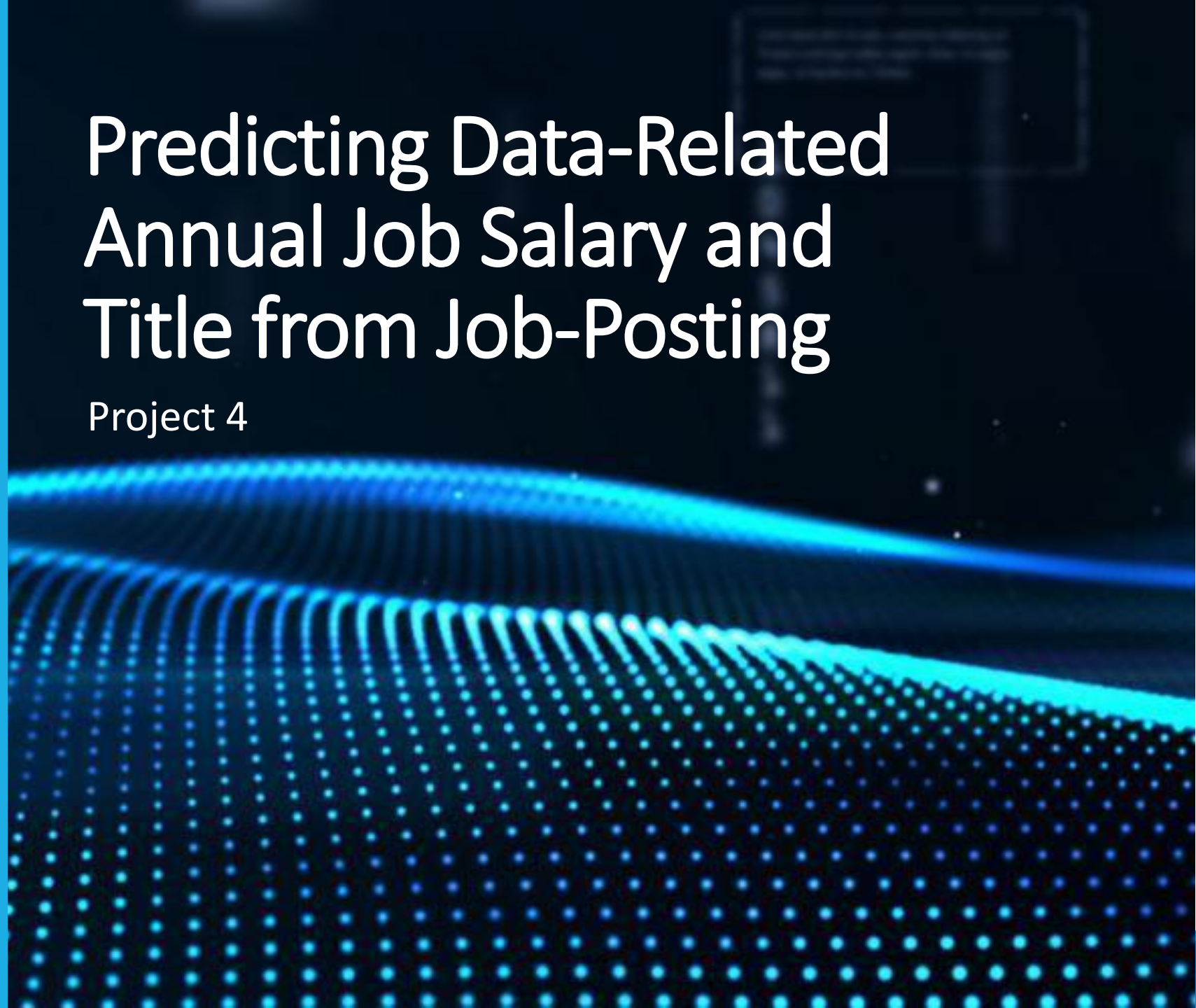


Predicting Data-Related Annual Job Salary and Title from Job-Posting

Project 4





Project Objectives

- **Identify factors that predict data-related role annual salaries**
- **Identify what distinguishes between data-related job categories/titles**

General approach to project

Web scrape
Job Listings



Natural
Language
Processing



Modelling
(Regression
and
Classification)



Findings and
Conclusion



Job listing sourced from

- **Key Libraries Used for Web Scraping and Extraction**

- Requests
- BeautifulSoup

- **Job Search Criteria:**

- Australia is the designated search location: selected as locally appropriate to client needs
- Search terms selected:
 - 'research+scientist',
 - data+scientist',
 - 'machine+learning+engineer',
 - 'data+analyst'



