HR Data Analytics

Presentation

By Alice Chang 9 Oct 2021, IOD Data Science & AI Course –Capstone Project

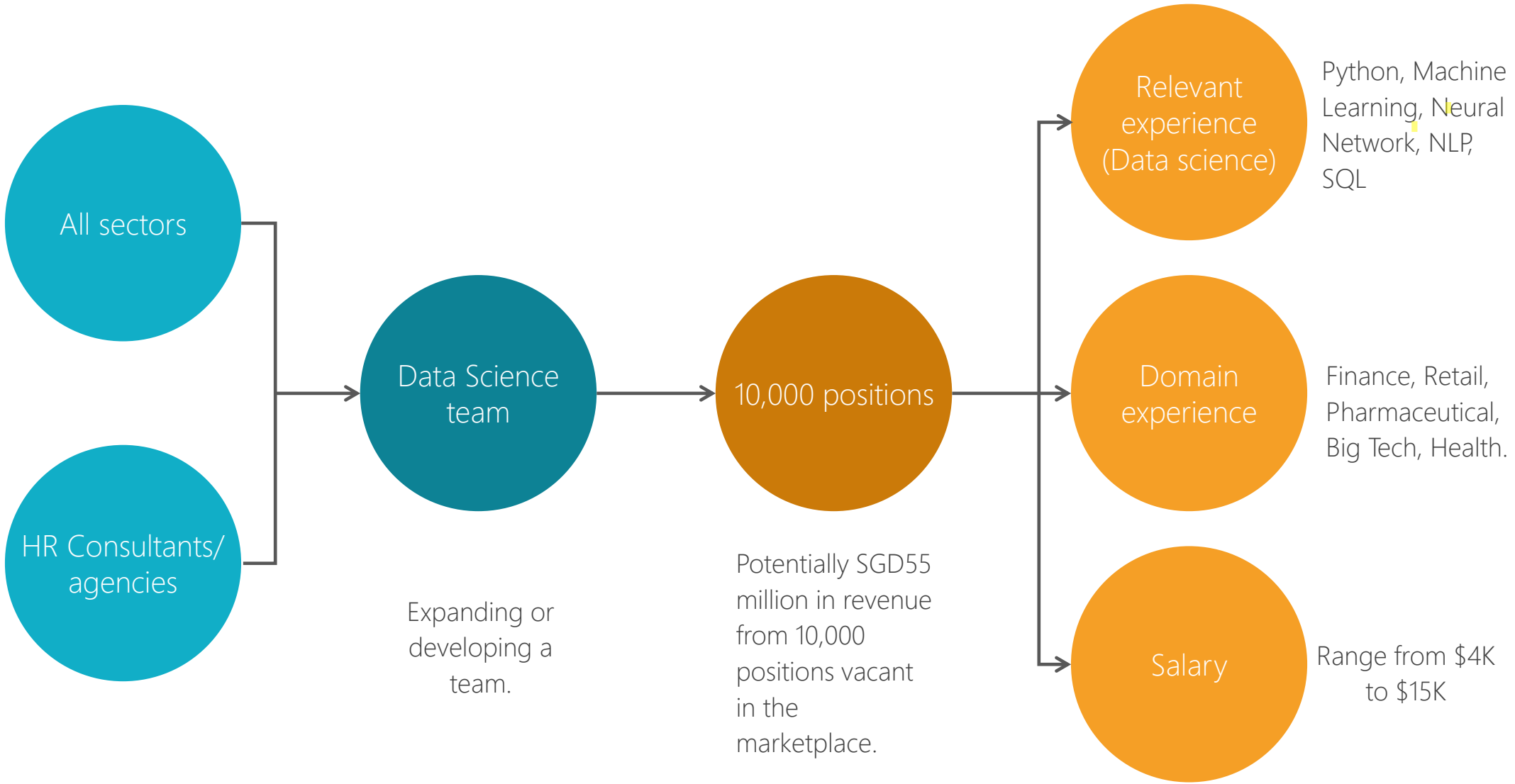
Project Analysis

1. MARKET ASSESSMENT
2. STAKEHOLDERS
3. DATA ANALYSIS
4. DATA SCIENCE
5. TESTING
6. IMPLEMENT PROJECT

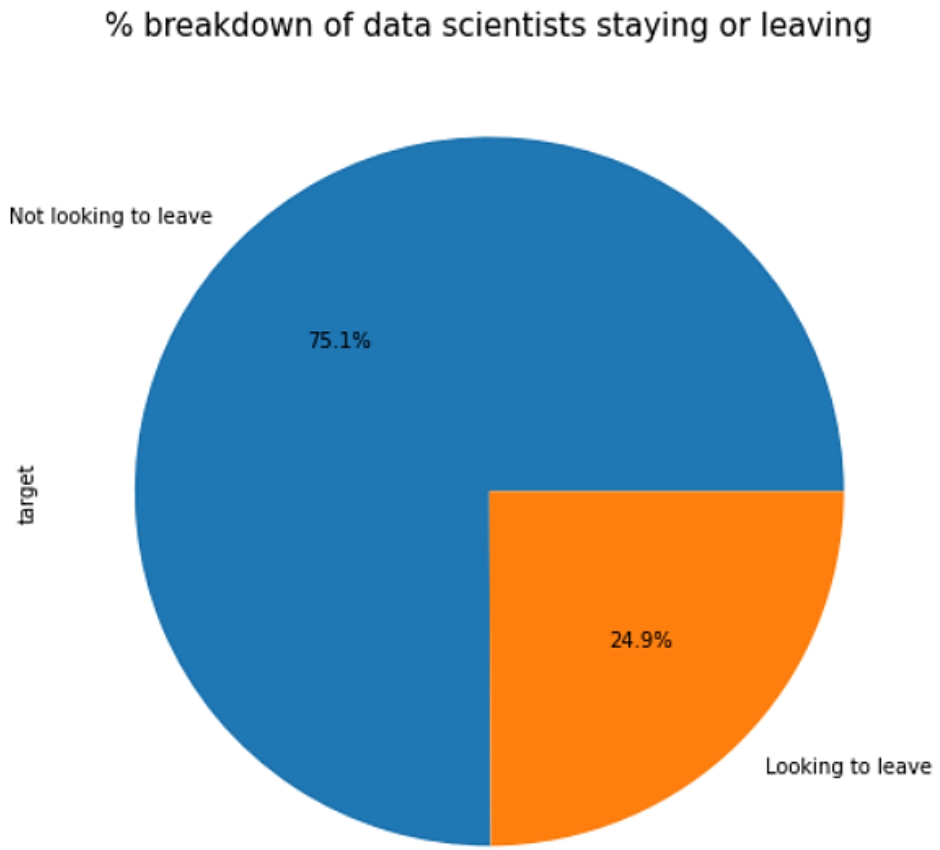
Market Analysis

	POSITIVE	NEGATIVE
INTERNAL	**STRENGTH** › Robust dataset where the candidates already have the right technical foundation and background to fulfil their role as data scientists/ analysts. › Educational background and training of candidates in the dataset are promising and highly qualified.	**WEAKNESS** › Dataset is highly imbalanced. › There are far too many males than females in the dataset. › Only 25% wants to quit vs 75% who have no intention to quit (yet).
EXTERNAL	**OPPORTUNITY** › As at 8 Oct 2021, there are 1729 'Data Science' and 1100 'Data Analyst' positions in Singapore advertised in LinkedIn. From LinkedIn alone, over 2500 positions to fill in the job market. › From JobStreet, there are over 7800 posts for Data Science related jobs in Singapore.	**THREAT** › Data Scientists who wish to quit may not have the domain knowledge of certain industries. For example, Bio-science research, finance and medical sectors require direct experience or domain knowledge. › COVID-19 means that candidates living abroad will find it challenging to relocate to Singapore.

