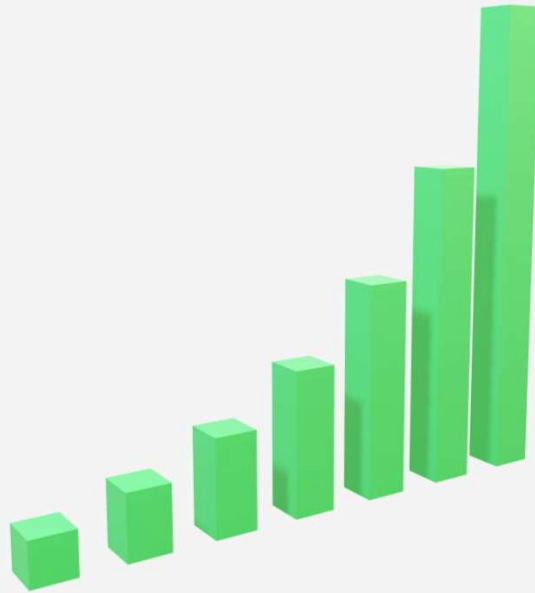
Pathrise project

• Working with Outliers

1. Exploration of data reveals that numerical columns of dataset including “**number of applications**”, “**number of interviews**” and “**program duration days**” have outliers
2. Interquartile range(IQR) is used to indicate the outlier
3. **Outliers are replaced by mean values.**

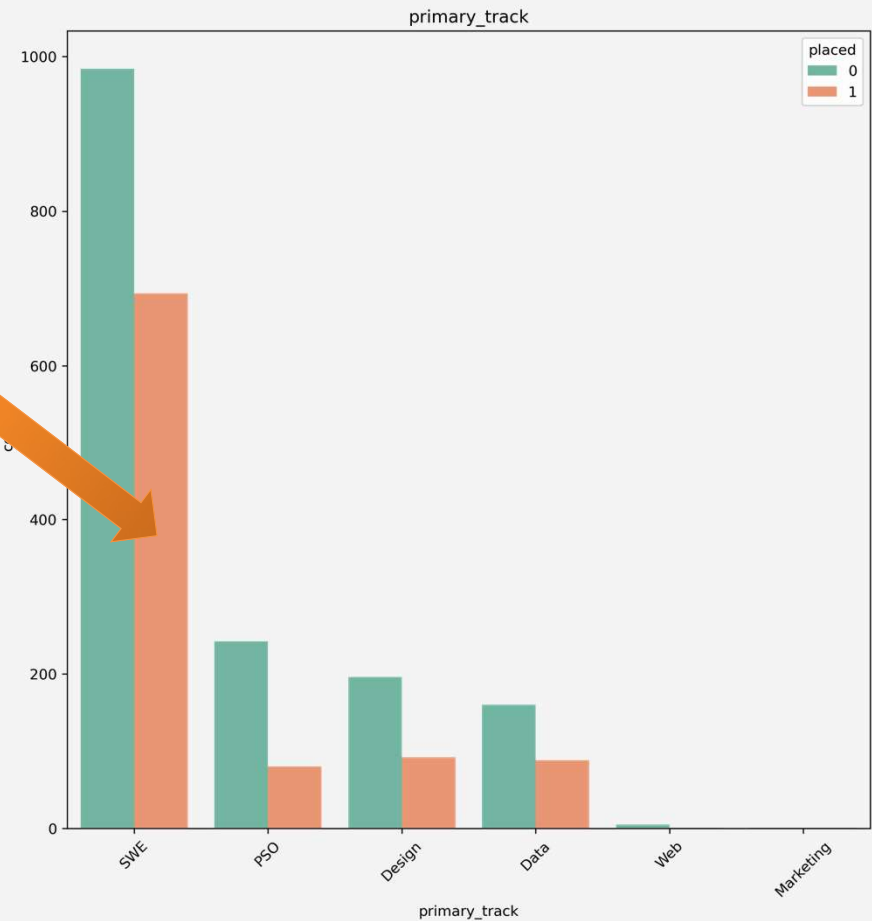
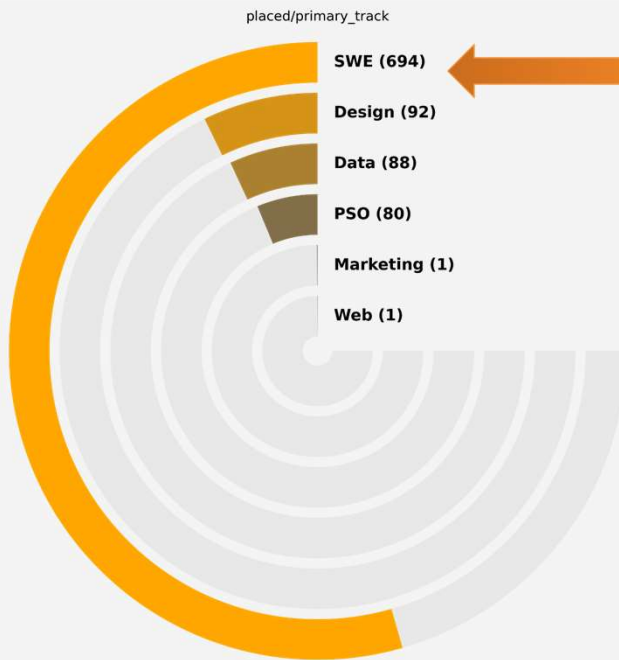


Exploratory data analysis(EDA)



Explorer in primary track

1. Which group of primary tracks have more population?
2. Which primary track is more successful to find a job than others?



Explorer in primary track

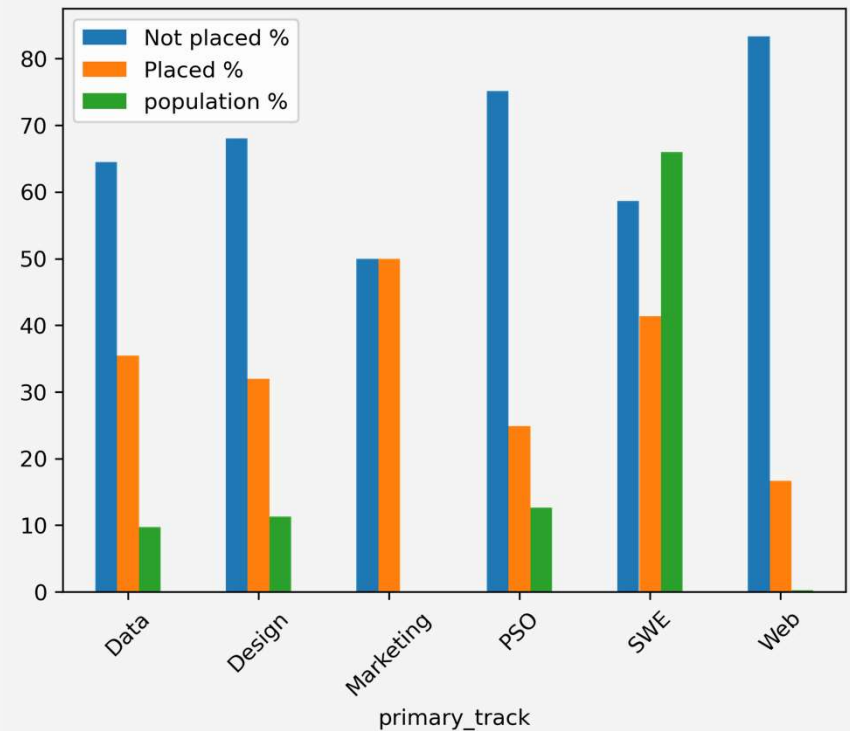
The percentage of people being successful to find a job changes if they are compared according to their population