Search Results

- 'research+scientist'

- 'Total jobs': 274
- 'Pages to Search': 5

- 'data+scientist'

- 'Total jobs': 638
- 'Pages to Search': 11

- 'machine+learning+engineer':

- 'Total jobs': 382
- 'Pages to Search': 7

- 'data+analyst'

- 'Total jobs': 3135
- 'Pages to Search': 20

TOTAL JOBS = 2357
WITH SALARY = 493

Job Title <'h3'>

Company

'div' 'class': 'icl-u-lg-mr--sm icl-u-xs-mr--xs'

Location

'span' 'class': 'jobsearch-JobMetadataHeader-iconLabel'

Salary

'span' 'class': 'jobsearch-JobMetadataHeader-iconLabel'

Description

'div',{id': 'jobDescriptionText'

**Locating HTML tags
containing relevant
information all in a
single Job
Information Frame**

IBM Research/Data Scientist - Health AI, Image Analytics

IBM ★★★★★ 29,011 reviews

[Apply On Company Site](#)

📍 Melbourne VIC 3001

💰 \$90,300 a year

Introduction

We are seeking a highly motivated and passionate researcher to join our AI for Health team. Our mission is to develop Artificial Intelligence (AI) technologies to help deliver top quality health care while mitigating the spiralling costs of the health care system. In order to make this a reality, we develop AI and advanced machine learning methodologies in collaboration with clinical researchers. We are seeking a proactive, enthusiastic researcher to contribute to our AI applications for health assessment and disease management. The candidate will dynamically work across disciplines and geographies to contribute towards developing the transformational ideas that will have impact on the world.

Your Role and Responsibilities

IBM Research/Data Scientist - Health AI, Image Analytics

Offered on a permanent regular employment basis or on a fixed term hire 24 Months basis

Compensation package from \$90,300 (full-time equivalent inclusive of superannuation) and will be determined based on successful candidates' relevant skills and experience.

We are seeking a highly motivated and passionate researcher to join our AI for Health team. Our mission is to develop Artificial Intelligence (AI) technologies to help deliver top quality health care while mitigating the spiralling costs of the health care system. In order to make this a reality, we develop AI and advanced machine learning methodologies in collaboration with clinical researchers. We are seeking a proactive, enthusiastic researcher to contribute to our AI applications for health assessment and disease management. The candidate will dynamically work across disciplines and geographies to contribute towards developing the transformational ideas that will have impact on the world.

Company Info



[Follow](#)

