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.

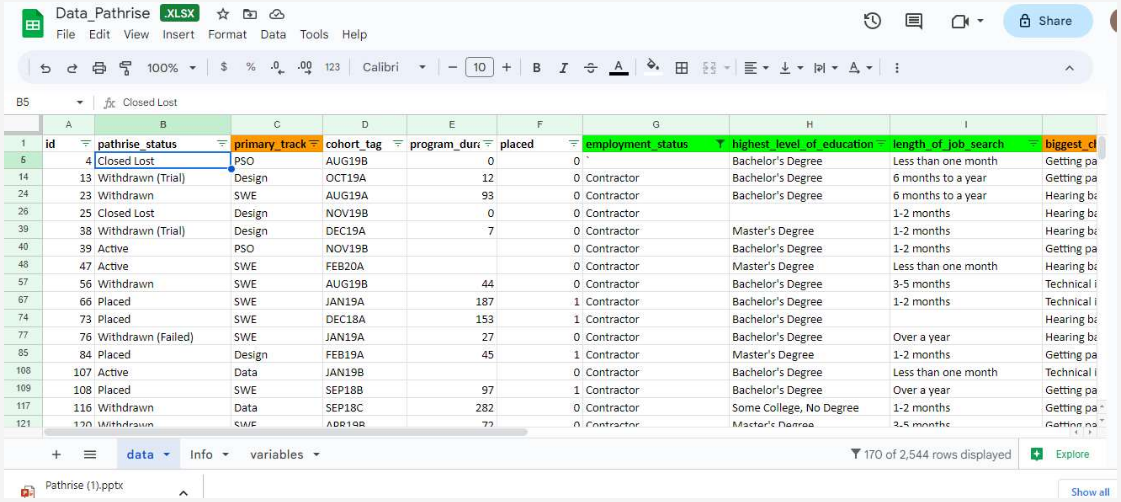
1. Preparing a model to predict whether people would find a job or not ?
2. Preparing a model to predict how long does it take to find a job?

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



	A	B	C	D	E	F	G	H	I	
1	id	pathrise_status	primary_track	cohort_tag	program_dur	placed	employment_status	highest_level_of_education	length_of_job_search	biggest_cl
5	4	Closed Lost	PSO	AUG19B	0	0	`	Bachelor's Degree	Less than one month	Getting pa
14	13	Withdrawn (Trial)	Design	OCT19A	12	0	Contractor	Bachelor's Degree	6 months to a year	Getting pa
24	23	Withdrawn	SWE	AUG19A	93	0	Contractor	Bachelor's Degree	6 months to a year	Hearing ba
26	25	Closed Lost	Design	NOV19B	0	0	Contractor		1-2 months	Hearing ba
39	38	Withdrawn (Trial)	Design	DEC19A	7	0	Contractor	Master's Degree	1-2 months	Hearing ba
40	39	Active	PSO	NOV19B		0	Contractor	Bachelor's Degree	1-2 months	Getting pa
48	47	Active	SWE	FEB20A		0	Contractor	Master's Degree	Less than one month	Hearing ba
57	56	Withdrawn	SWE	AUG19B	44	0	Contractor	Bachelor's Degree	3-5 months	Technical i
67	66	Placed	SWE	JAN19A	187	1	Contractor	Bachelor's Degree	1-2 months	Technical i
74	73	Placed	SWE	DEC18A	153	1	Contractor	Bachelor's Degree		Hearing ba
77	76	Withdrawn (Failed)	SWE	JAN19A	27	0	Contractor	Bachelor's Degree	Over a year	Hearing ba
85	84	Placed	Design	FEB19A	45	1	Contractor	Master's Degree	1-2 months	Getting pa
108	107	Active	Data	JAN19B		0	Contractor	Bachelor's Degree	Less than one month	Technical i
109	108	Placed	SWE	SEP18B	97	1	Contractor	Bachelor's Degree	Over a year	Getting pa
117	116	Withdrawn	Data	SEP18C	282	0	Contractor	Some College, No Degree	1-2 months	Getting pa
121	120	Withdrawn	SWE	APR19B	72	0	Contractor	Master's Degree	3-5 months	Getting pa

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%	-

Data preparation approach

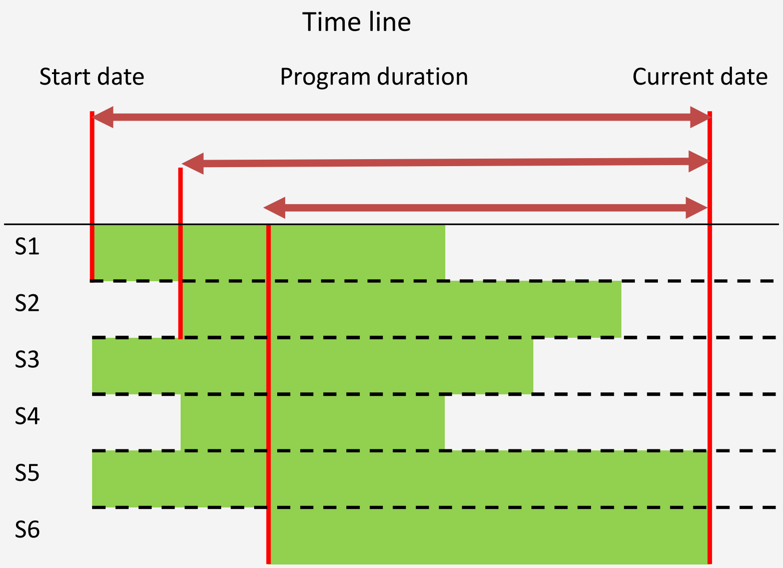
Remove people who did not get involved in Pathrise program

pathrise_status	status of a fellow in the program
Break	on a temporary break
Closed Lost	didn't accept our offer
Deferred	accepted our offer, but willing to start later
MIA	missed in action - joined the program, but stopped being involved

According to variable definition for pathrise_status these people did not participate in Pathrise Program actually

Pathrise project

How to calculate program duration for missing values



*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**

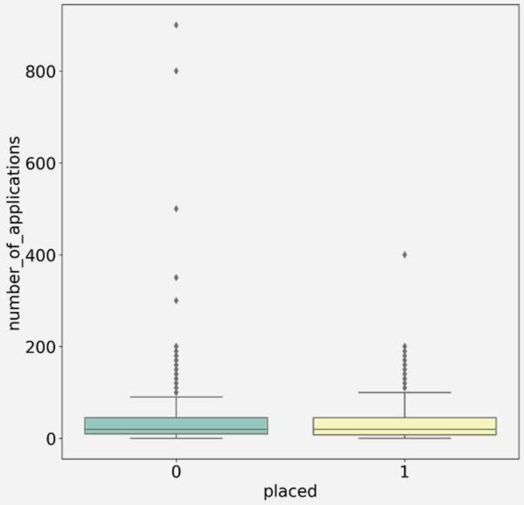
Calculating Program duration for current student

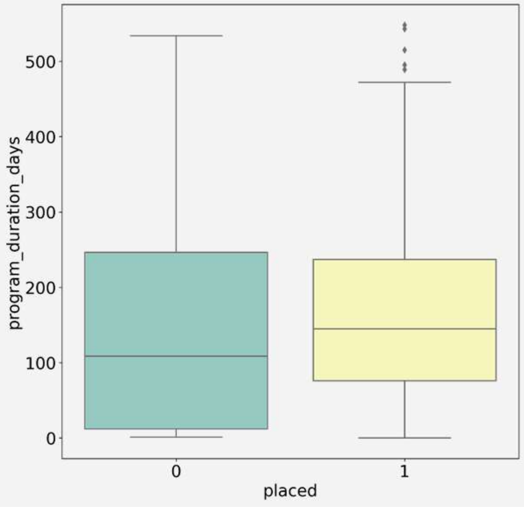
Difference between start program date and current date is considered as program duration days for **current student**