Primary_track	Data	Design	Marketing	PSO	SWE	Web
Not placed %	64.5	68.1	50.0	75.2	58.6	83.3
Placed %	35.5	31.9	50.0	24.8	41.4	16.7
population %	9.7	11.3	0.1	12.7	66.0	0.2

Most successful

Most popular

The number of people whose primary track is marketing only two, so this group should be ignored as a insufficient evidence

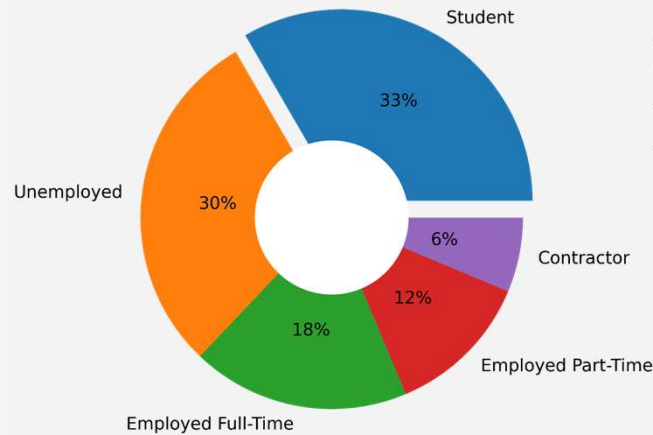


Explorer in employment status

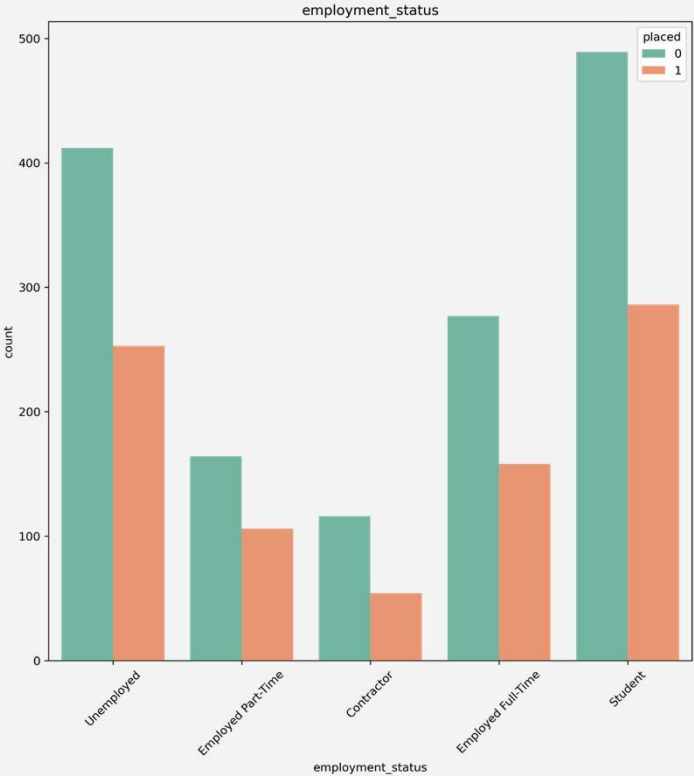
1. Which group of people had more chance to find a career?

	Contractor	Employed Full-Time	Employed Part-Time	Student	Unemployed
Not placed %	68.2	63.7	60.7	63.1	62.0
Placed %	31.8	36.3	39.3	36.9	38.0
population %	7.3	18.8	11.7	33.5	28.7

placed/employment_status



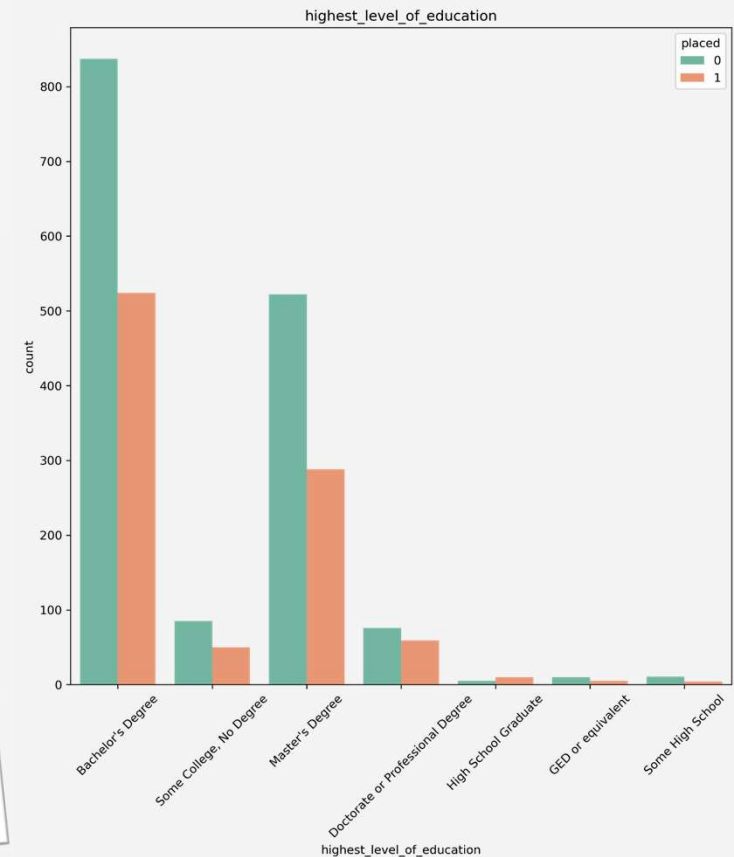
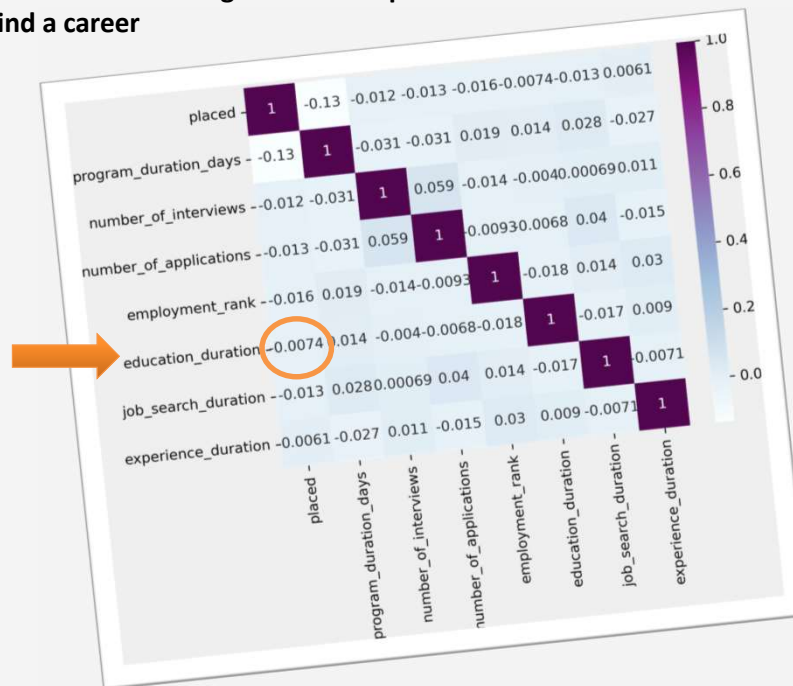
Although the students made up the majority of people participating in Pathrise Program, people with part time job and unemployed people had a bit more chance to find a job