Get job updates from IBM

IBM

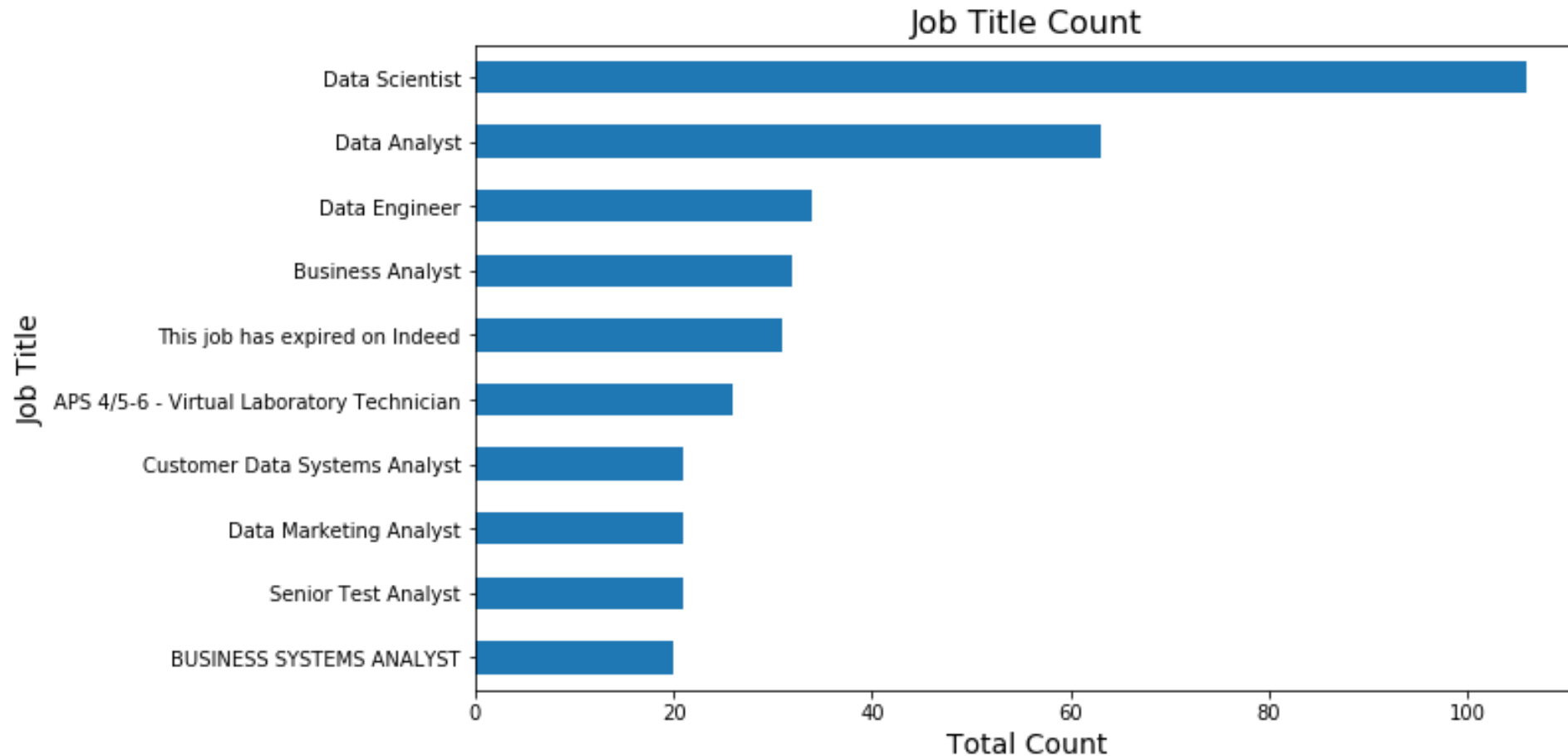
★★★★★ 29,011 reviews

At IBM, work is more than a job—it's a calling. To build. To design. To code. To consult. To think along with clients and sell. To make m...

