HR Analytics: Job Change of Data Scientists

Team 8: Aash Gohil, Caroline Guo, Caroline Lun, Chris Chang, Jacinto Lemarroy

TABLE OF CONTENTS

01

Intro to Problem

- Background and Objective
- Dataset Summary

Overview

Background

Company active in data science wants to hire people who received training.

Company wants to analyze factors affecting a candidate's decision on staying or looking for a new job after training.

Objective

To analyze and predict whether a data scientist candidate will look for a new employment or wants to work for the company after training.

Bigger Picture?

To optimize HR costs and increase efficiencies.

Dataset Summary

- enrollee_id : Unique ID for candidate
- city_ development _index : Developement index of the city (scaled)
- gender: Gender of candidate
- relevent_experience: Relevant experience of candidate
- enrolled_university: Type of University course enrolled if any
- education_level: Education level of candidate
- major_discipline :Education major discipline of candidate
- experience: Candidate total experience in years
- company_size: No of employees in current employer's company
- company_type : Type of current employer
- lastnewjob: Difference in years between previous job and current job
- training_hours: training hours completed
- target: 0 – Stay in the "Company" after Training, 1 – Looking for a New job after Training

Overview

12,477 rows and 13 features

Dataset Processing

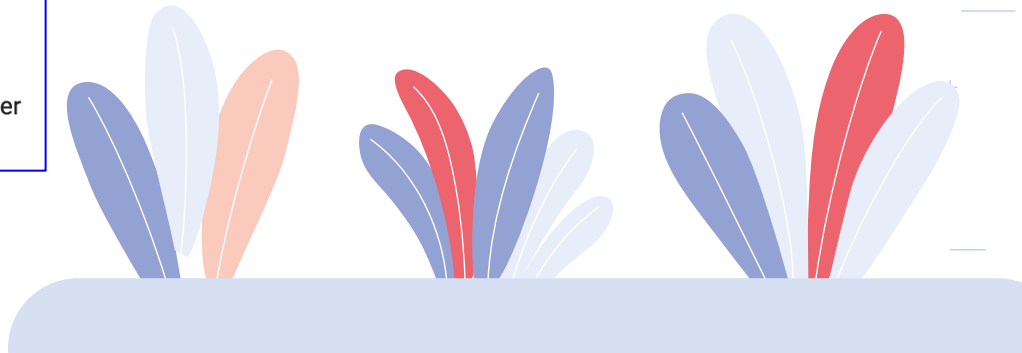
Data Preparation for ML

Created dummy variables and categorized groups within features

- gender->0,1,
- experience->0~21,
- relevent_experience: 0,1 binary
- enrolled_university: get dummy, 3 columns
- education_level: get dummy, 3 columns
- major_discipline: 0,1 STEM, NOT STEM
- company_size: small, large, medium, unknown
- company_type: Unknown, startupGroup, Private, public, NGO, other
- lastnewjob-> turn into integer: 0,1,2,3,4,5;

Data Cleaning for EDA

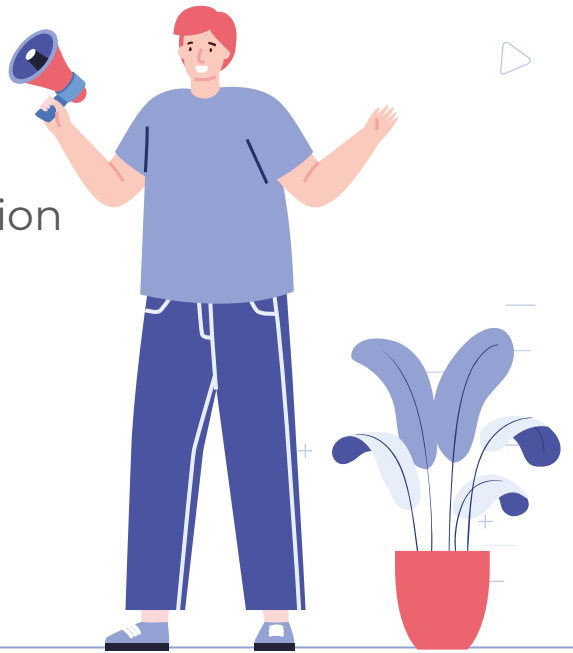
Assigned ‘Unknown’ to null values then dropped nulls that didn’t make much impact in each feature - only kept unknowns in company size and type.