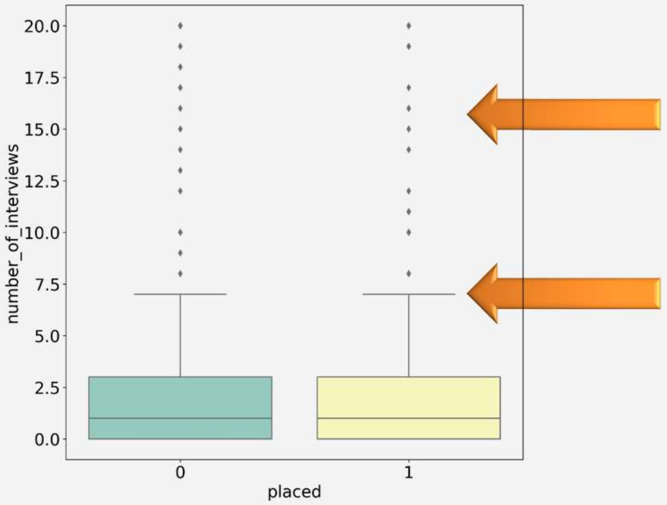
Pathrise project

- **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. **<u>Interquartile range(IQR)</u>** is used to indicate the outlier
3. **Outliers are replaced by <u>mean values</u>.**



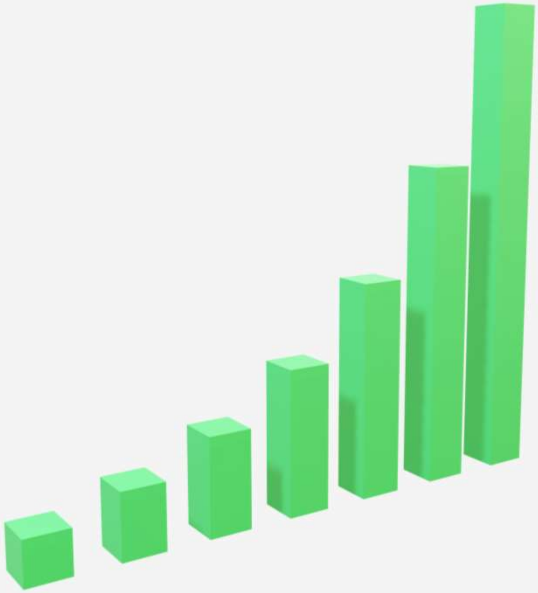




Outliers

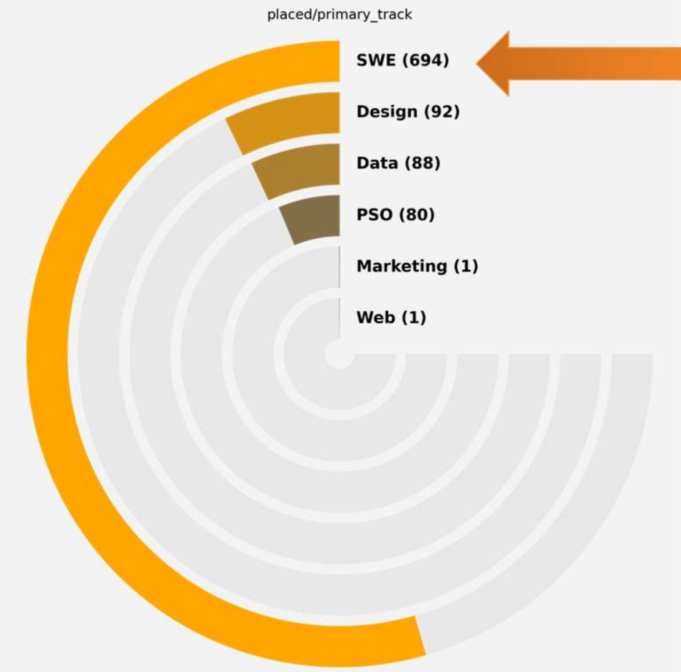
Upper limit Q3+1.5IQR

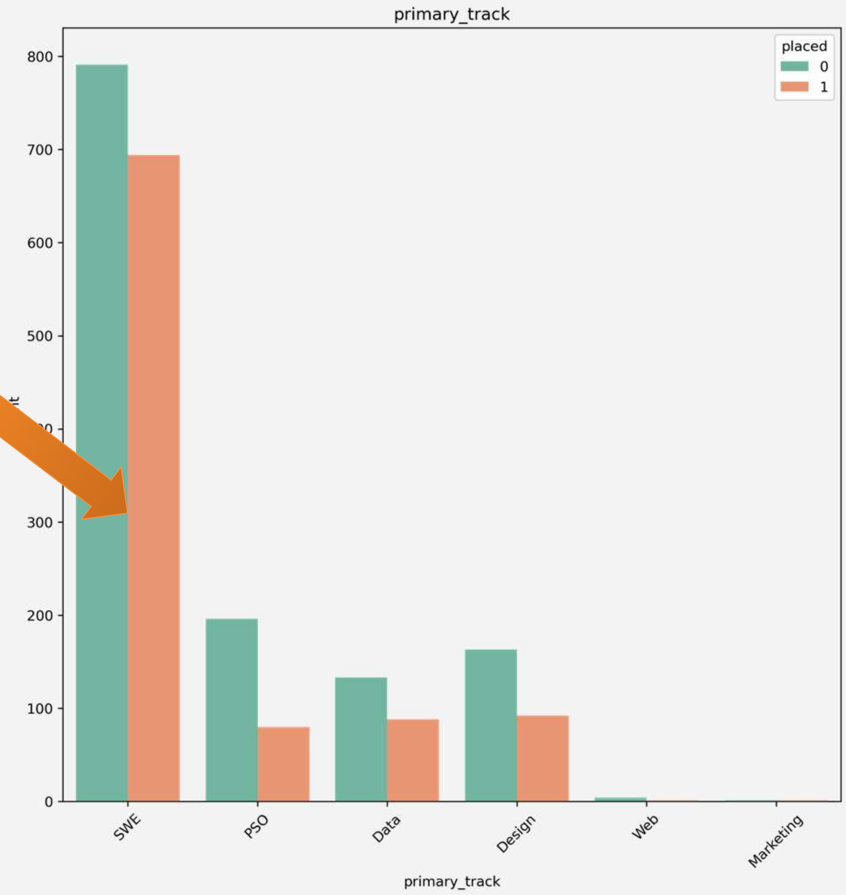
Exploratory data analysis(EDA)