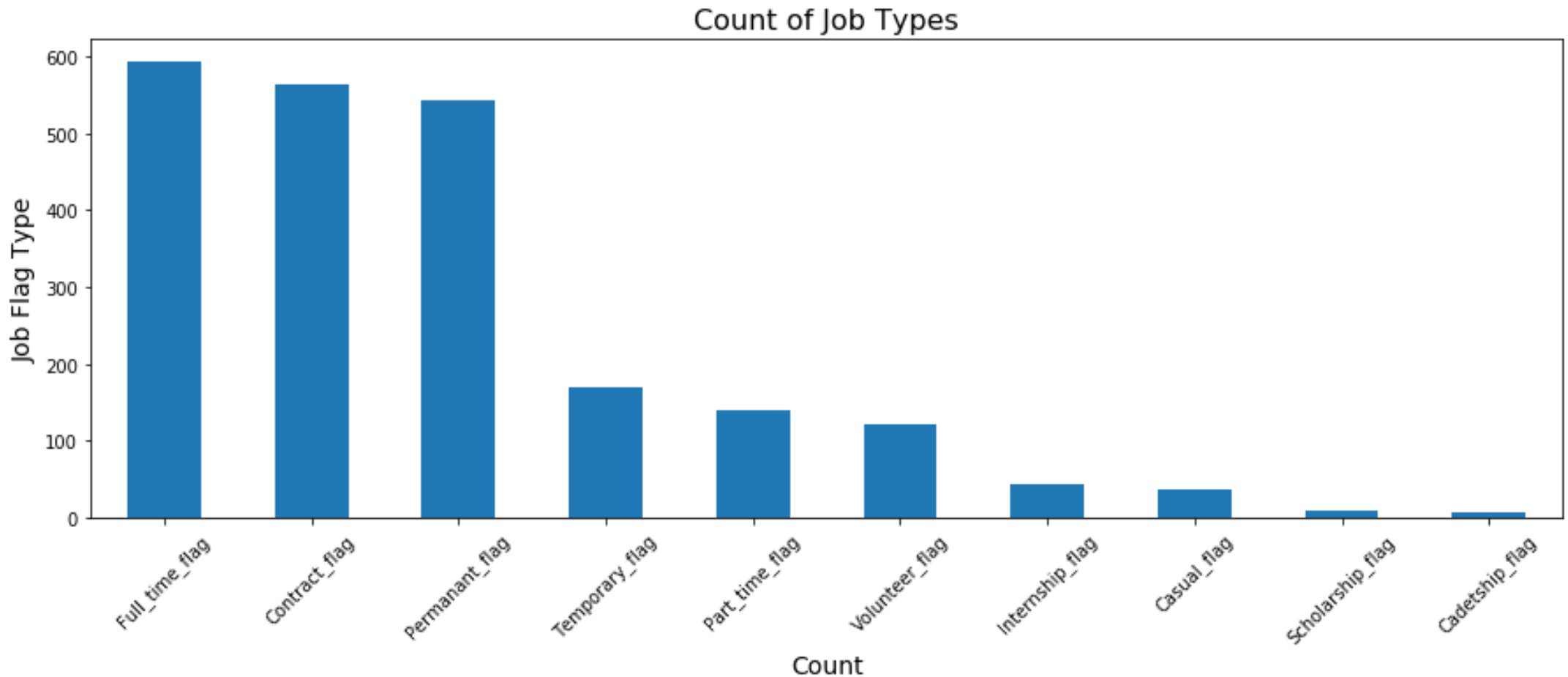
Job Information Frame

'div',{class': 'jobsearch-ViewJobLayout-jobDisplay icl-Grid-col icl-u-xs-span12 icl-u-lg-span7"

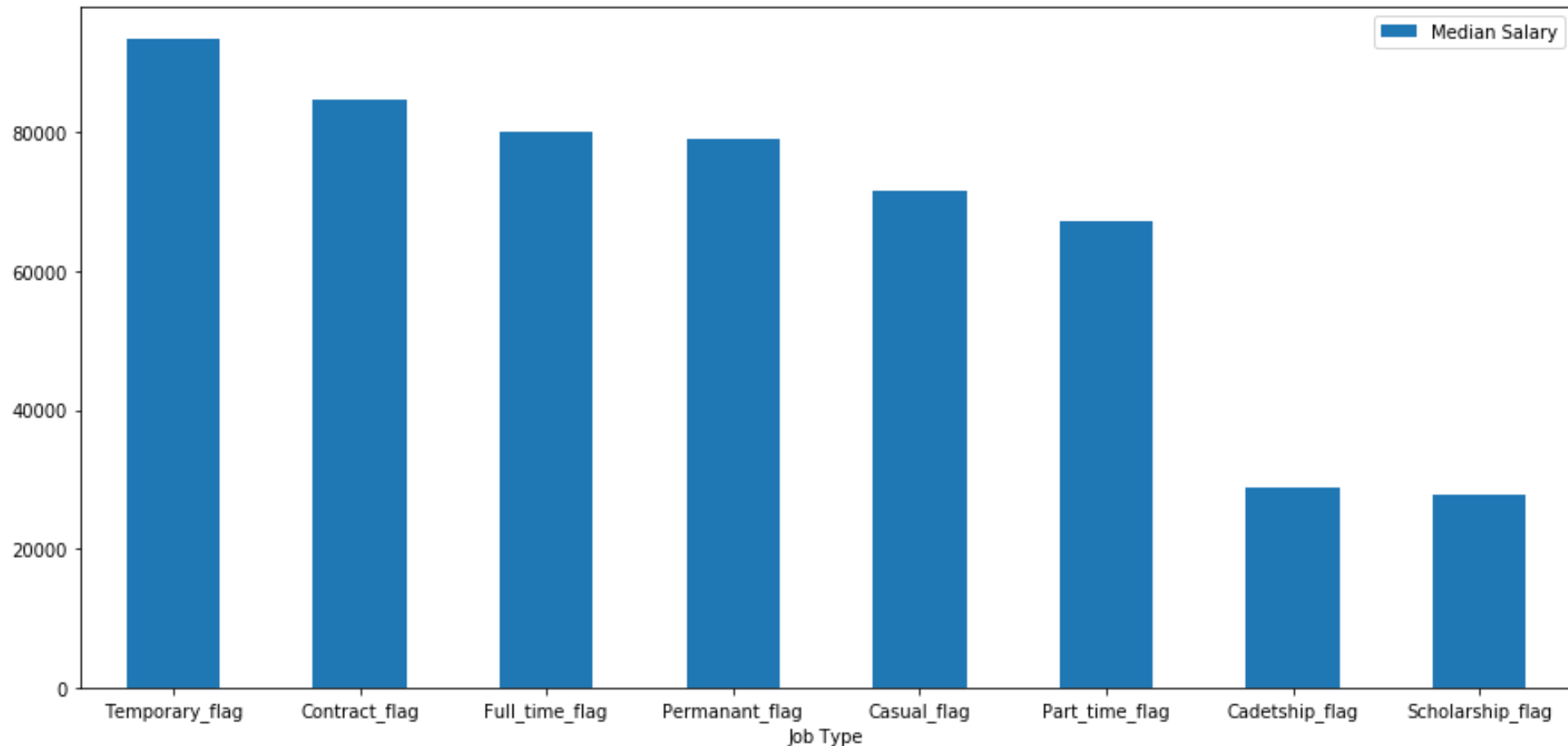
Data scientist, Data Analyst and Data Engineer were the top-3 job titles that were extracted from the job listings