02

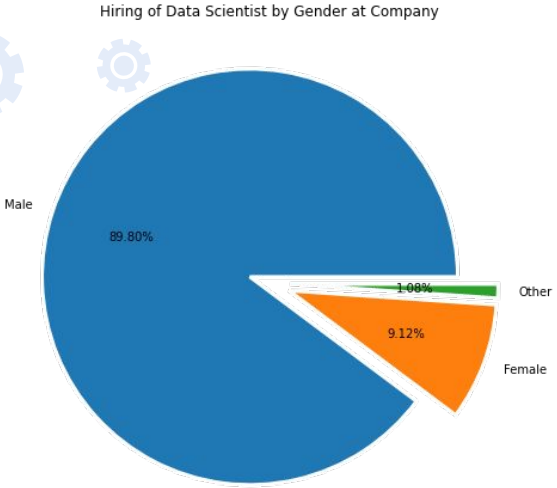
Exploratory Data Analysis

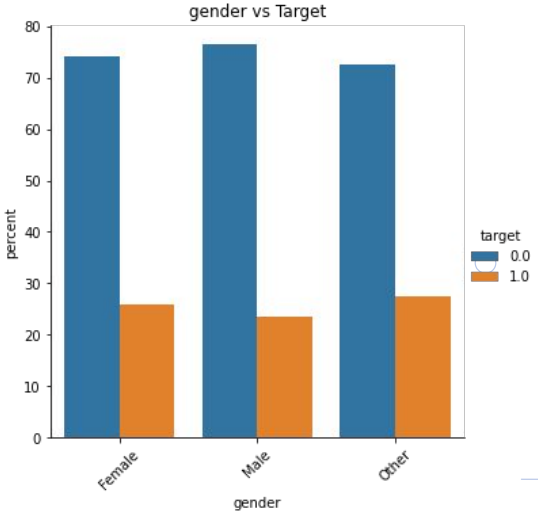
- 2.1. Demographics
  - 2.1.1. Gender
  - 2.1.2. City
  - 2.1.3. Education
  - 2.1.4. Job History
- 2.2. Engagement and Retention



Gender

- Is the hiring of data scientists gender biased?
- What is the impact of gender on attrition?

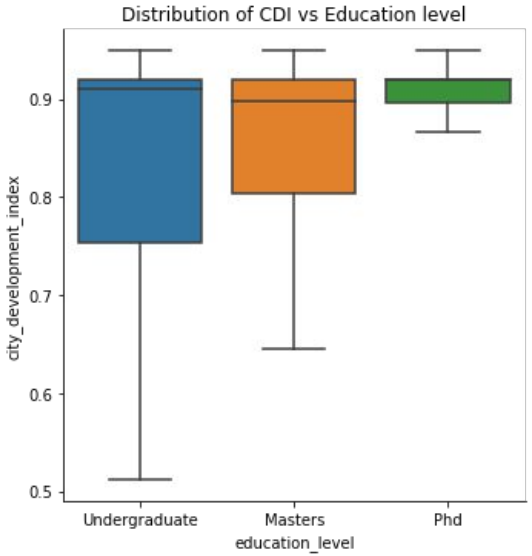


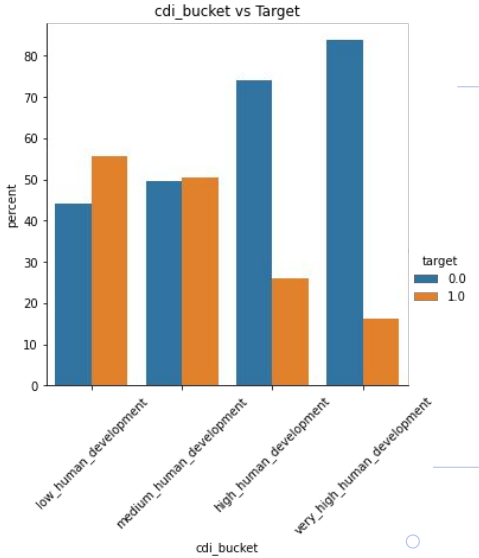


City

- Top 10 cities the company hires from and their corresponding CDI
- How is CDI correlated with an individual's education level?

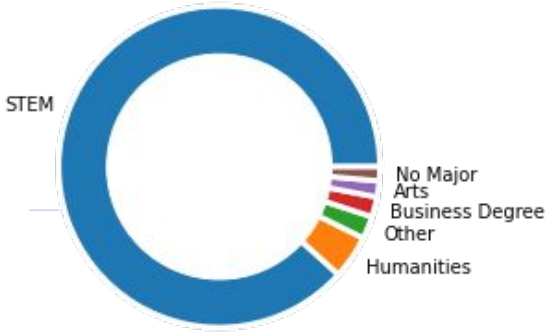
	city	city_development_index	cdi_bucket	count
0	city_103	0.920	very_high_human_development	3262.0
1	city_21	0.624	medium_human_development	1480.0
2	city_16	0.910	very_high_human_development	1093.0
3	city_114	0.926	very_high_human_development	801.0
4	city_160	0.920	very_high_human_development	619.0
5	city_136	0.897	very_high_human_development	405.0
6	city_67	0.855	very_high_human_development	277.0
7	city_75	0.939	very_high_human_development	218.0
8	city_104	0.924	very_high_human_development	190.0
9	city_102	0.804	very_high_human_development	190.0