Explorer in highest level of education

1. How much can the education level help people find a job?

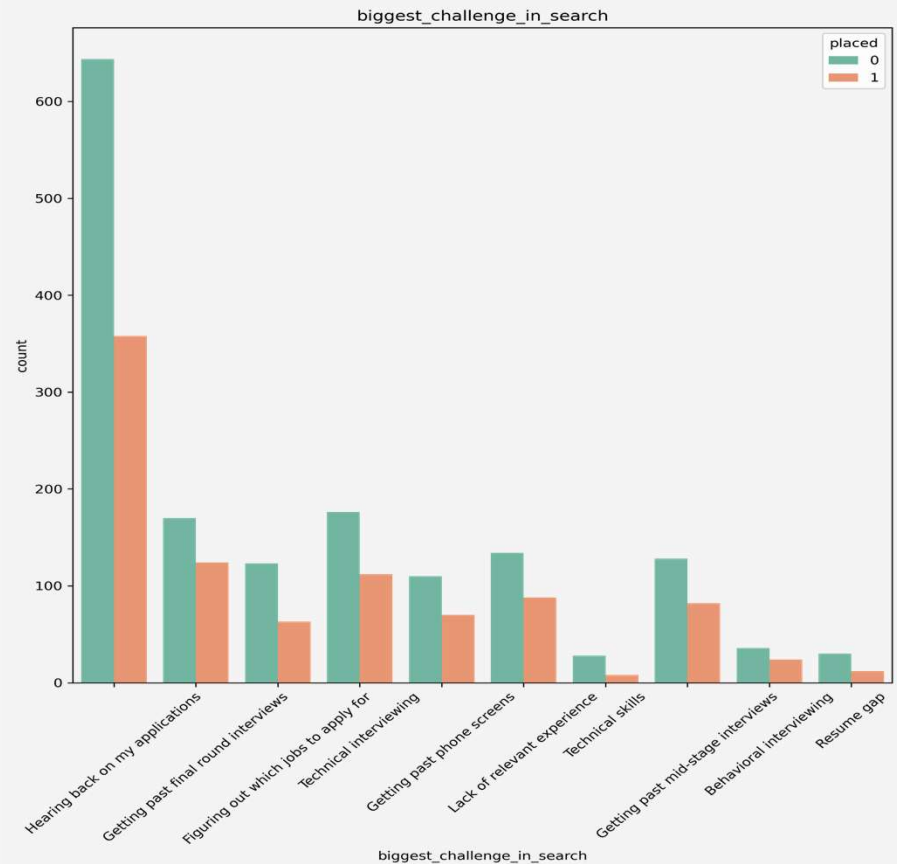
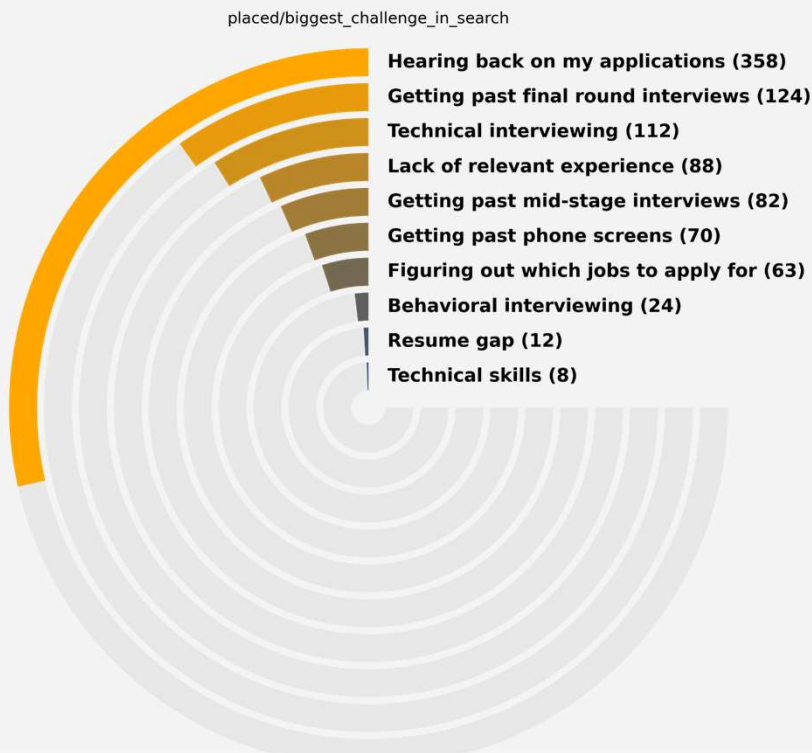
As is can be seen, different levels of education have not enough population to assess the influence of education level on the opportunity of finding a job. However, individuals with Bachelor's and master degree made up the most number of people being successful to find a career



Heat map correlation plot shows very small negative number -0.0074 correlation coefficient between level of education and placed

Explorer in biggest challenge in search

“Hearing back on my application” was the prevalent challenging issue for both groups (placed and not placed)