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

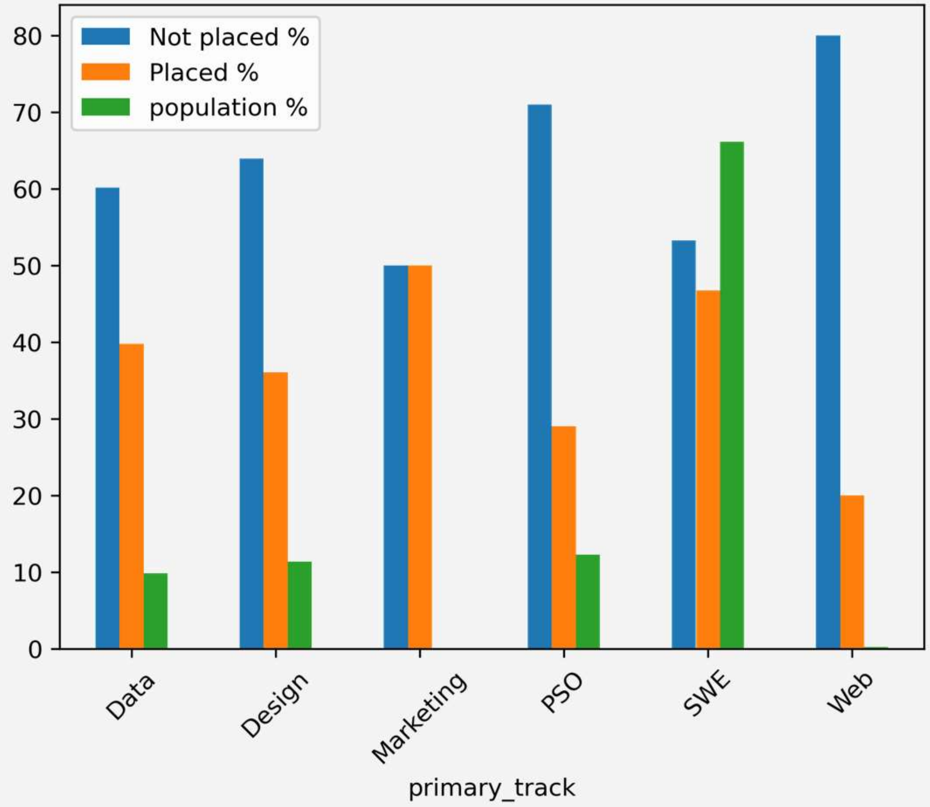
The percentage of people being successful to find a job changes if the they are compared according to their population

Most successful

	Data	Design	Marketing	PSO	SWE	Web
Not placed %	60.2	63.9	50.0	71.0	53.3	80.0
Placed %	39.8	36.1	50.0	29.0	46.7	20.0
population %	9.8	11.4	0.1	12.3	66.2	0.2

Most popular

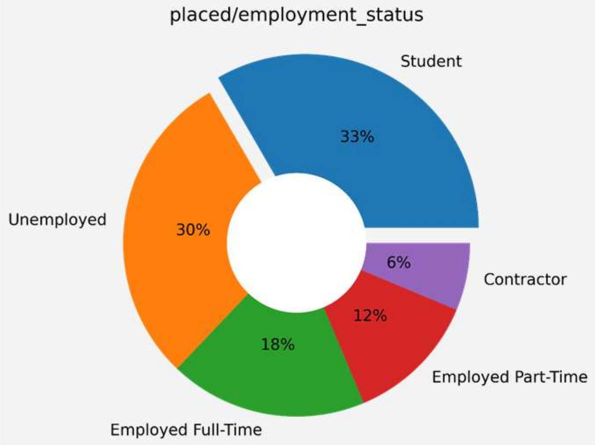
The number of people whose primary track is marketing only two, so this group should be ignored as a insufficient evidence



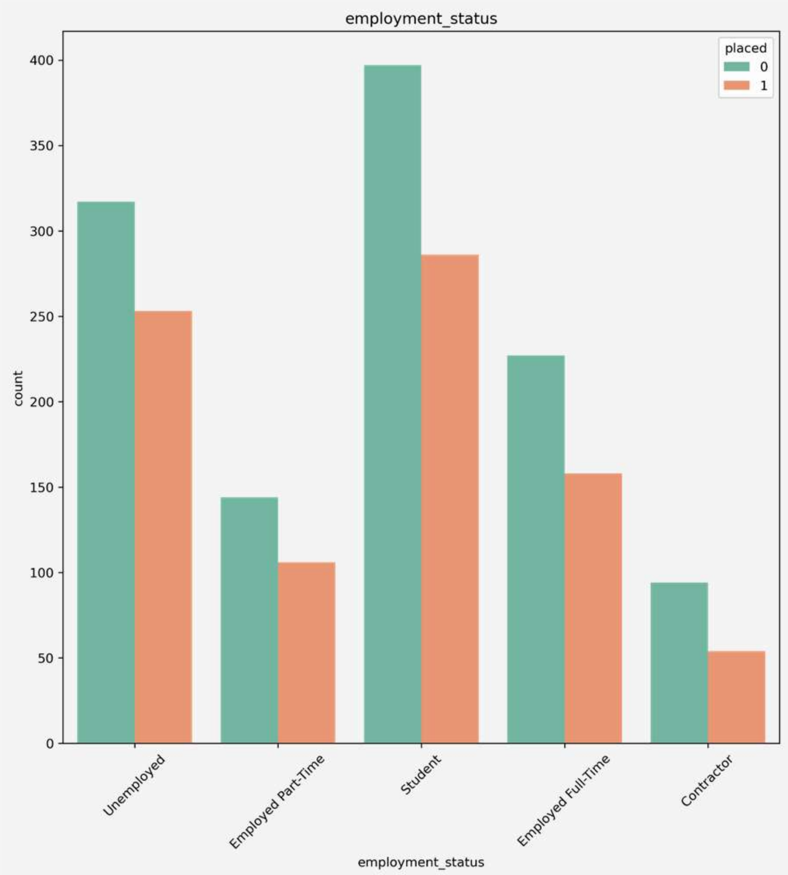
Explorer in employment status

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

	Contractor	Employed Full-Time	Employed Part-Time	Student	Unemployed
Not placed %	63.5	59.0	57.6	58.1	55.6
Placed %	36.5	41.0	42.4	41.9	44.4
population %	7.3	18.9	12.3	33.5	28.0



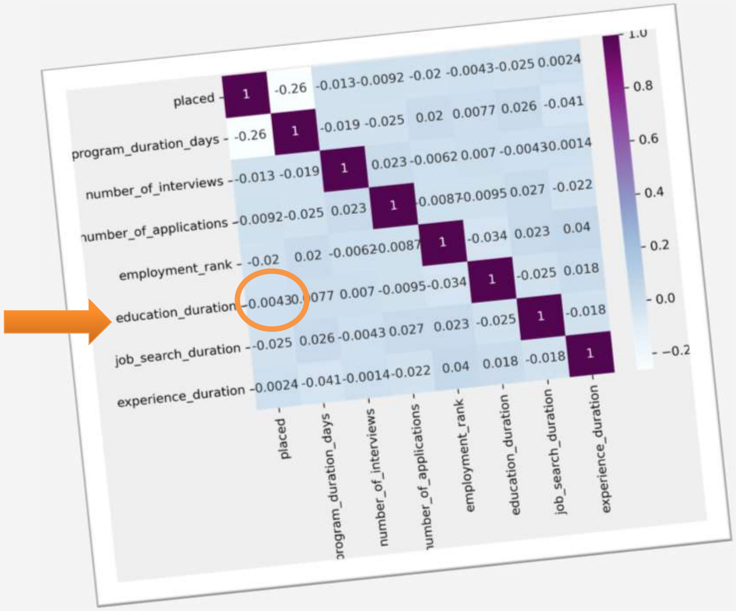
Although the **students** made up the majority of people participating in Pathrise Program, people with **part time job** and **unemployed** people had a bit more chance to find a job