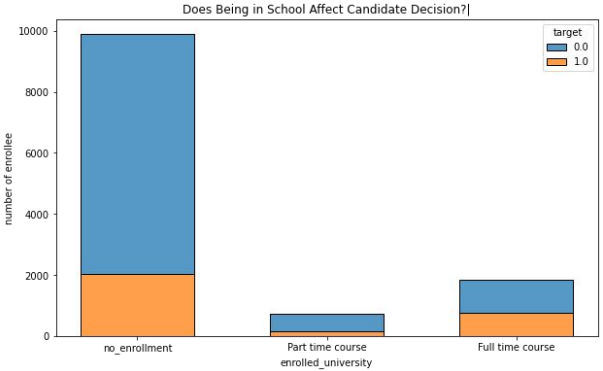


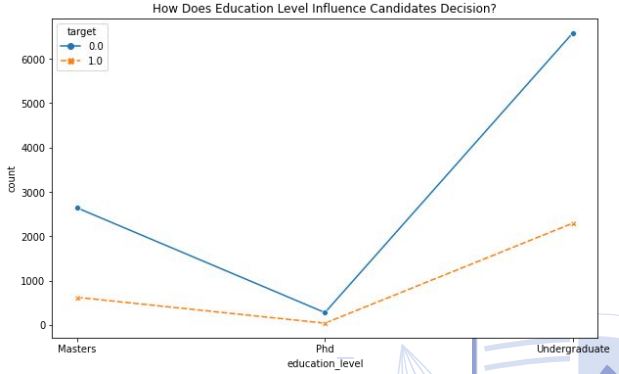


Education

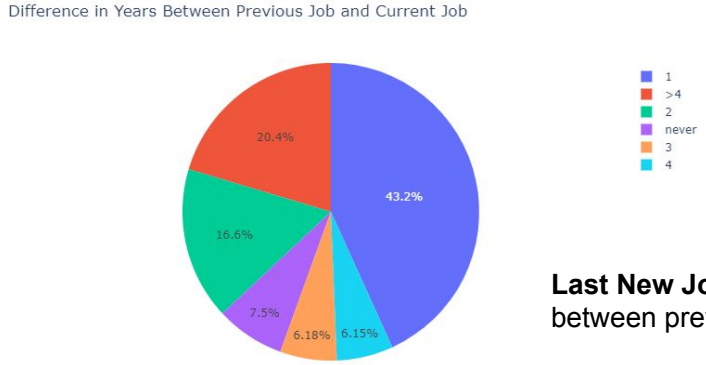
- What are some education characteristics for those candidates who are staying?

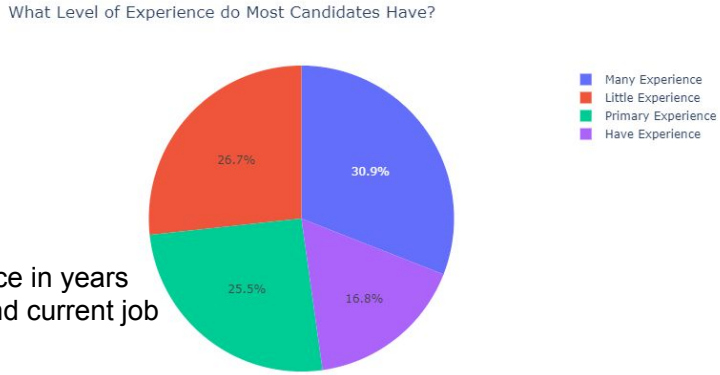




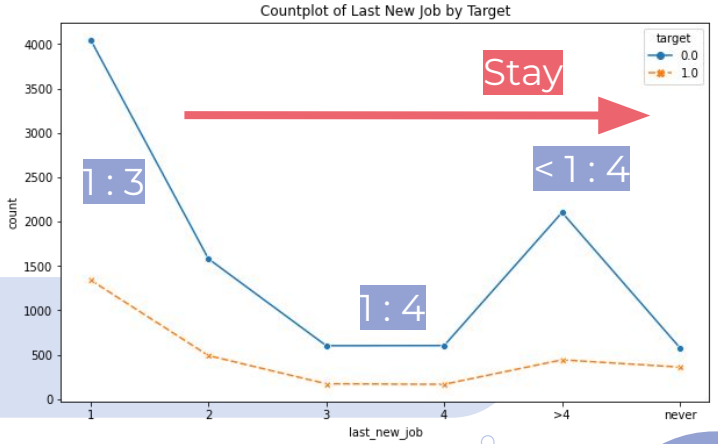


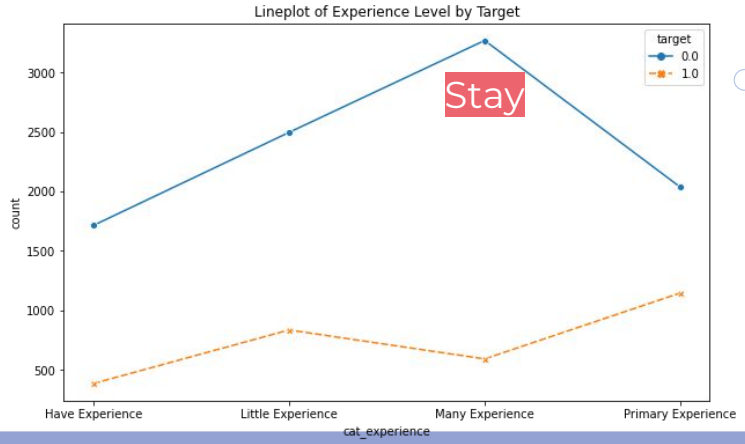
Gap Years/Relevant Experience





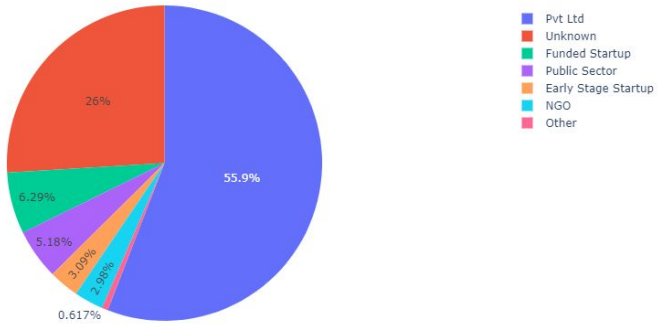
Last New Job: Difference in years between previous job and current job