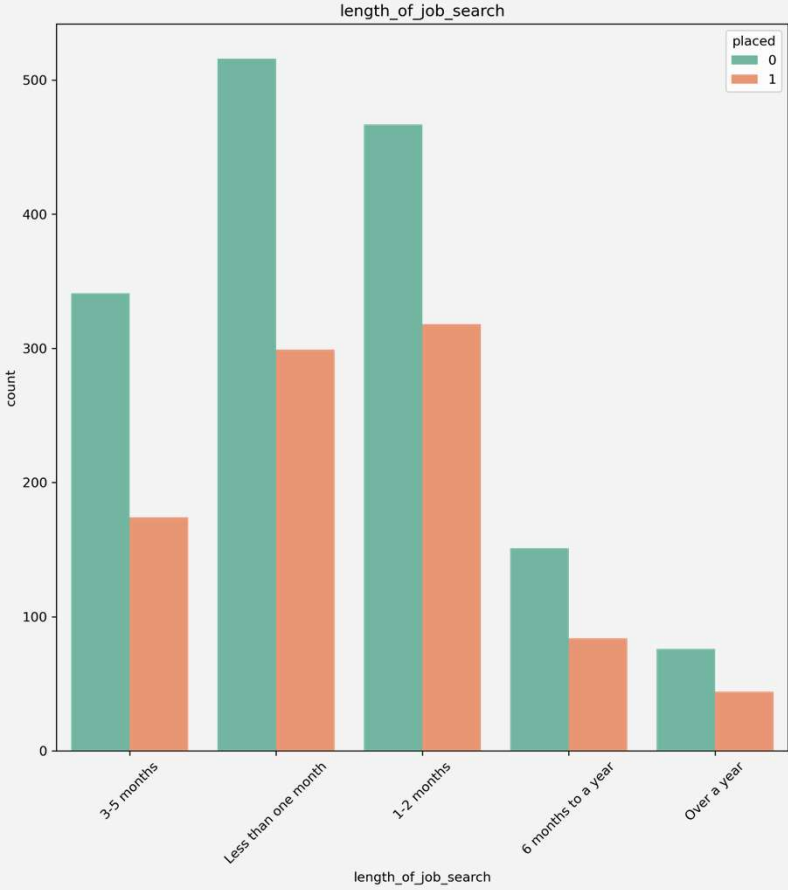
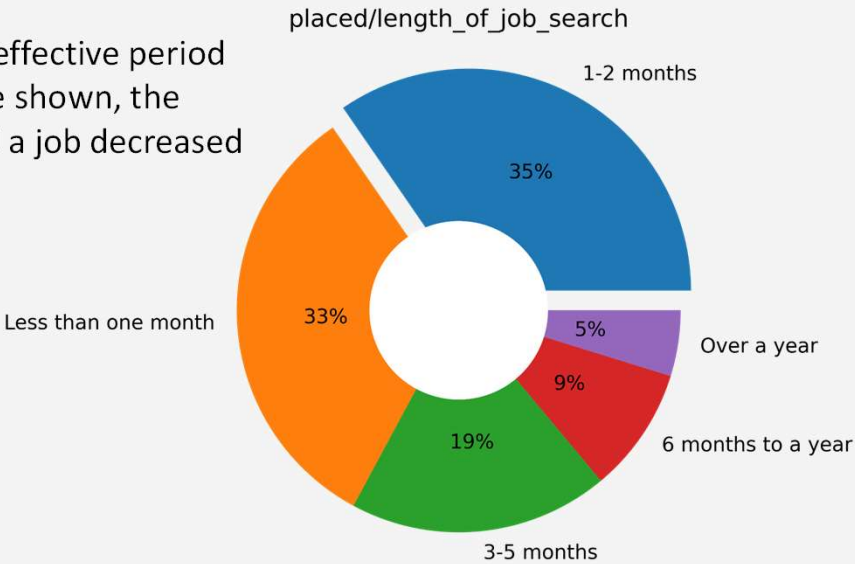


Explorer in length of job search

1. What is the most effective period of time to seek a job?

	1-2 months	3-5 months	6 months to a year	Less than one month	Over a year
Not placed %	59.5	66.2	64.3	63.3	63.3
Placed %	40.5	33.8	35.7	36.7	36.7
population %	31.8	20.9	9.5	33.0	4.9

1-2 months is the most effective period to find a job, As it can be shown, the chance of people to find a job decreased after this time

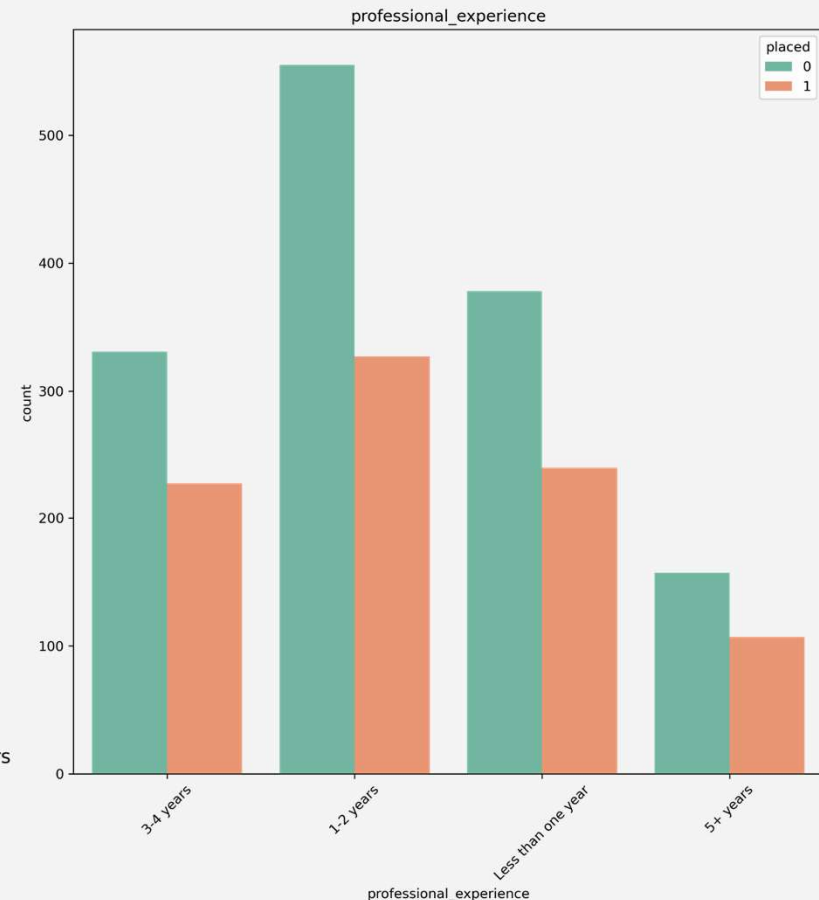
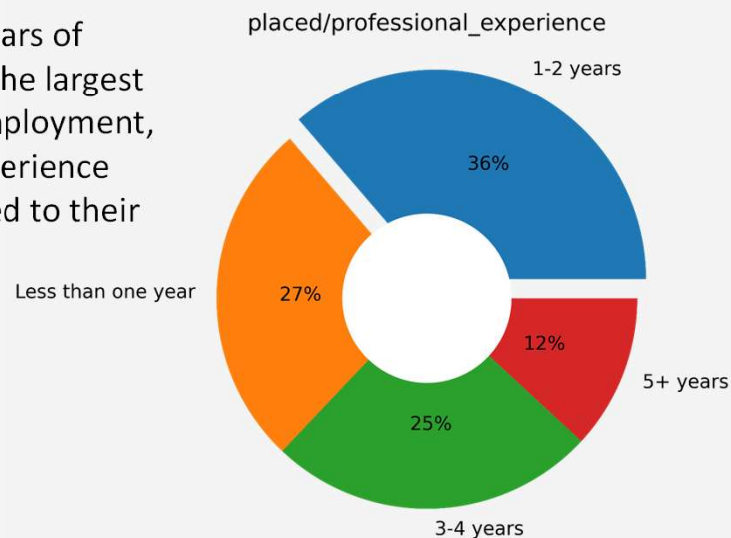


Explorer in professional experience

How much does people's professional experience help them find a job?