Stakeholder Analysis



Data Analysis

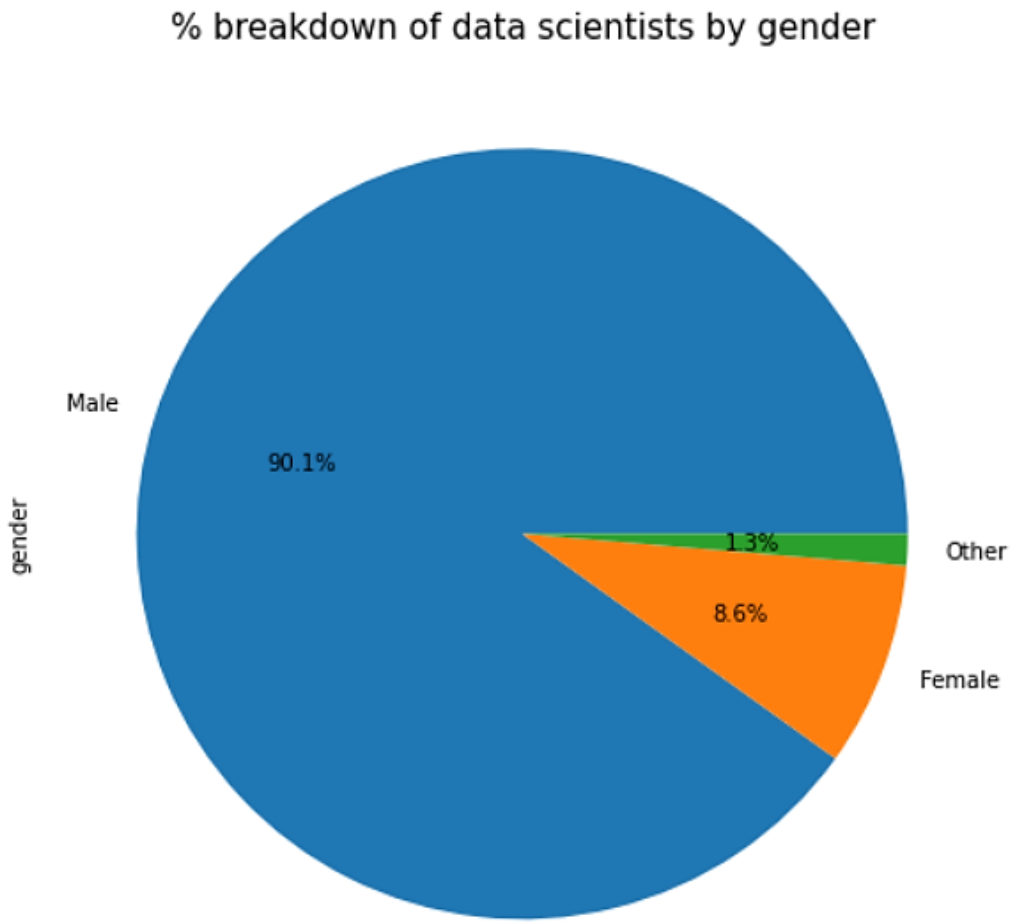


25% wants to quit their present job.

75% are not quitting their present job yet.

Total number: 19,158.