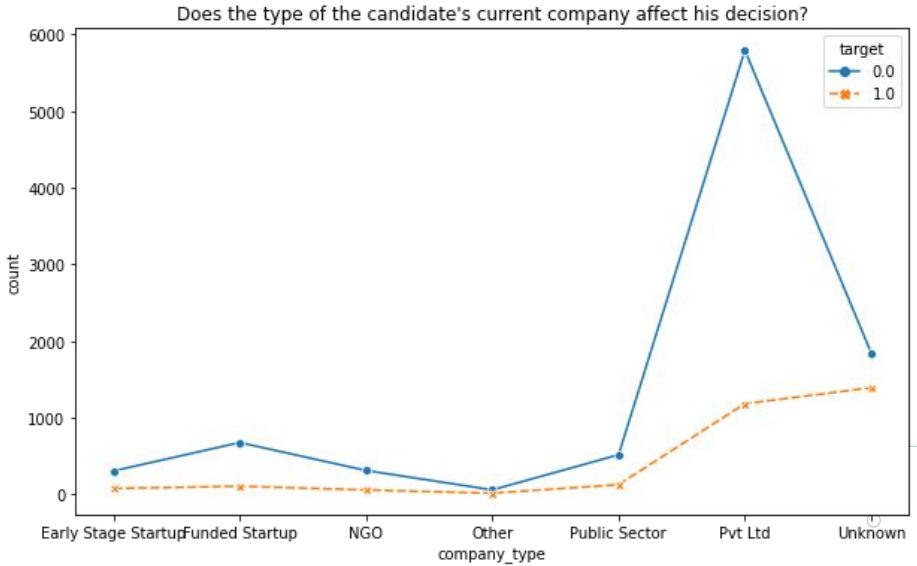
Company Type

What is the most prevalent company type in the dataset?



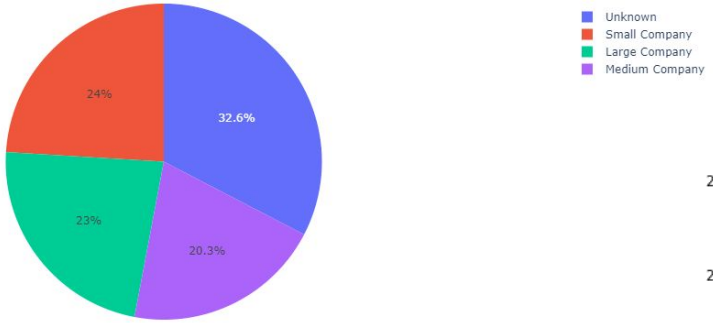
56% of Candidates from **Private Company**

Most candidates from **Private Company** seems like to stay after training



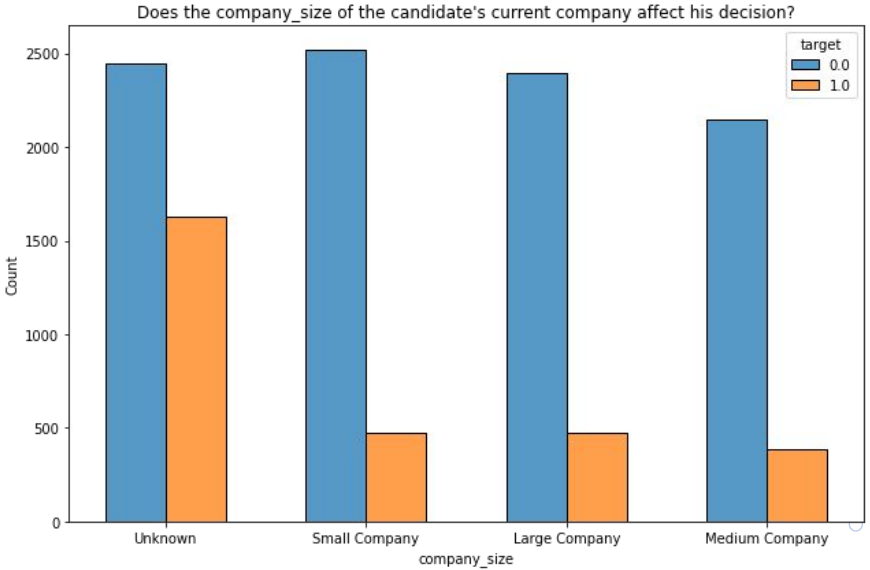
Company Size

What is the most prevalent company size in the dataset?



Candidates come from **different size** companies.