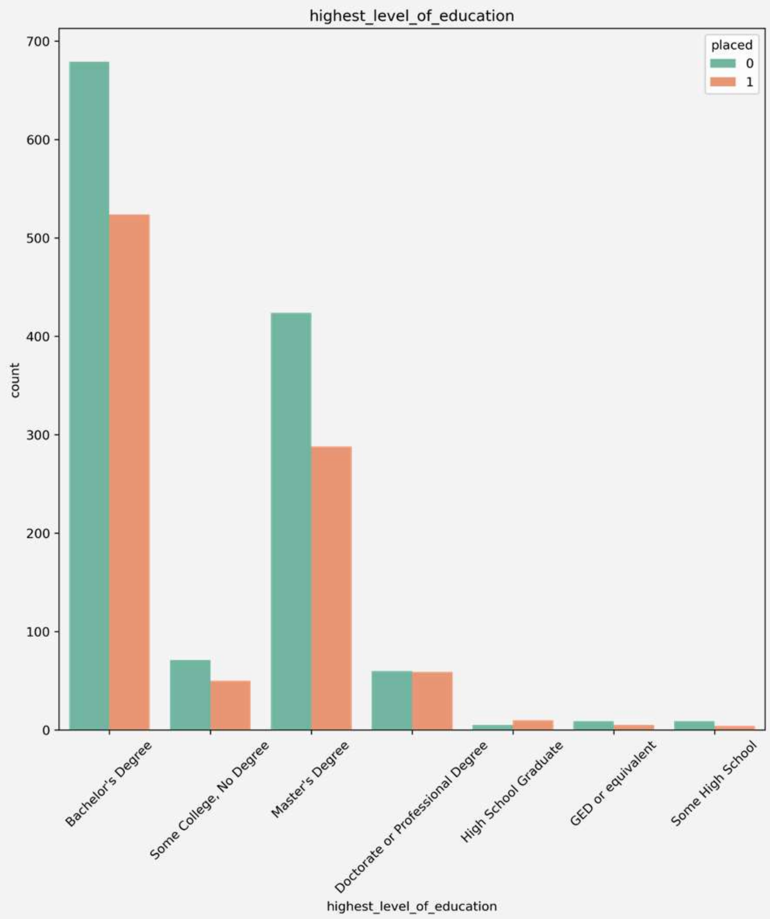
Explorer in highest level of education

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

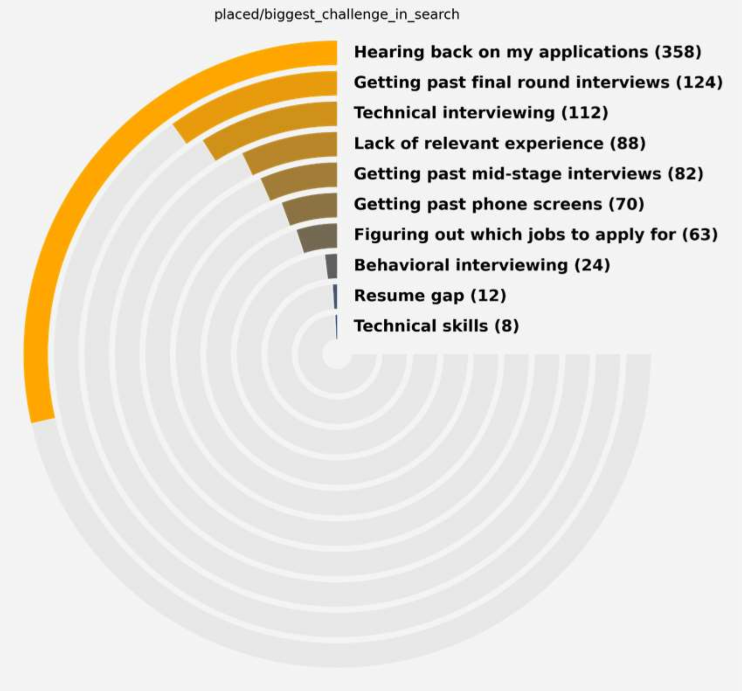


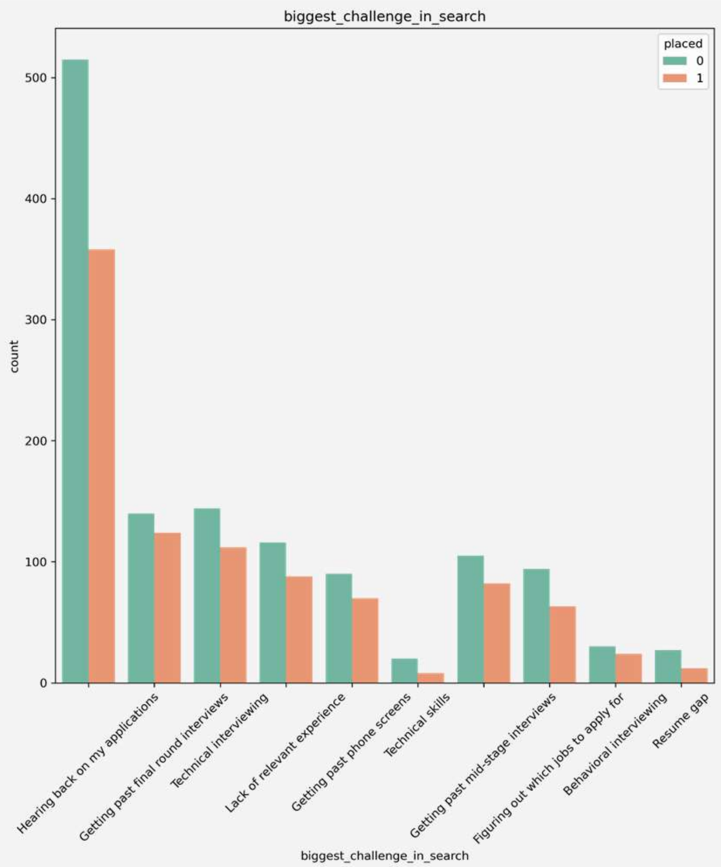
Heat map correlation plot shows very small negative number -0.0043 correlation coefficient between level of education and placed