Data Analysis

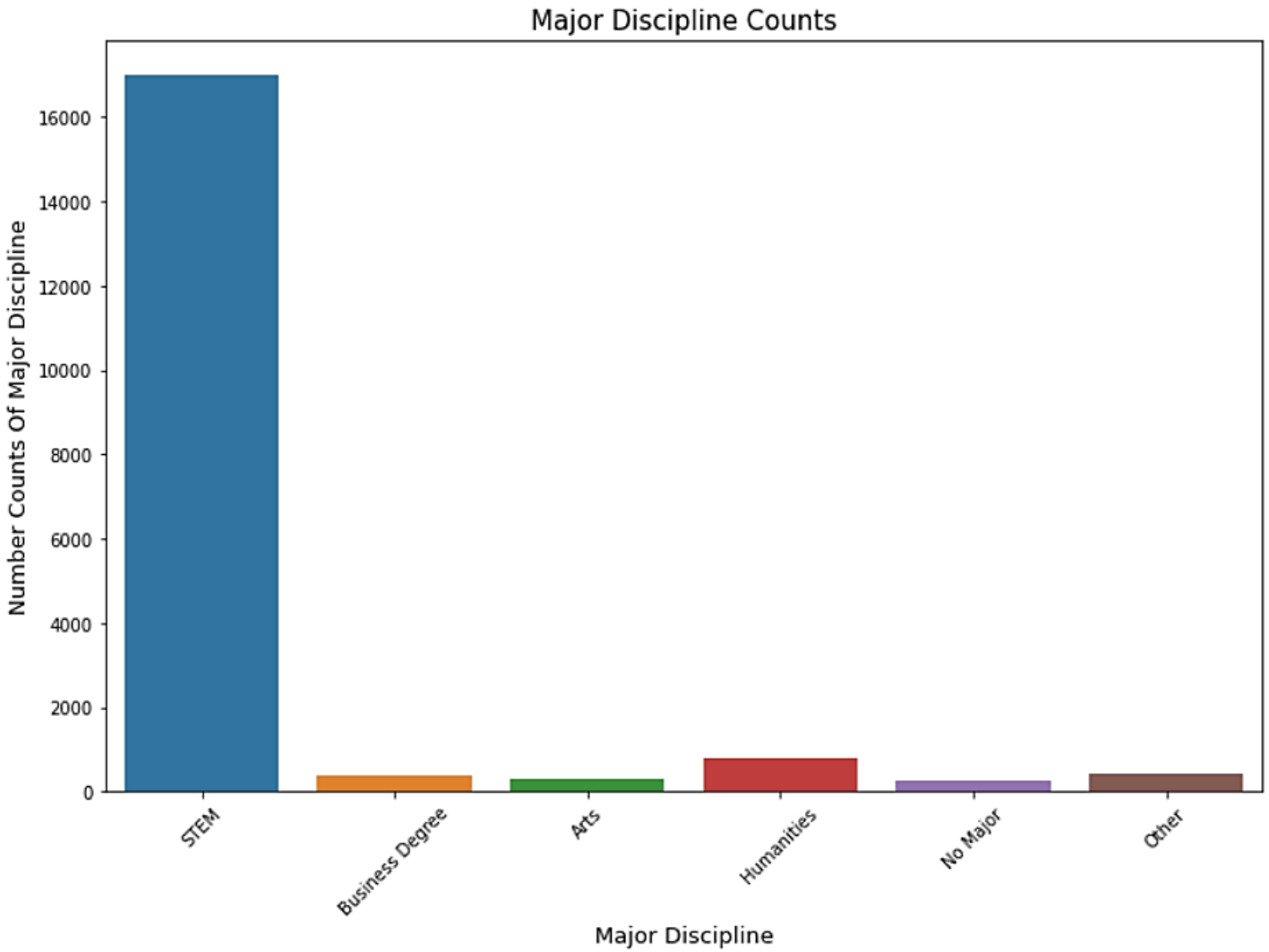


Less than 10% were female.

Data science is still dominated by male (90%).

Total number: 19,158.