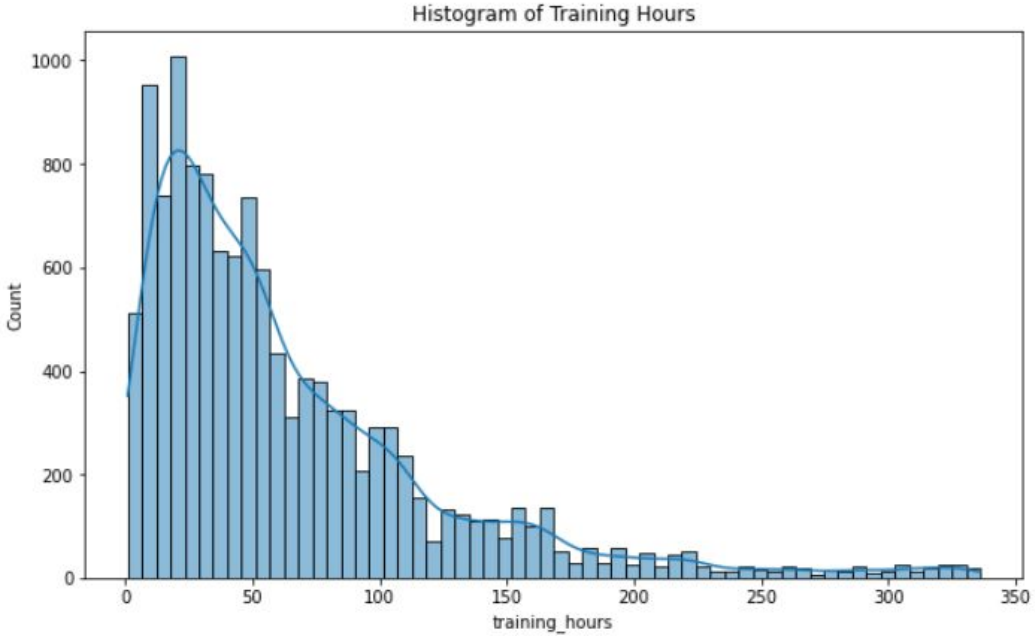
Candidates coming from **small companies** tend to **stay** with the "company" after training



Training and Retention

Training Hours:

- How much training hours does the company invest in its future employees?

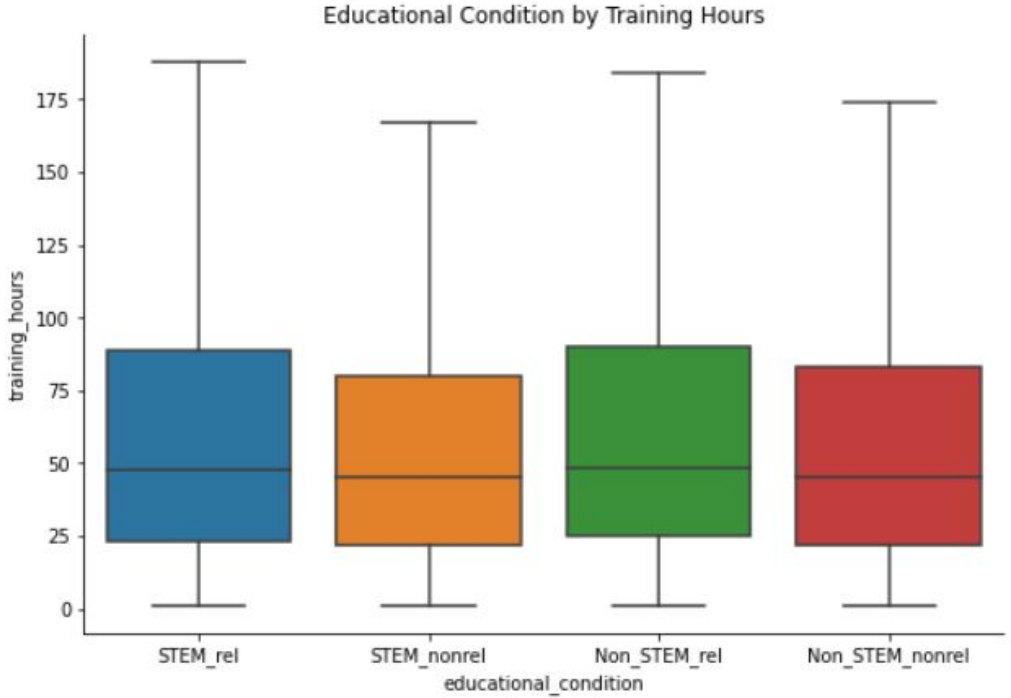


mean	64.927306
std	59.732622
min	1.000000
25%	23.000000
50%	47.000000
75%	88.000000
max	336.000000