Explorer in biggest challenge in search

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



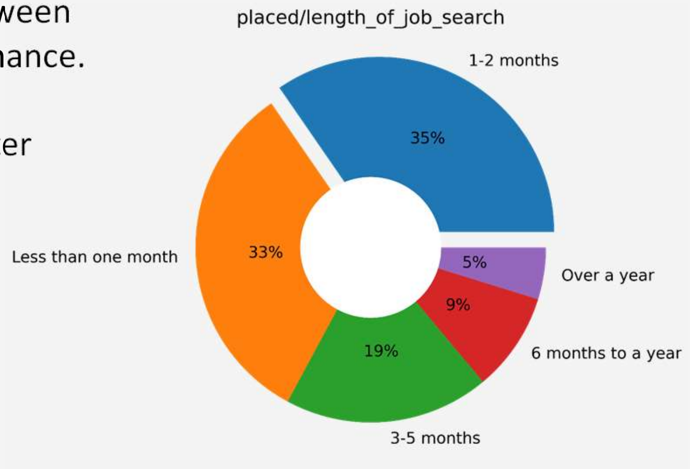


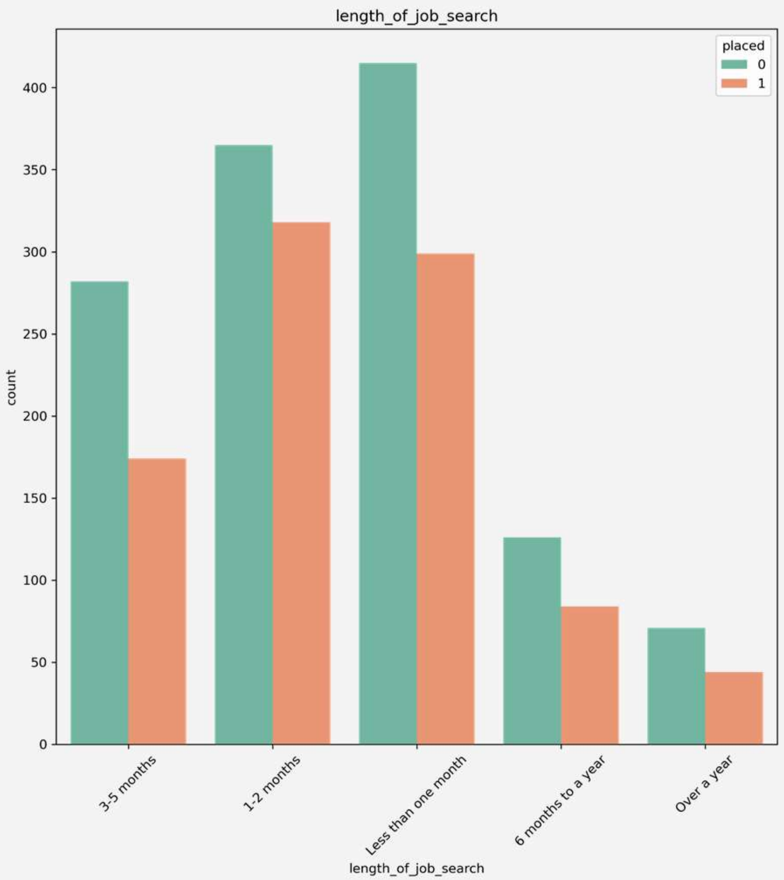
Explorer in length of job search

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 %	53.4	61.8	60.0	58.1	61.7
Placed %	46.6	38.2	40.0	41.9	38.3
population %	31.4	20.9	9.6	32.8	5.3

People who searched the job between **1 to 2 months** had better performance. As it can be shown, the chance of people to find a job decreased after this time



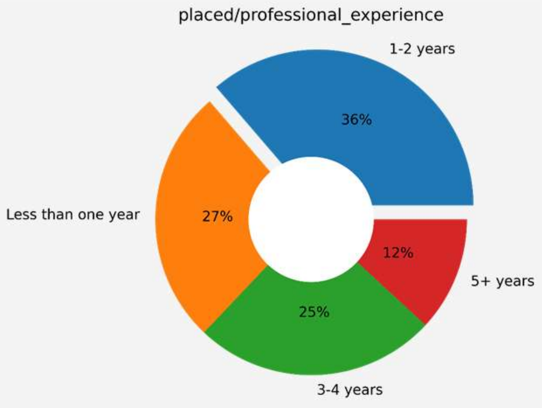


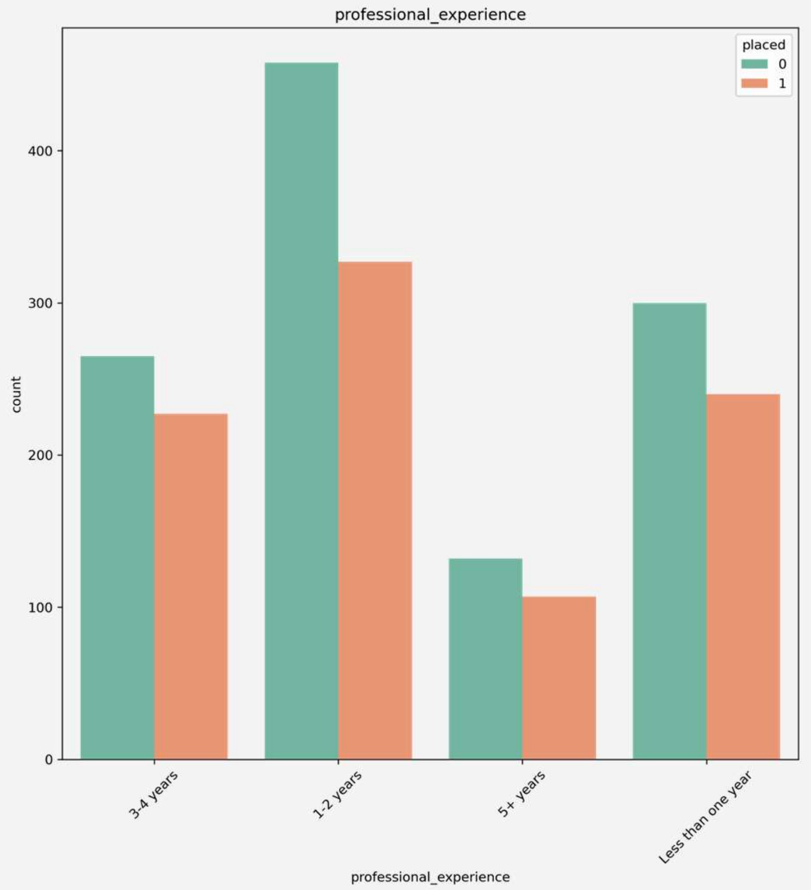
Explorer in professional experience

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

	1-2 years	3-4 years	5+ years	Less than one year
Not placed %	58.3	53.9	55.2	55.6
Placed %	41.7	46.1	44.8	44.4
population %	38.2	23.9	11.6	26.3

Although people with **1 to 2** years of professional experience were the largest group of people who found employment, people with more than **5 years** of experience or **less then one year** were more successful compared to their population



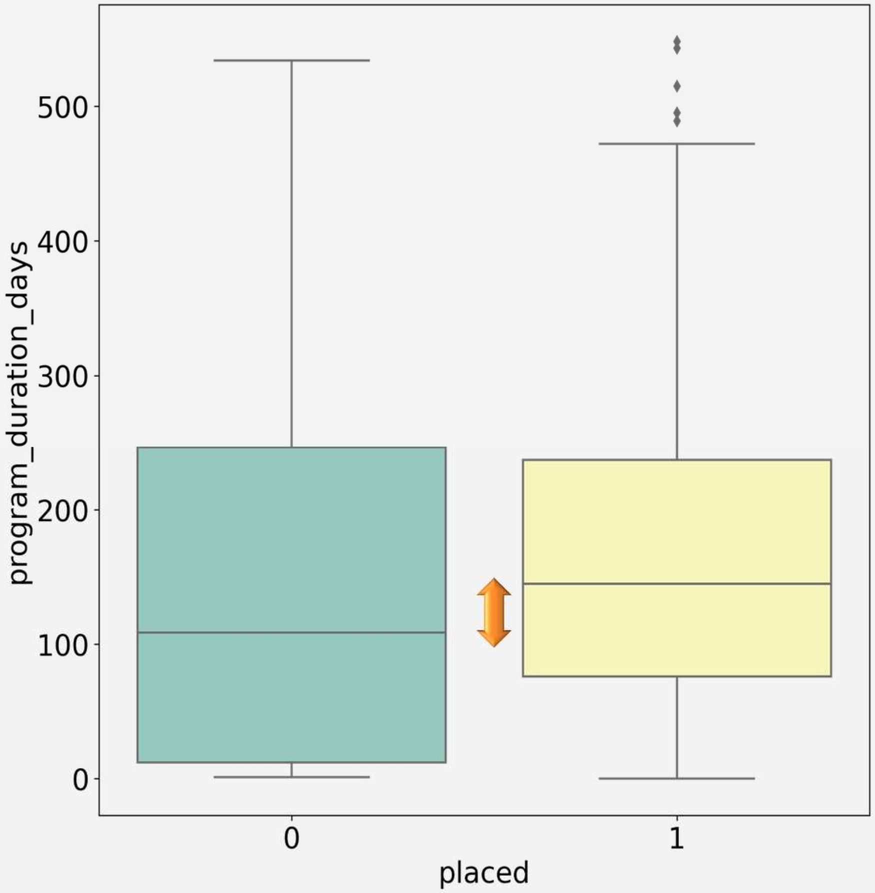


Explorer in program duration days

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

There is no meaningful difference between average time which successful people and unsuccessful people spend on Pathrise Program