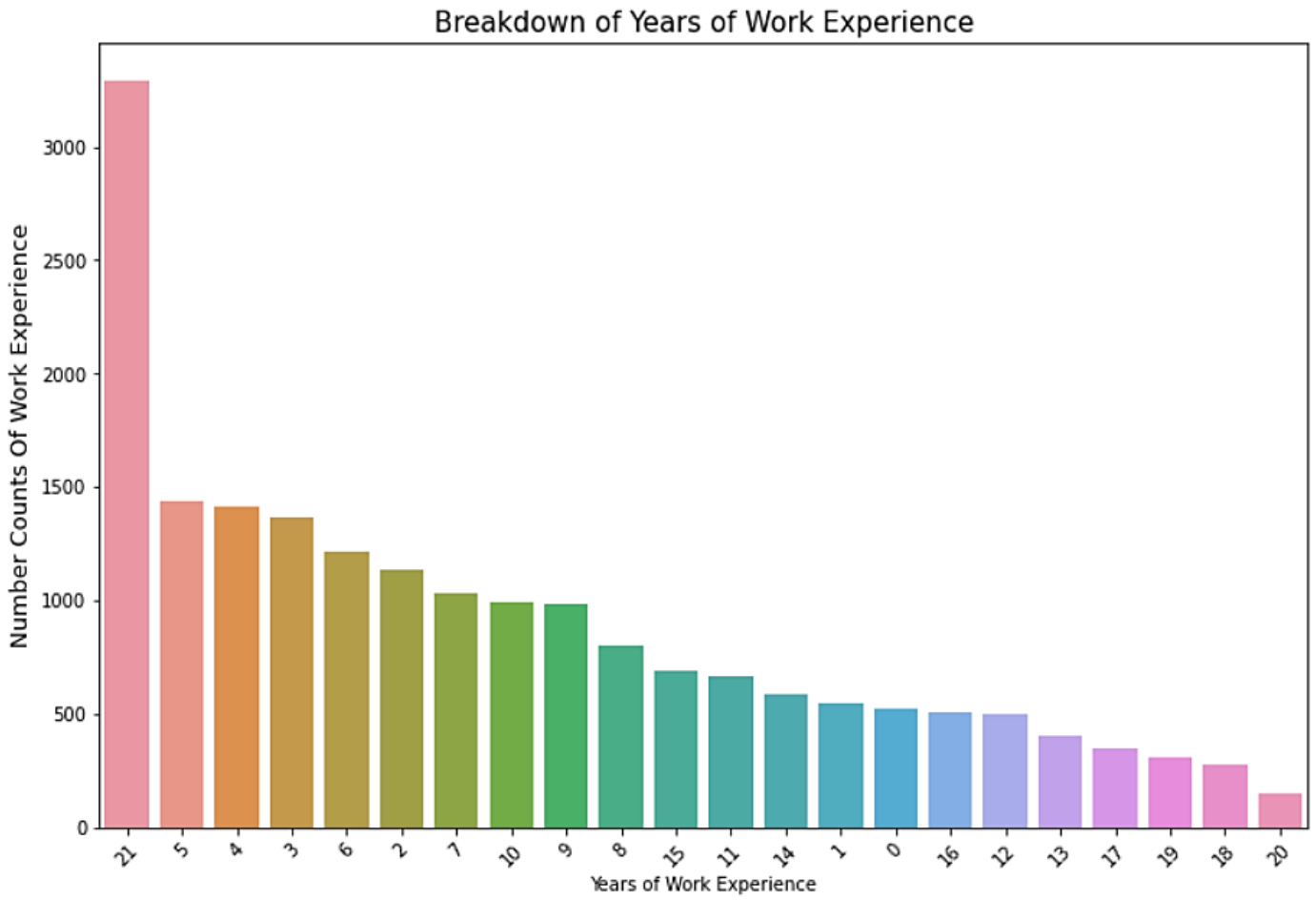
Data Analysis



Majority of data scientists had a major in STEM (Science, Technology, Engineering and Maths).

Total number: 19,158.