Training and Retention

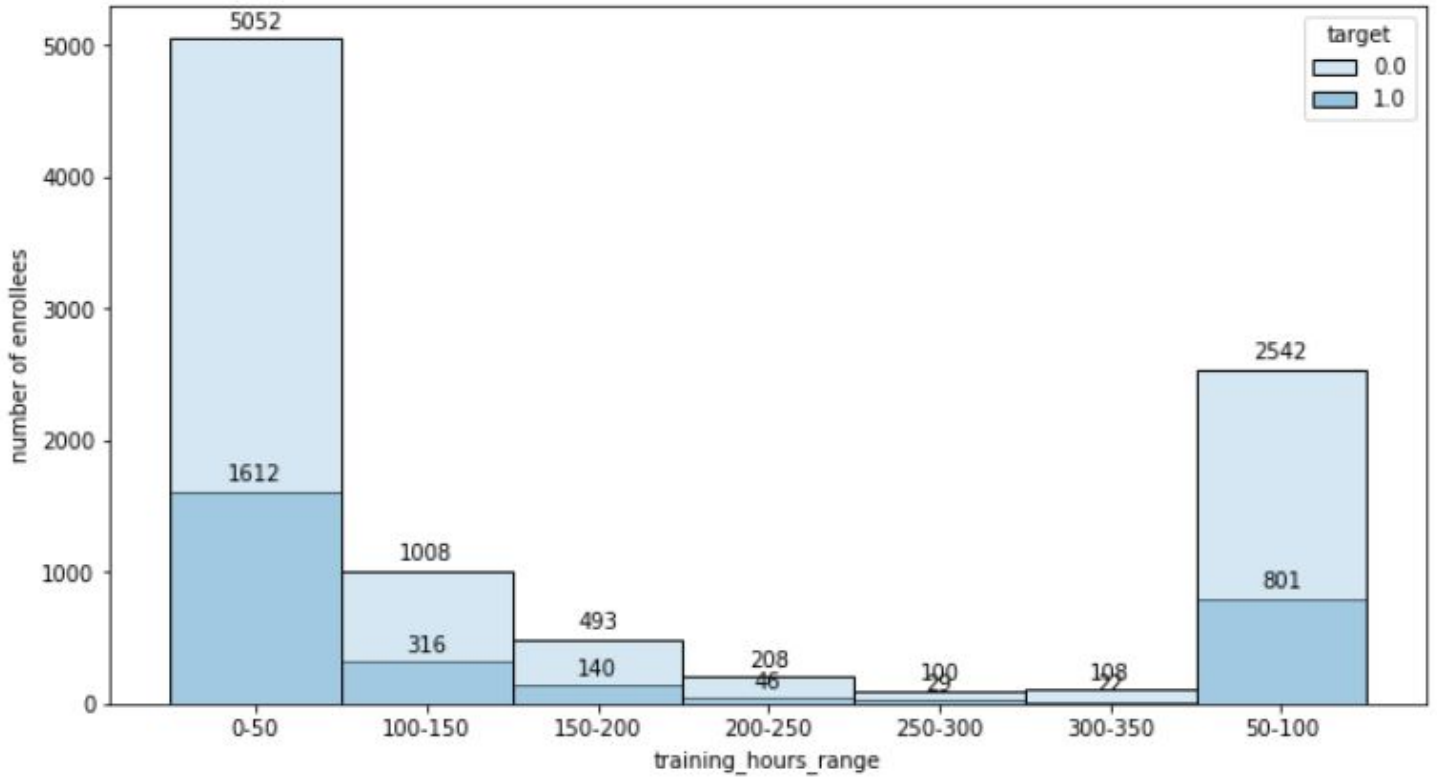
Training Hours:

- Is training dependant on candidates past background?



Training and Retention

- Does training hours impact retention?