	Less than one year	1-2 years	3-4 years	5+ years
Not placed %	61.2	62.9	59.3	59.5
Placed %	38.8	37.1	40.7	40.5
population %	26.6	38.0	24.0	11.4

Although people with **1 to 2** years of professional experience were the largest group of people who found employment, people with **3 to 4** years of experience were more successful compared to their population

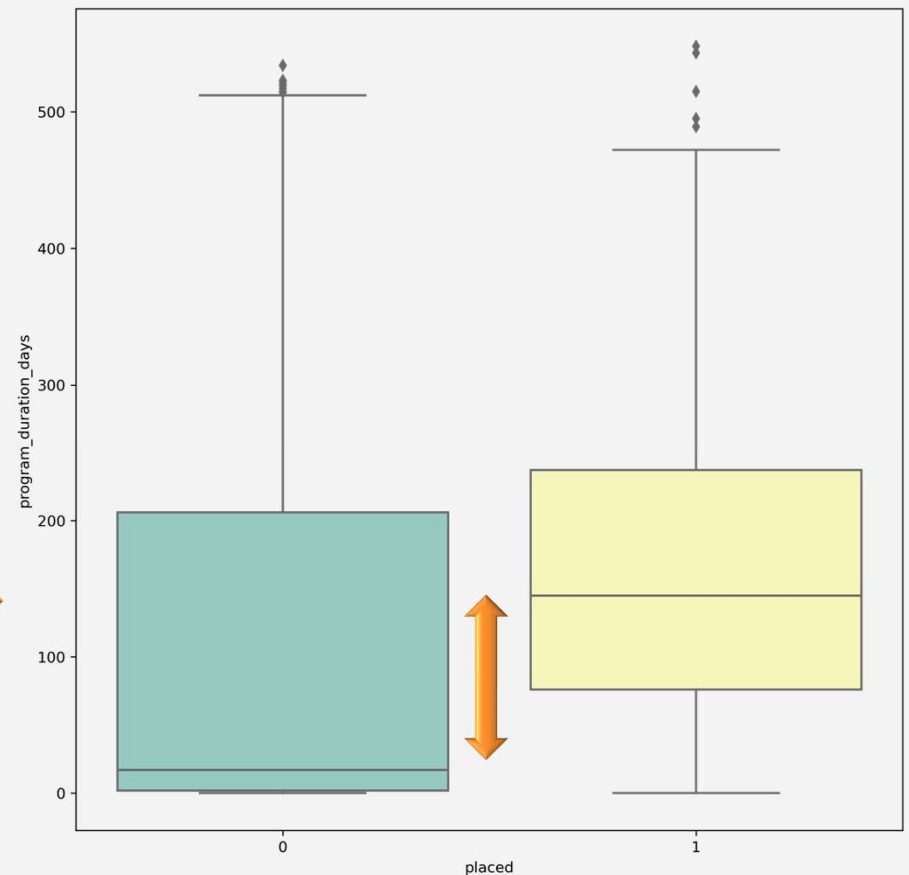


Explorer in program duration days

1. How much can the Pathrise program help the people find a job?

There is significant difference between average time which successful people and unsuccessful people spend on Pathrise Program

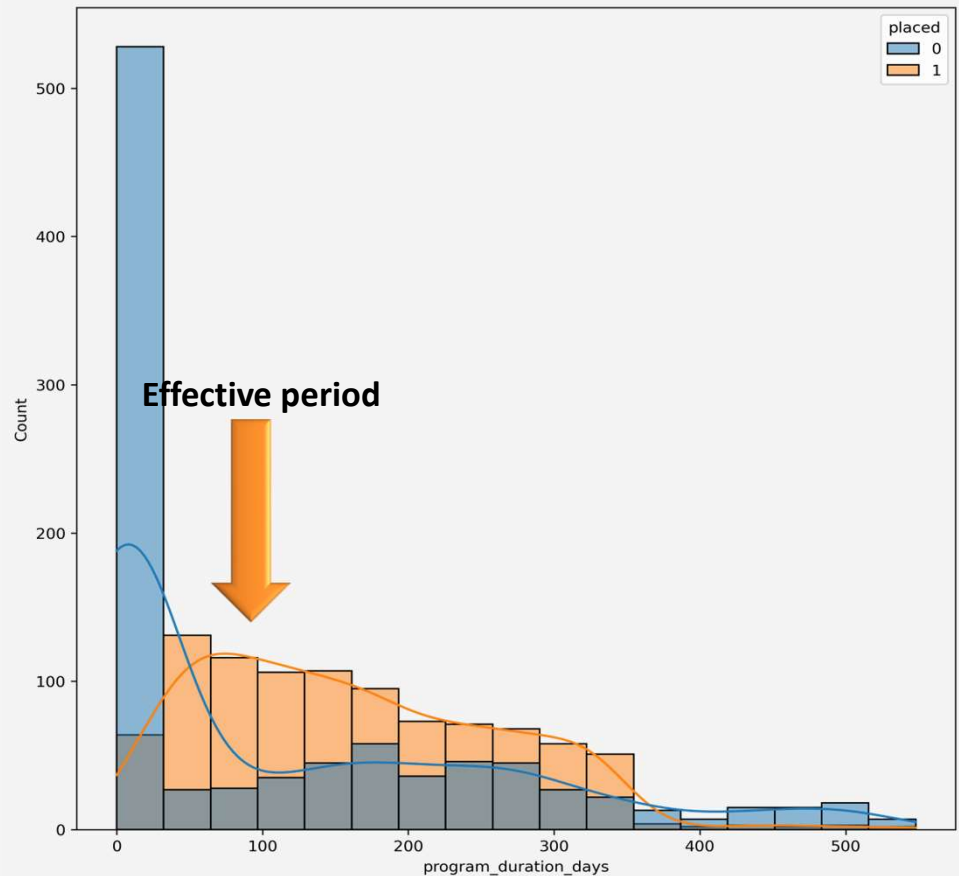
Mean differences between two groups shows Program effectiveness



Explorer in program duration days

2. How much should the people spend time on Pathrise program?

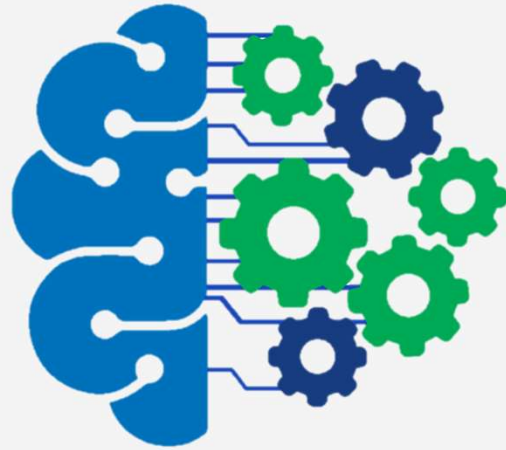
Effectiveness of program decreases after almost 100 days



Summary of EDA results

- Software engineering (**SWE**) was the most common primary track and people with this primary track was the most successful group.
- **Employed part-time**, unemployed and student People were the most group of people who find a job respectively
- There is insufficient evidence to show the relationship between **level of education** and opportunity of finding a job. However, correlation examination shows vary small negative correlation coefficient.
- “Hearing back on my application” was the prevalent **challenging** issue for both groups (placed and not placed)
- **1-2 months** is the most effective period to find a job
- Having **3 to 4 years professional** experience increase the chance of people a bit more to be successful in this program
- **Following the Pathrise program in 100 days** have remarkable effect to increase the opportunity of people to find a job.