Data Analysis

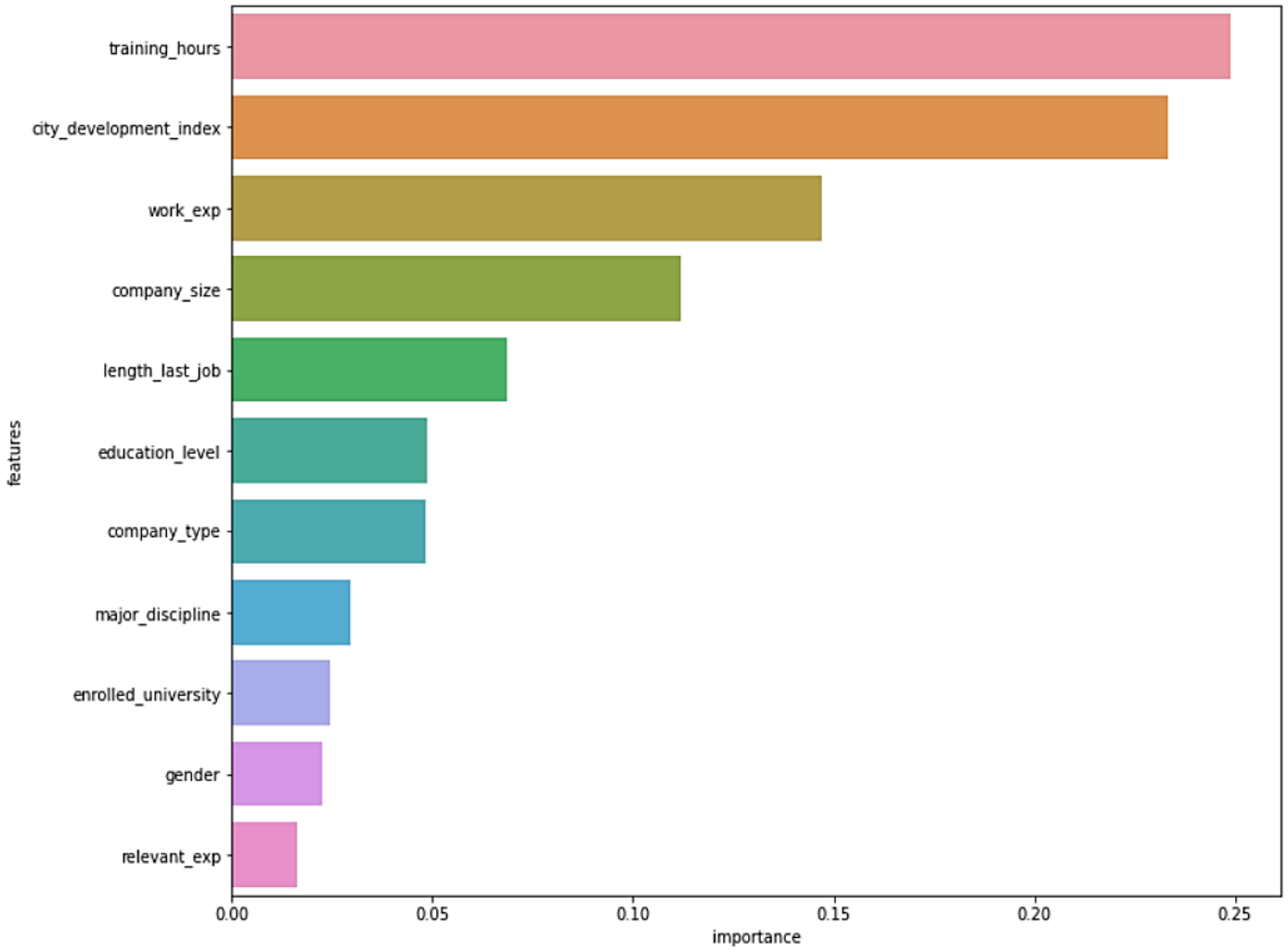


Those who had over 20 years of overall work experience made up the largest group. This tells us quite a number of data scientists started off with another discipline but adapted into data science as a mid career change.

The next group (2 – 10 years) formed a substantial number who have chosen data science as a choice career quite early in their work life.

Total number: 19,158.

Feature Importance



The 2 features that had most significance on the results:

1. Training hours (the more training in data science, the more competent and confident the candidate is in their job)
2. City Development Index (high-tier or mid-tier city determines the lifestyle they want and availability of career opportunities.)

The next 2 features of importance:

3. Work experience (not necessarily relevant experience).
4. Company size (prospects for growth/ promotion)

Machine Learning

	model	accuracy	specificity	sensitivity
0	GradientBoostClassifier	0.78	0.90	0.41
1	AdaBoostClassifier	0.78	0.92	0.35
0	StackingClassifier	0.78	0.90	0.40
4	RandomForestClassifier	0.77	0.91	0.36
6	MLPClassifier	0.77	0.92	0.34
3	KNeighborsClassifier	0.74	0.87	0.35
5	XGBClassifier	0.74	0.87	0.36
2	DecisionTreeClassifier	0.71	0.82	0.37

ACCURACY SCORE

78%

GradientBoostClassifier
AdaBoostClassifier
StackingClassifier

SPECIFICITY (For non-quitters)