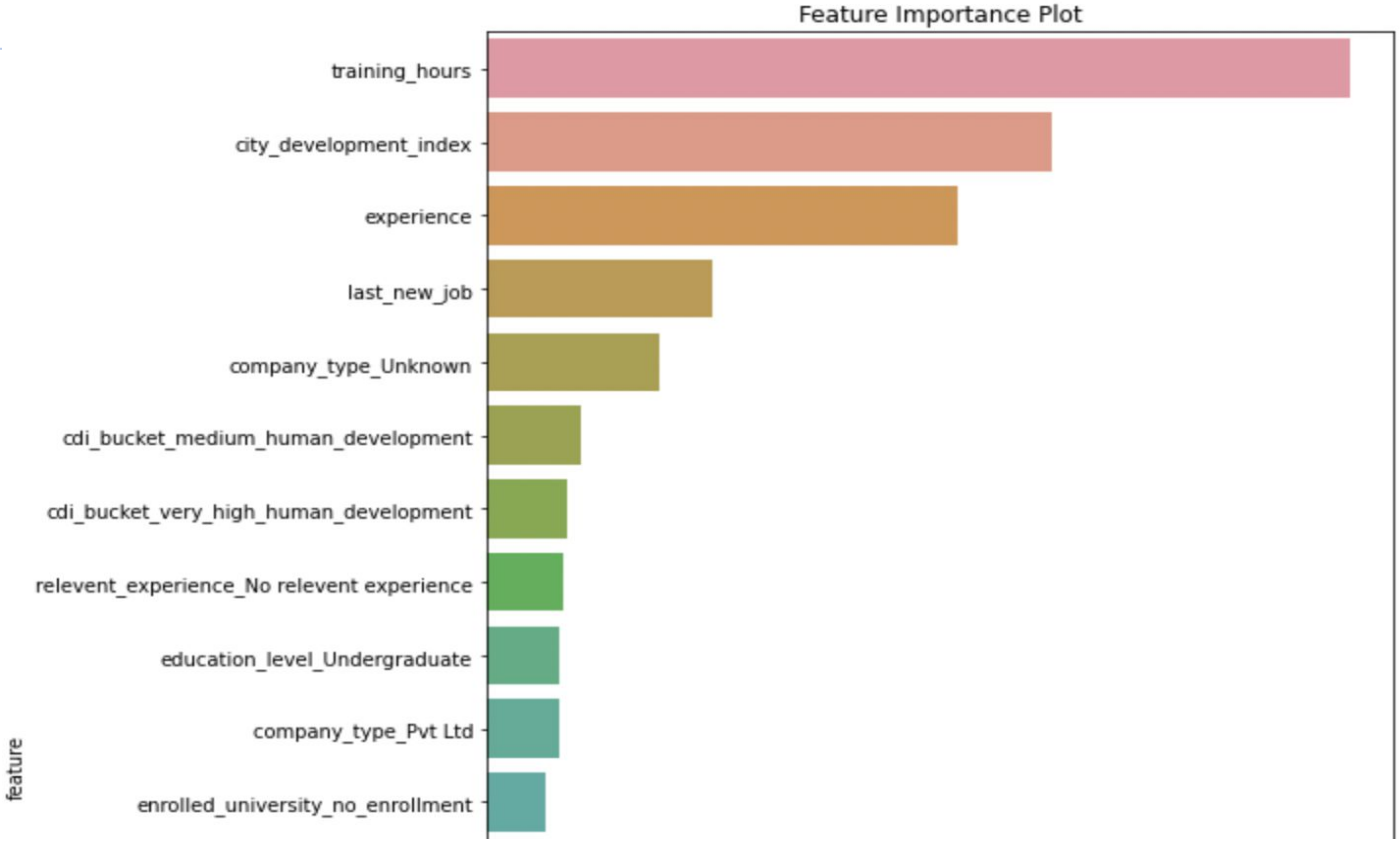


03

Machine Learning Models & Results

- Gaussian Naive Bayes
- Logistic Regression
- Random Forest
- XGBoost
- Multilayer Perceptron Classifier

Main Results of ML Models: Feature Engineering



Main Results of ML Models

	Accuracy	Sensitivity	Specificity	AUC
Gaussian Naive Bayes	0.7476	0.8201	0.5759	0.75
Logistic Regression	0.7769	0.1956	0.9527	**0.78**
Random Forest	0.7891	0.4154	0.8880	0.7681
XGBoost	0.8037	0.4833	0.9005	0.69
Multilayer Perceptron Classifier	0.88	0.8333	0.9231	**0.78**

04

Conclusion

- Main Results Overall
- Recommendation for HR hiring

Main Results & Advices For HR

For the Exploratory Data Analysis:

A. Demographics

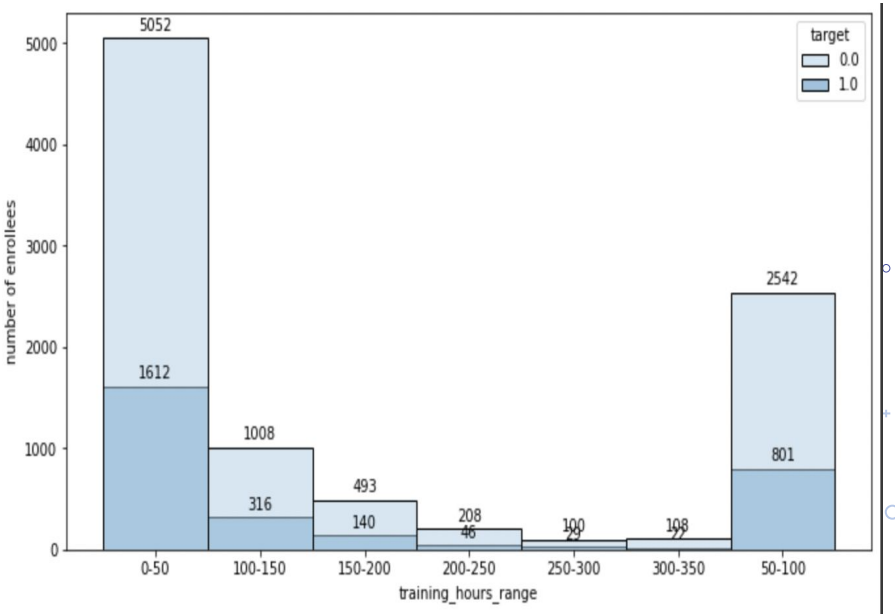
- No **gender** bias.
- Candidates with higher “quality” come from cities with higher **CDI.**
- STEM **majored** candidates have higher chance to stay.
- With **gap** between last and new job increases, possibility of stay goes up as well.

Main Results & Advices For HR

For the Exploratory Data Analysis:

B. Engagement and Retention

- **Ideal range of training hours**: 50-150 hours training range has higher percentage of retention about 76%.
- STEM major or not, has relevant experience or not do not impact training hours.