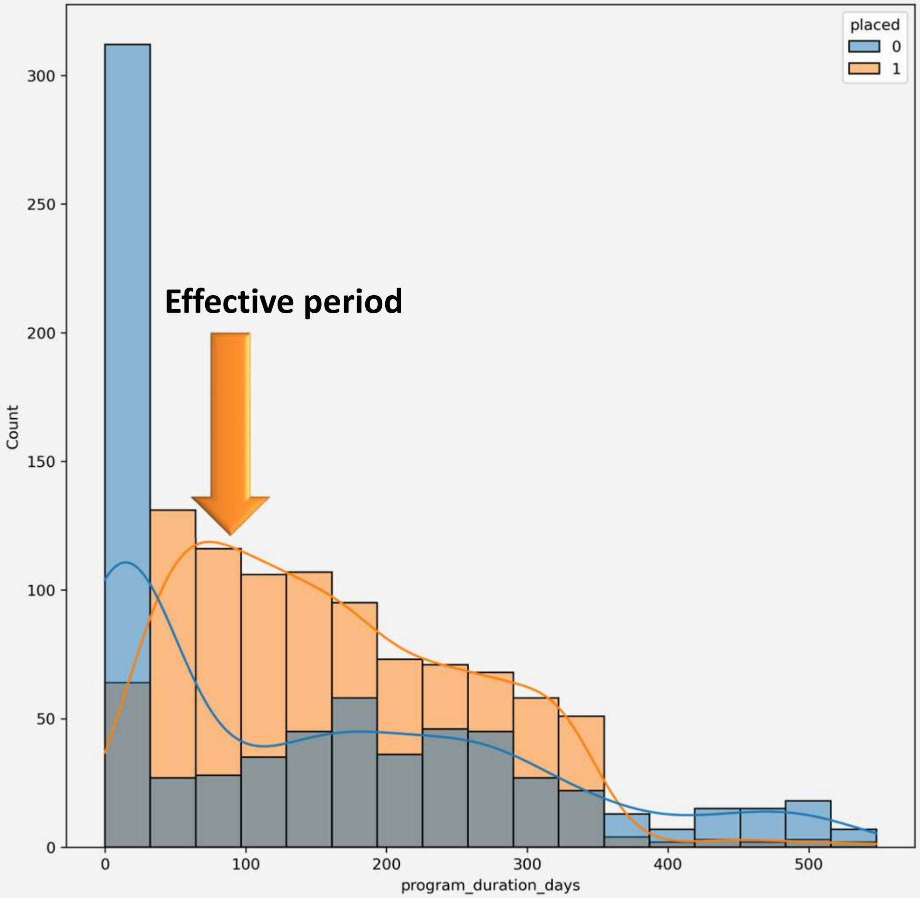
Mean differences between two groups shows Program duration is not critical factor



Explorer in program duration days

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

Effectiveness of program decreases after almost **<u>100 days</u>**

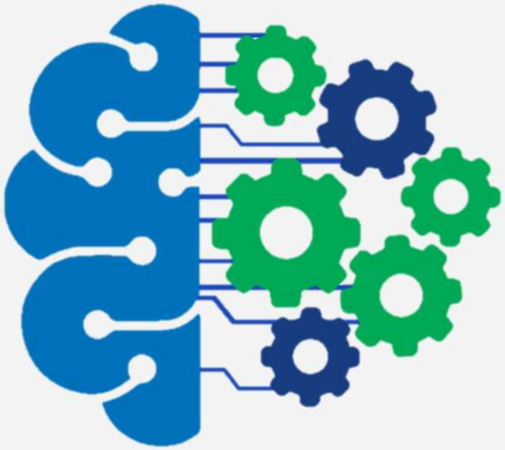


Effective period

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.

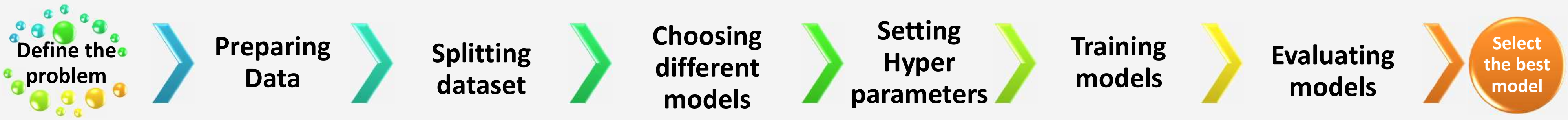
Performing Machine learning



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



Define the problem

- **“Placed”** column is target column for classification model
- **program duration days** is the target column for regression model

Preparing Data

- Working with missing values
- Convert categorical values to numerical values
- Working with outliers
- Standardizing the values

Splitting dataset

- Dataset is splited to two dataset, training and test dataset
- A part of the data which the value of "placed" column form them is

Choosing different models

➢ Candidates for classification model

- logistic regression
- Decision tree
- Super vector machine
- K nearest neighborhood

➢ Candidates for regression model

- Linear Regression
- Support Vector Regression
- Decision Tree Regression'

Setting Hyper parameters

Setting different hyper-parameters for each model

Training models

Training models by taring dataset and applying grid search to find out the best hyperparameter