92%

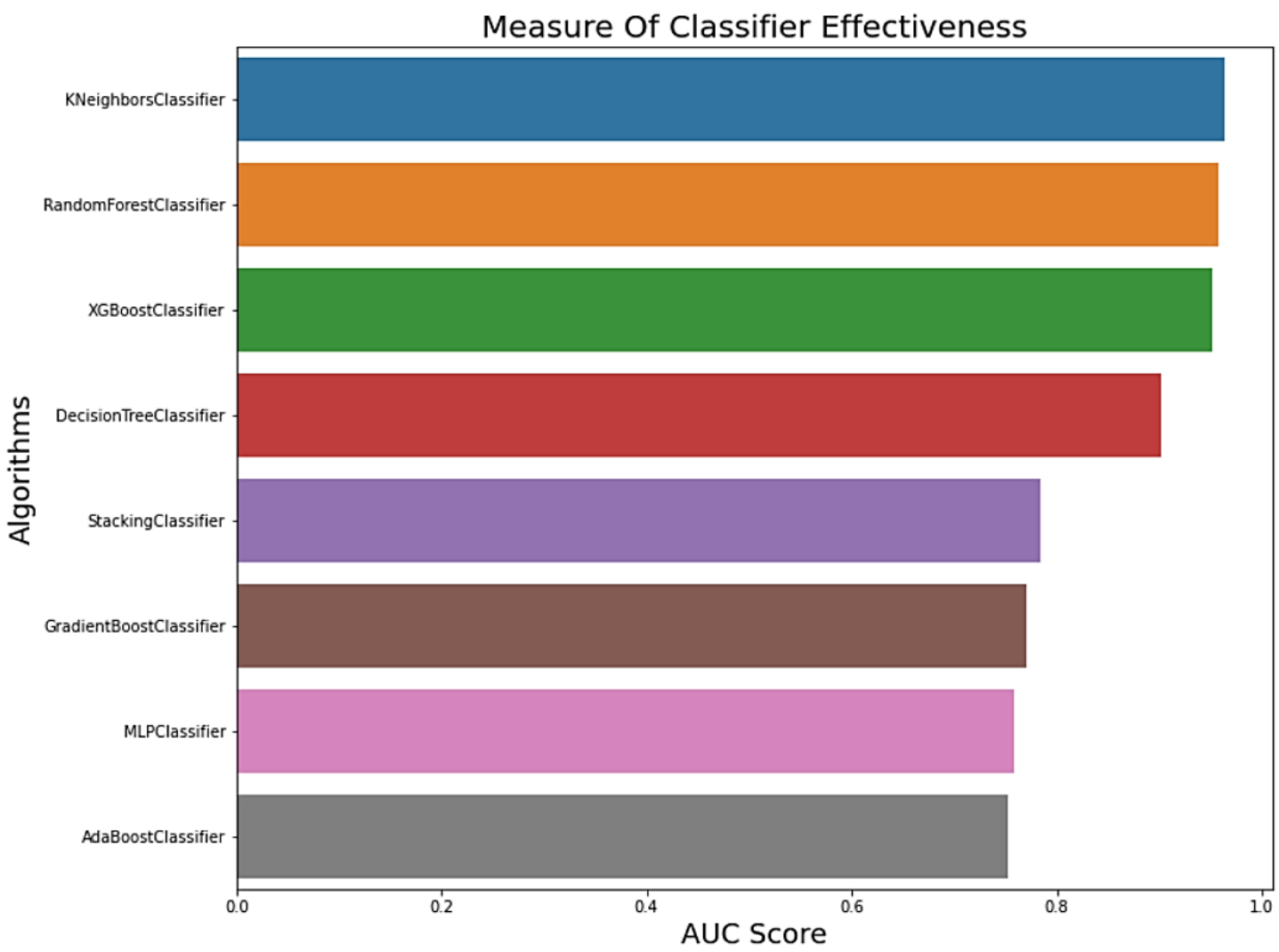
AdaBoostClassifier
MLPClassifier

SENSITIVITY (For quitters)