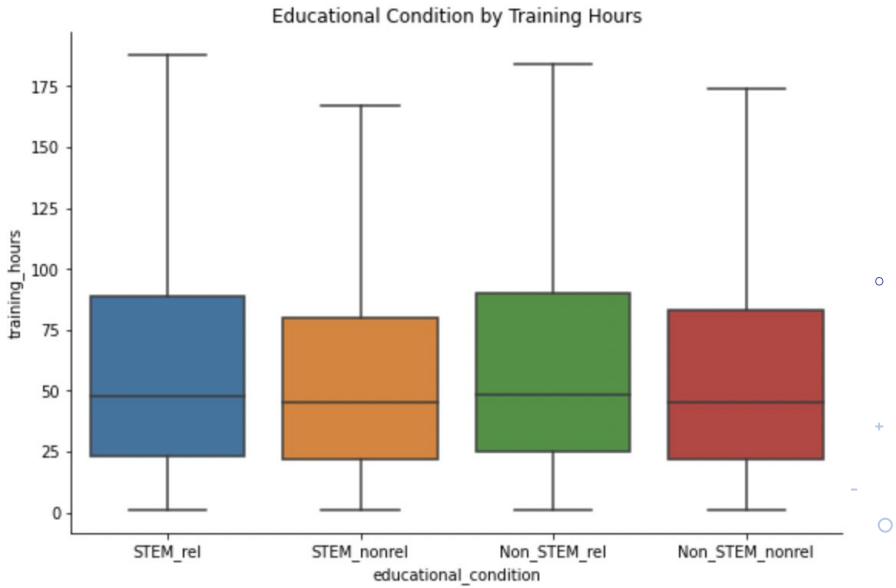


Main Results & Advices For HR

For the Exploratory Data Analysis:

B. Engagement and Retention

- **Ideal range of training hours**: 50-150 hours training range has higher percentage of retention about 76%.
- STEM major or not, has relevant experience or not do not impact training hours.



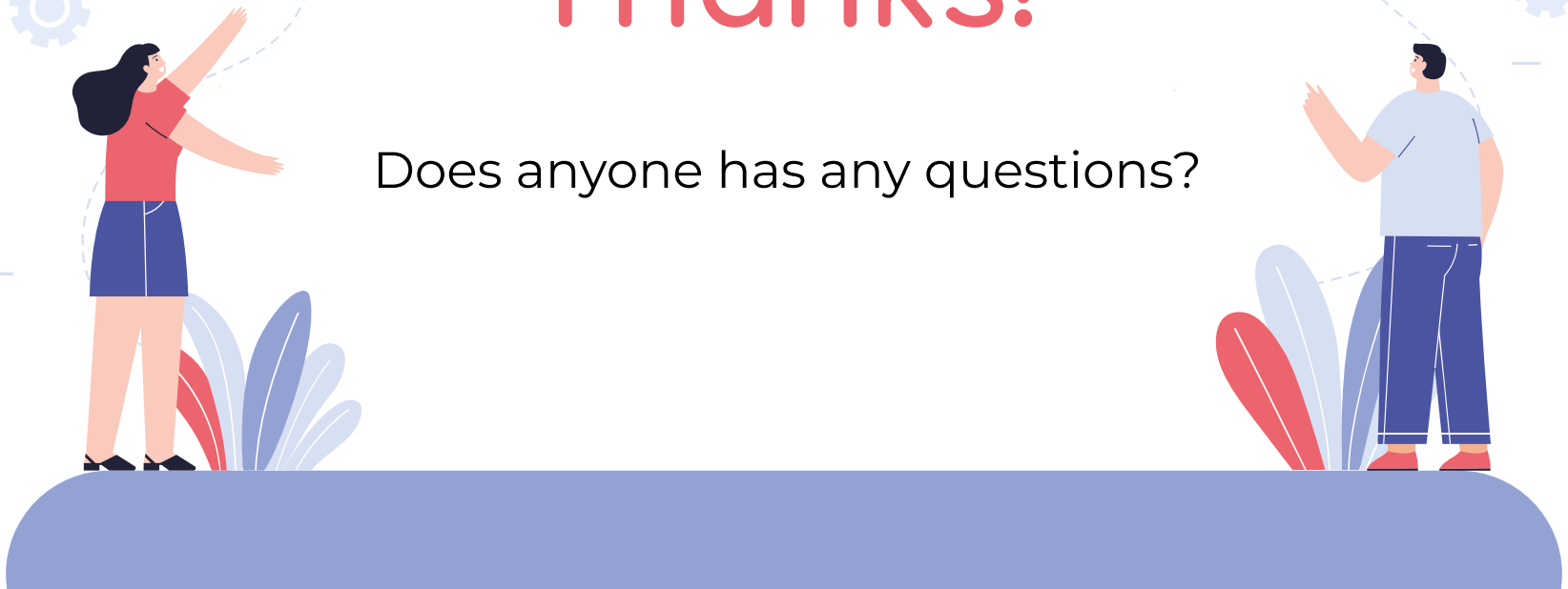
Main Results & Advices For HR

For the Machine Learning Model:

- Logistic and MLP has the best performance, with AUC score of 0.78.
- Training hours, city development and candidates' year of experience play a vital role in predicting whether the candidates stay or not after training.

Thanks!

Does anyone has any questions?



Multilayer Perceptron Classifier: An artificial neural network with a forward structure.