Evaluating models

➢ classification metrics

- confusion matrix
- Accuracy
- Recall
- Precision
- F1-score
- ROC curve

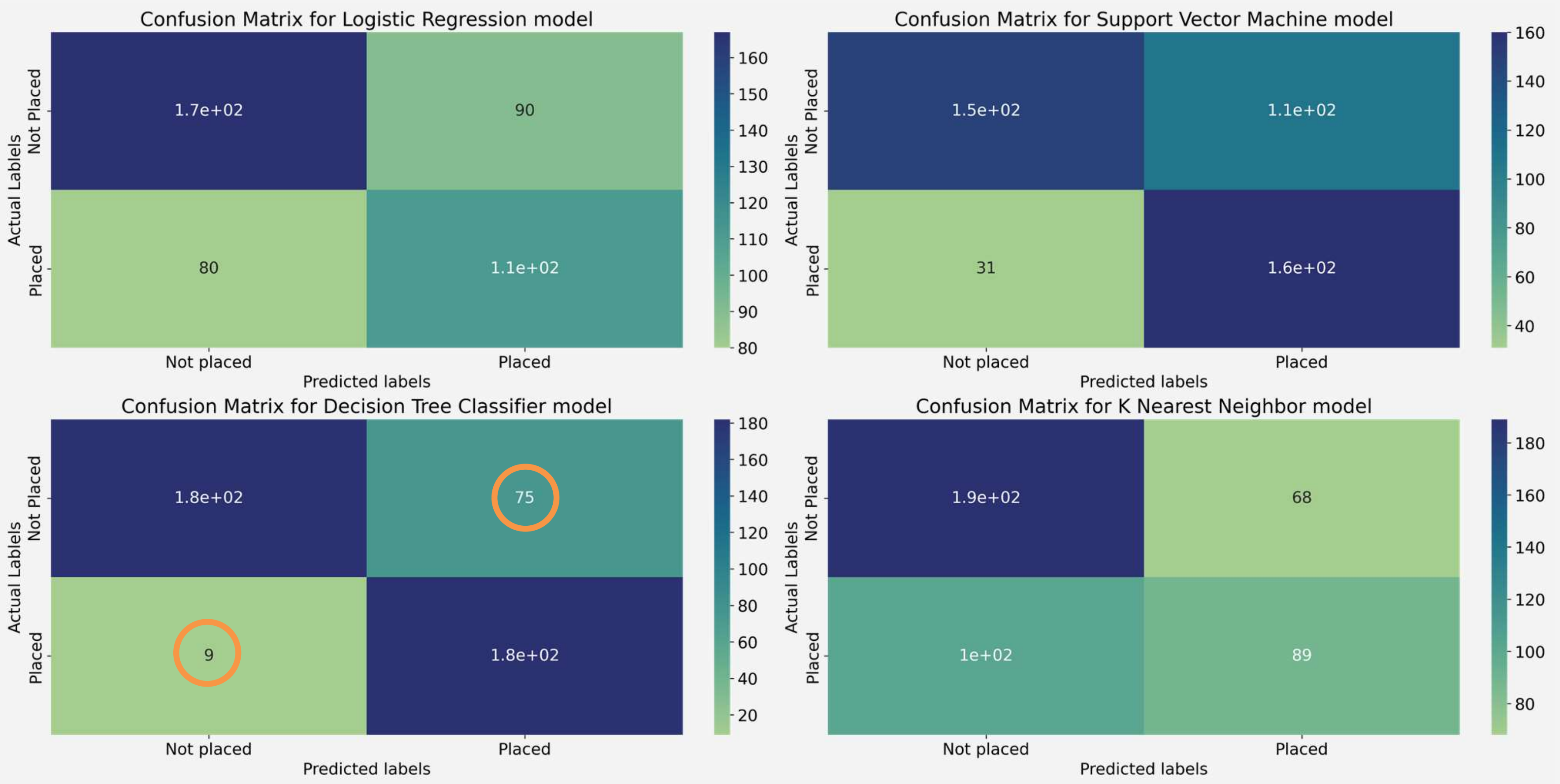
➢ Regression metrics

- MAE
- MSE
- RSME

Select the best model

Classification results

- Comparing results of models' confusion matrix



Confusion Matrix for Logistic Regression model

Confusion Matrix for Support Vector Machine model

Confusion Matrix for Decision Tree Classifier model

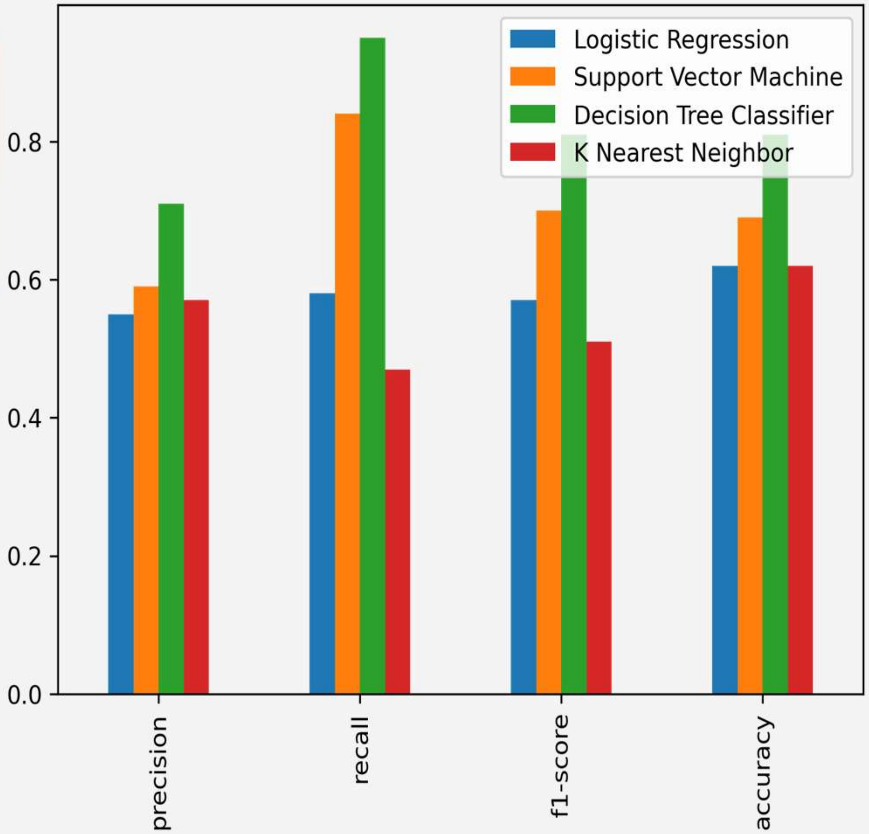
Confusion Matrix for K Nearest Neighbor model

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	0.55	0.59	0.71	0.57
recall	0.58	0.84	0.95	0.47
f1-score	0.57	0.7	0.81	0.51
accuracy	0.62	0.69	0.81	0.62

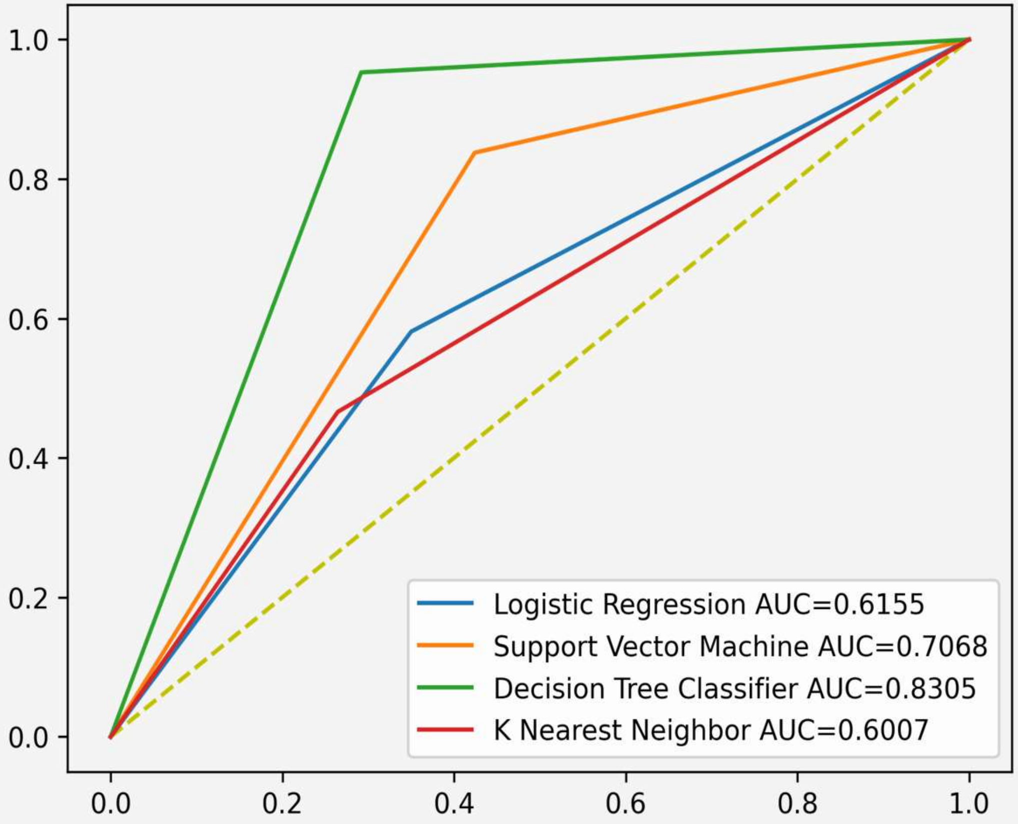
As it can be seen **decision tree** shows the least error and the best accuracy



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.