41%

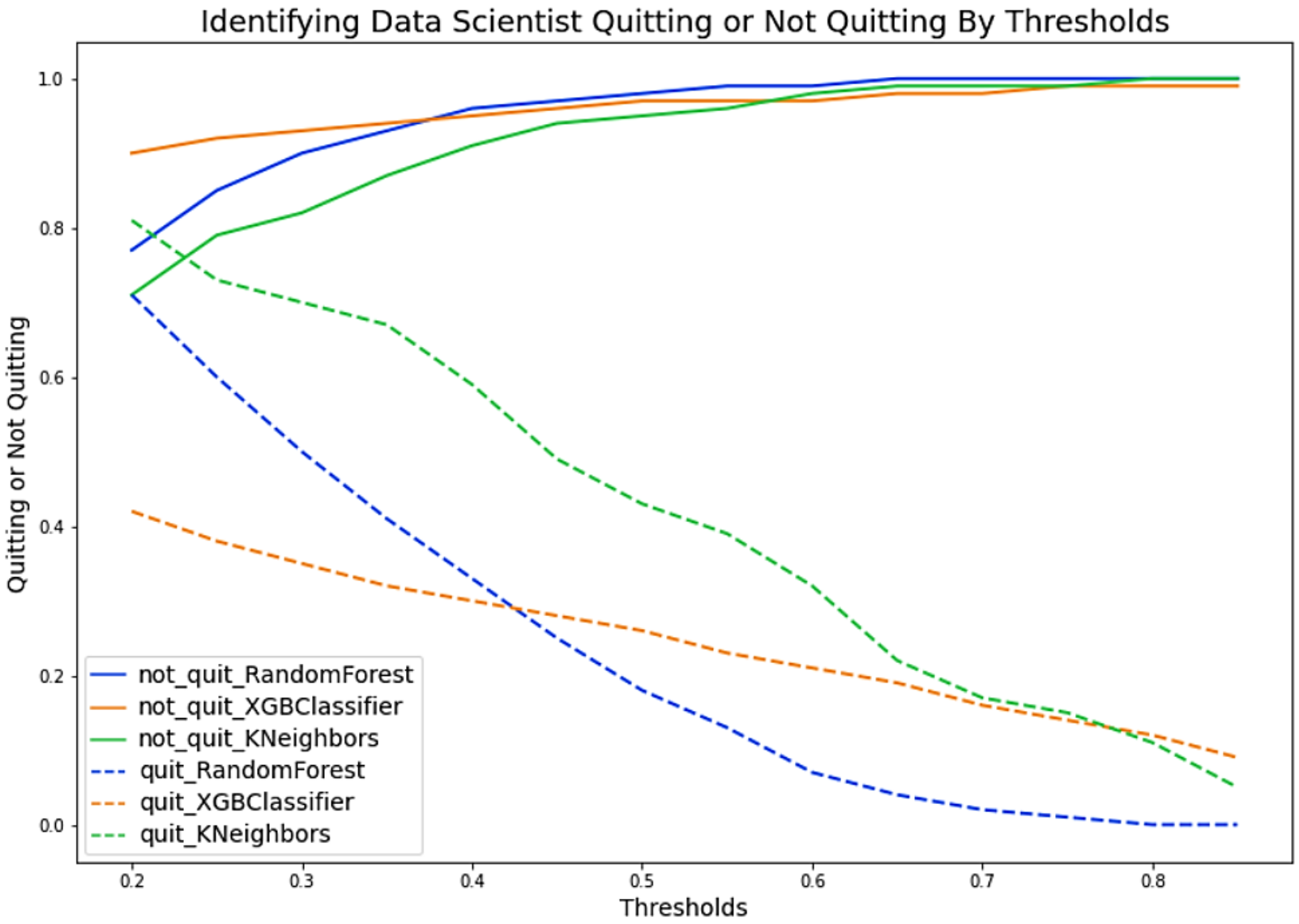
GradientBoostClassifier.

Testing (AUC ROC Score)



	Models	ROC AUC SCORE
4	KNeighborsClassifier	0.963129
6	RandomForestClassifier	0.957559
7	XGBoostClassifier	0.951514
3	DecisionTreeClassifier	0.902327
0	StackingClassifier	0.783812
1	GradientBoostClassifier	0.770357
5	MLPClassifier	0.757838
2	AdaBoostClassifier	0.752410

Testing (Threshold)



Implementation: deciding on thresholds

	20%	25%	30%	35%	40%	45%	50%	Yes/No
KNeighborsClassifier	1,837	33	33	33	13	0	0	Yes
RandomForestClassifier	3,750	1,086	233	38	4	2	0	Yes
XGBoostClassifier	63	46	25	15	9	6	4	No
DecisionTreeClassifier	296	296	296	292	292	270	270	Yes
StackingClassifier	277	8	0	0	0	0	0	No
GradientBoostClassifier	2,887	1,409	0	0	0	0	0	Yes
MLPClassifier	34	20	15	12	8	5	3	No
AdaBoostClassifier	19,158	19,158	19,158	19,158	19,158	19,158	0	No

Thank You