Performing Machine learning

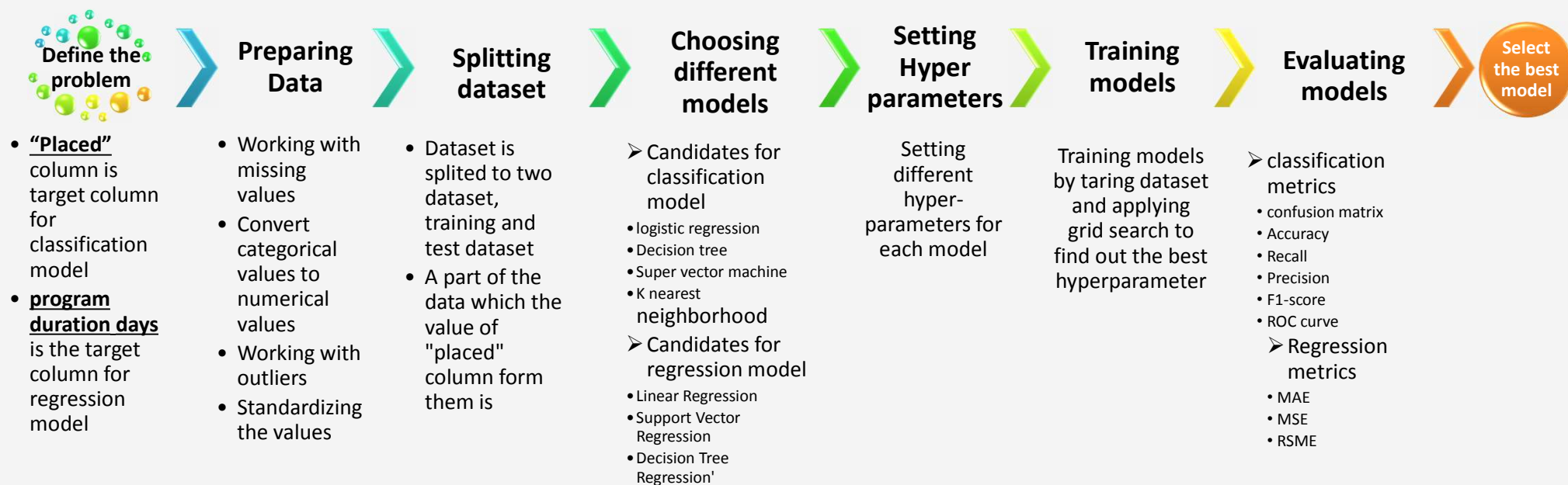


Machine learning



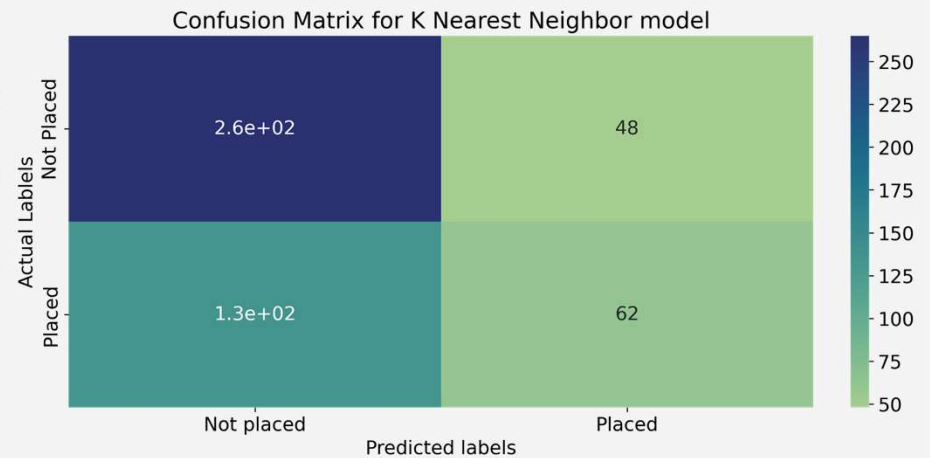
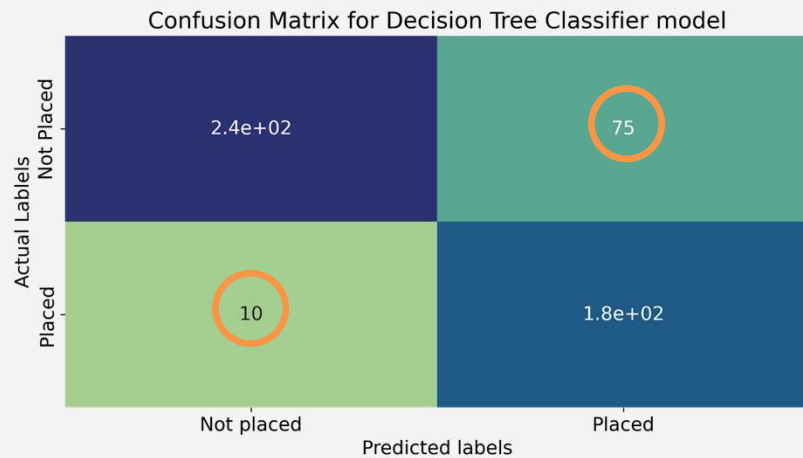
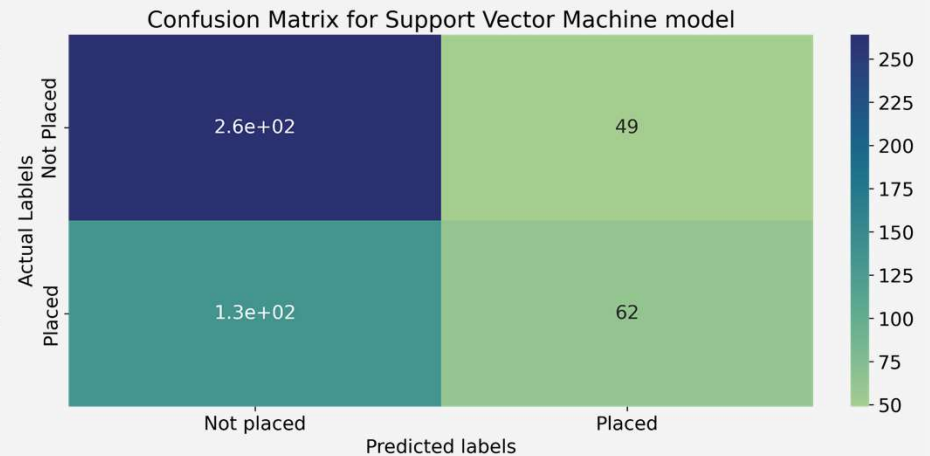
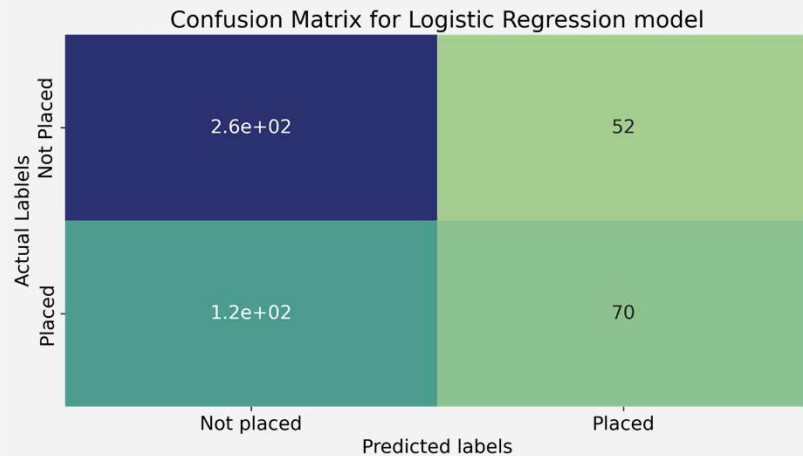
- ❖ There are two questions which are aimed to answer it by preparing supervised machine learning models
 - ❖ Preparing a model would be able to predicting whether or not someone participating in Pathrise program would be successful to find a job. This a classic **supervised classification machine learning**.
 - ❖ Training a model would be able to predicting how long a person participating in Pathrise program would find a job. This is classic **regression machine learning**