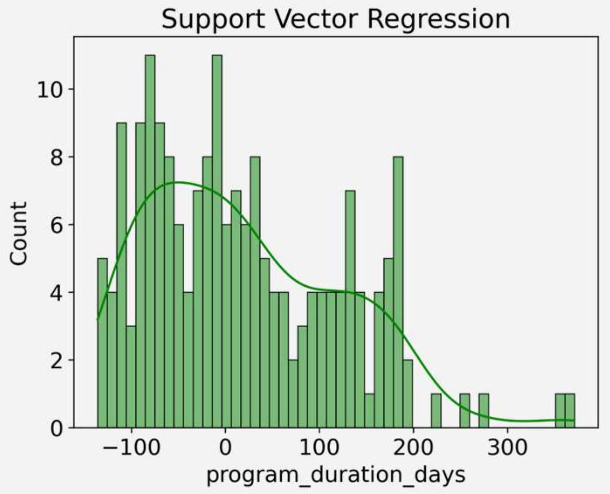


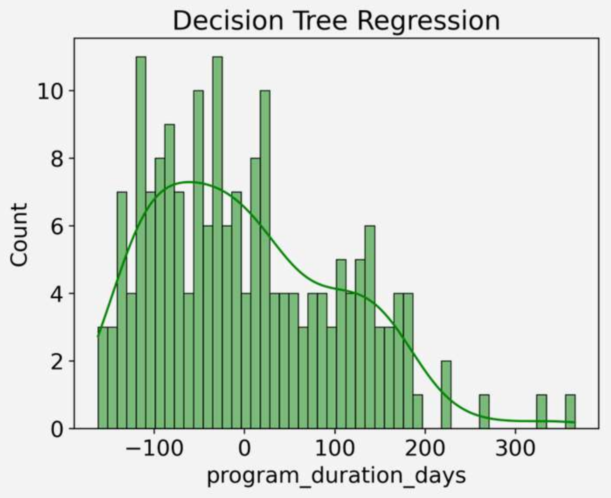
Regression results

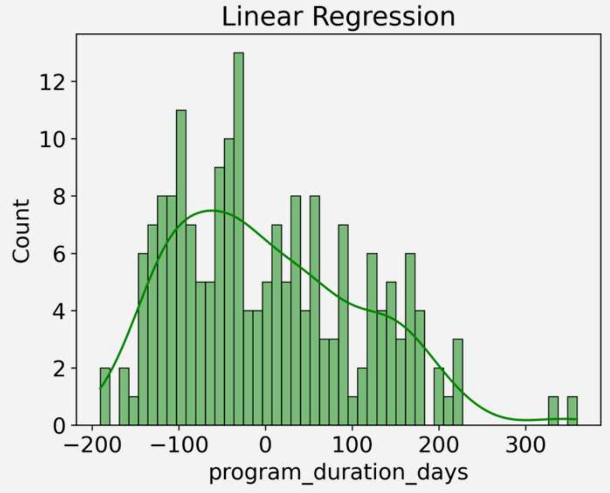
- Comparison of residual distributions for different models

	Linear Regression	Decision Tree Regression	Support Vector Regression
MAE	88.1	85.5	83.8
MSE	11,296.5	10,846.8	10,949.2
RMSE	106.3	104.1	104.6

<u>Support vector regression</u> model shows a bit better performance

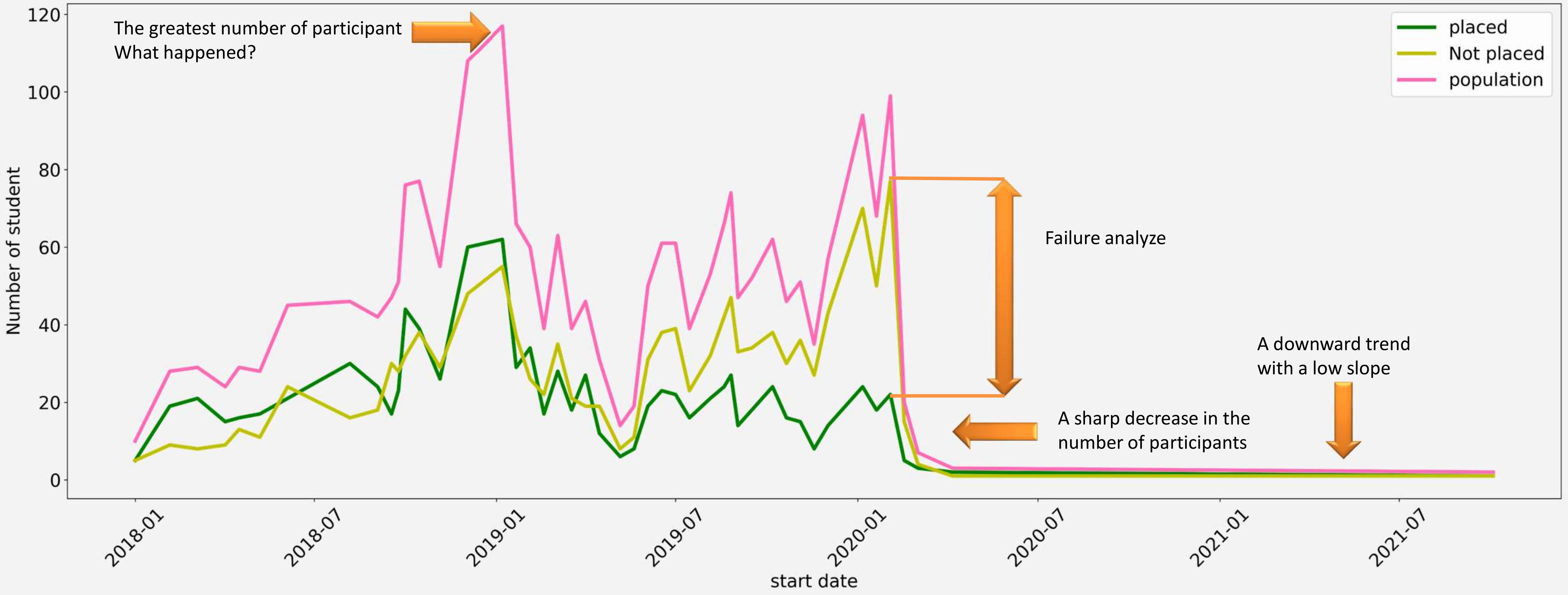

Support Vector Regression


Decision Tree Regression


Linear Regression

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