Process of preparing a machine learning model



Classification results

- Comparing results of models' confusion matrix

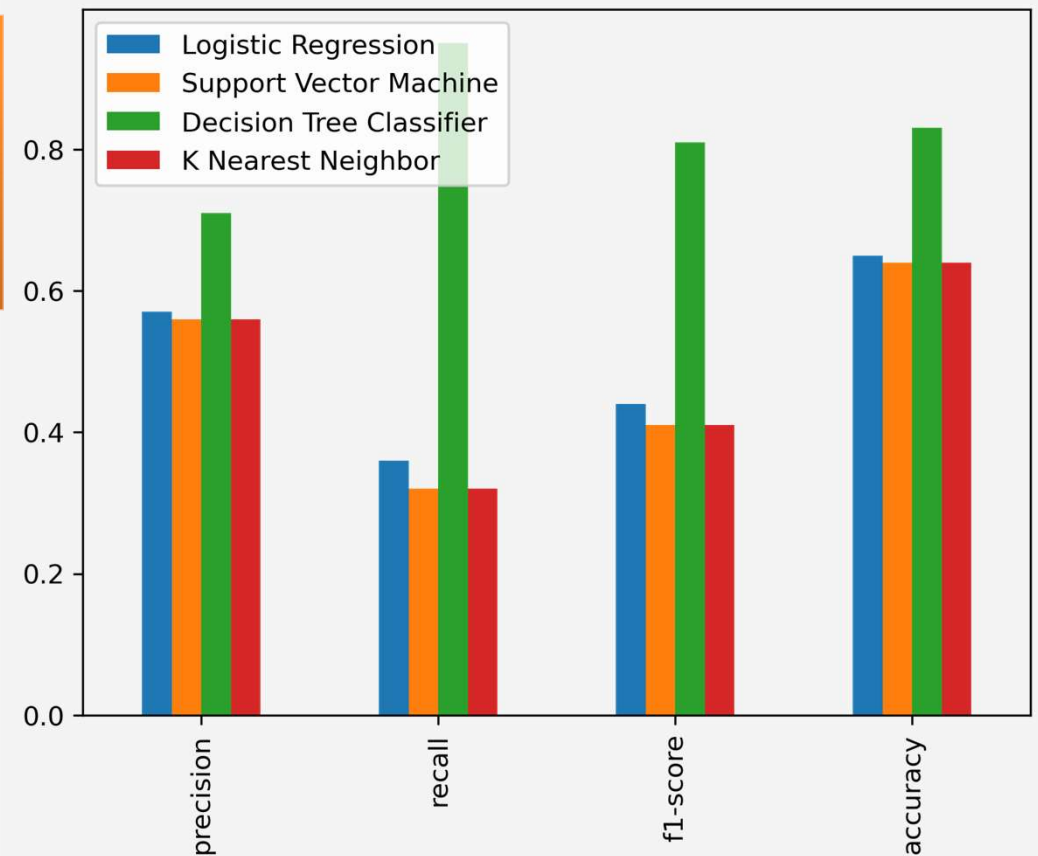


Classification results

- Comparison of different metrics for each model in “placed=1” state

	Logistic Regression	Support Vector Machine	Decision Tree Classifier	K Nearest Neighbor
precision	0.57	0.56	0.71	0.56
recall	0.36	0.32	0.95	0.32
f1-score	0.44	0.41	0.81	0.41
accuracy	0.65	0.64	0.83	0.64

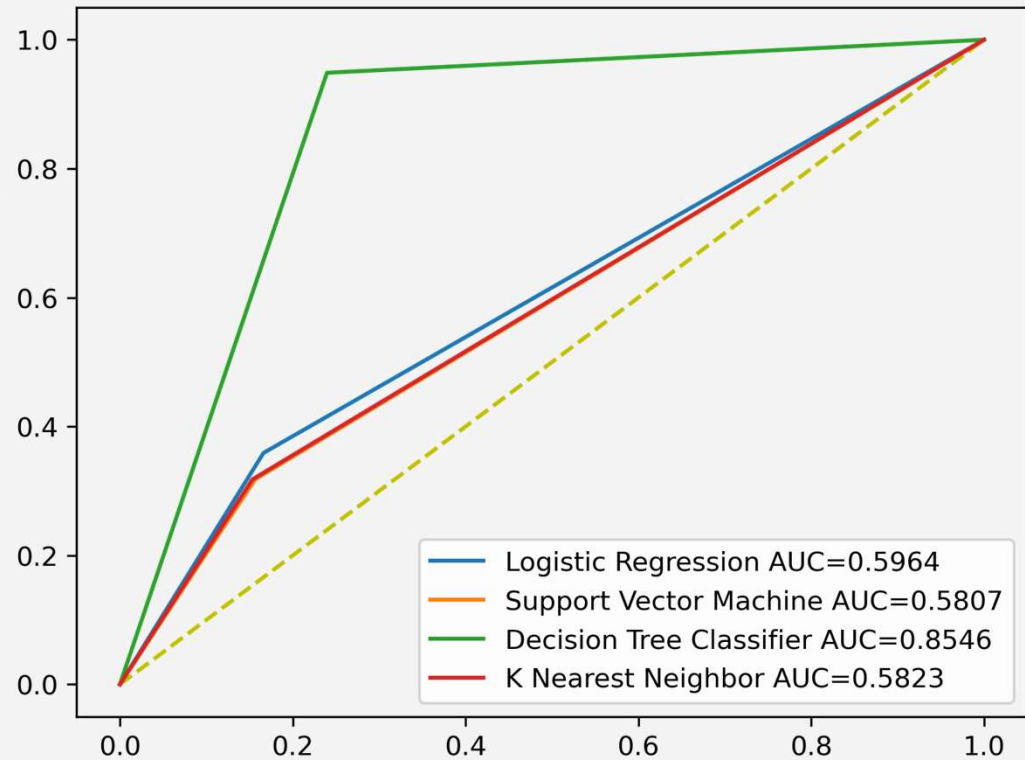
As it can be seen **decision tree** shows the least error and the best accuracy



Classification results

- Comparison of ROC curve for each model in “placed=1” state

The area under curve (AUC) for decision tree model is more than others and it shows better performance to predict and answer our question.



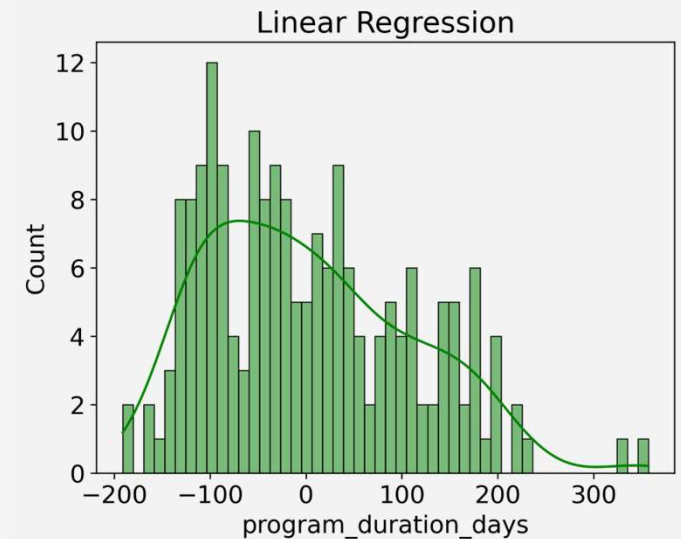
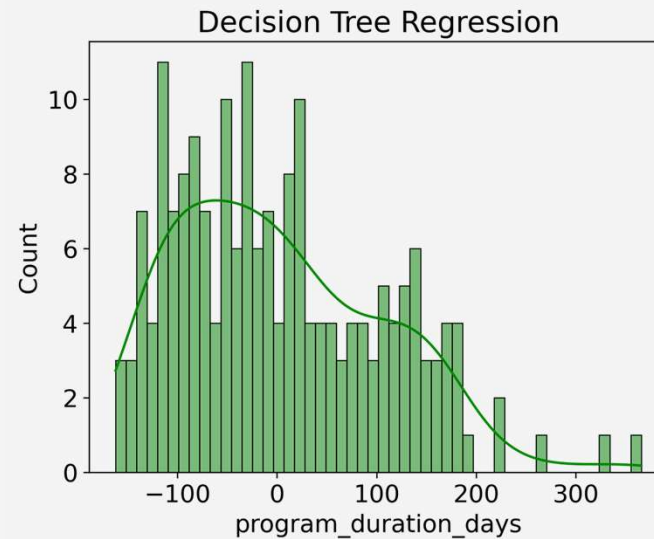
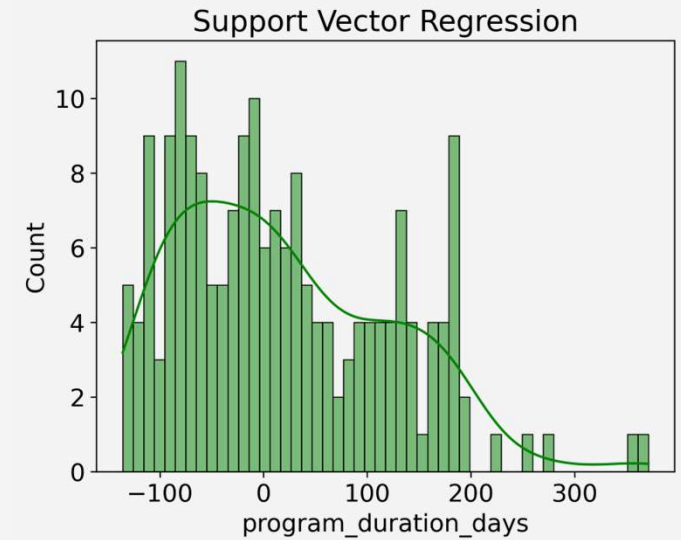


Regression results

- Comparison of residual distributions for different models

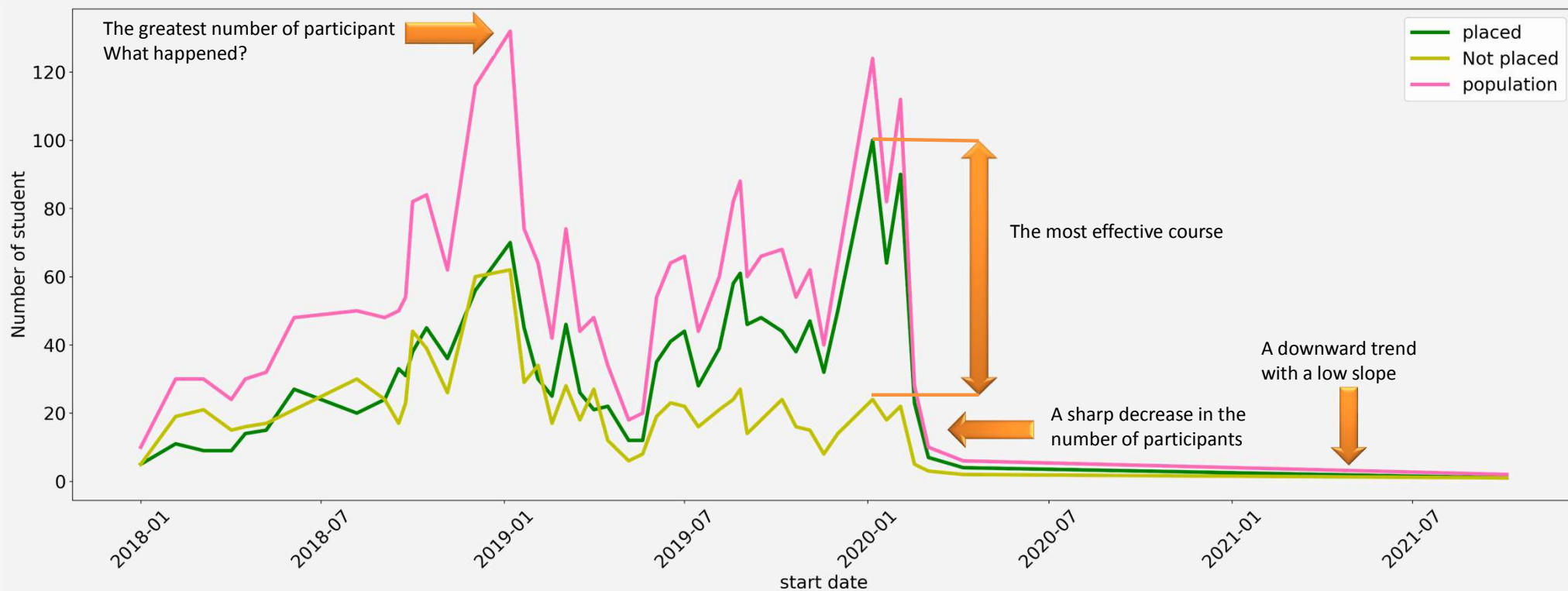
	Linear Regression	Decision Tree Regression	Support Vector Regression
MAE	87.7	85.5	83.7
MSE	11338	10846.8	10950
RMSE	106.5	104.1	104.6

Support vector regression model shows a bit better performance



Some proposals for future Research

Examining the number of different groups of people over time



Thanks For watching

For more information please see my GitHub

<https://github.com/Rezassp/Pathrise>