Most roles were full-time, contract, or permanent, with smaller counts for temporary and part-time positions



Highest median salary was for short-term positions (temp/contract), which may be compensation for lack of other benefits (e.g. job security)



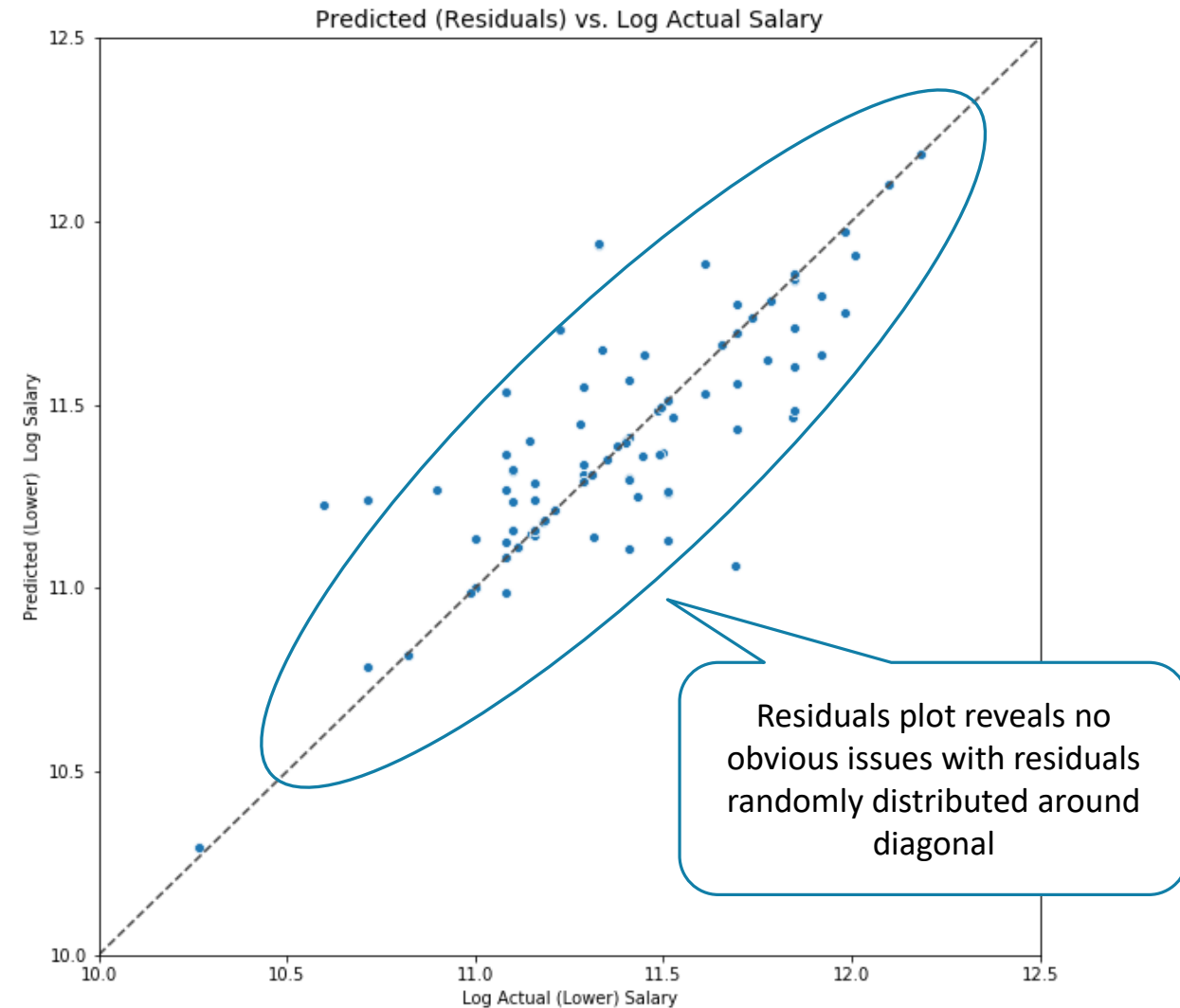
Predicting Salary from Job Postings

Question 1



Linear Regression

Linear Regression model did quite well in predicting **(log) salary** in hold out test set, with an R^2 value of **70.9%**

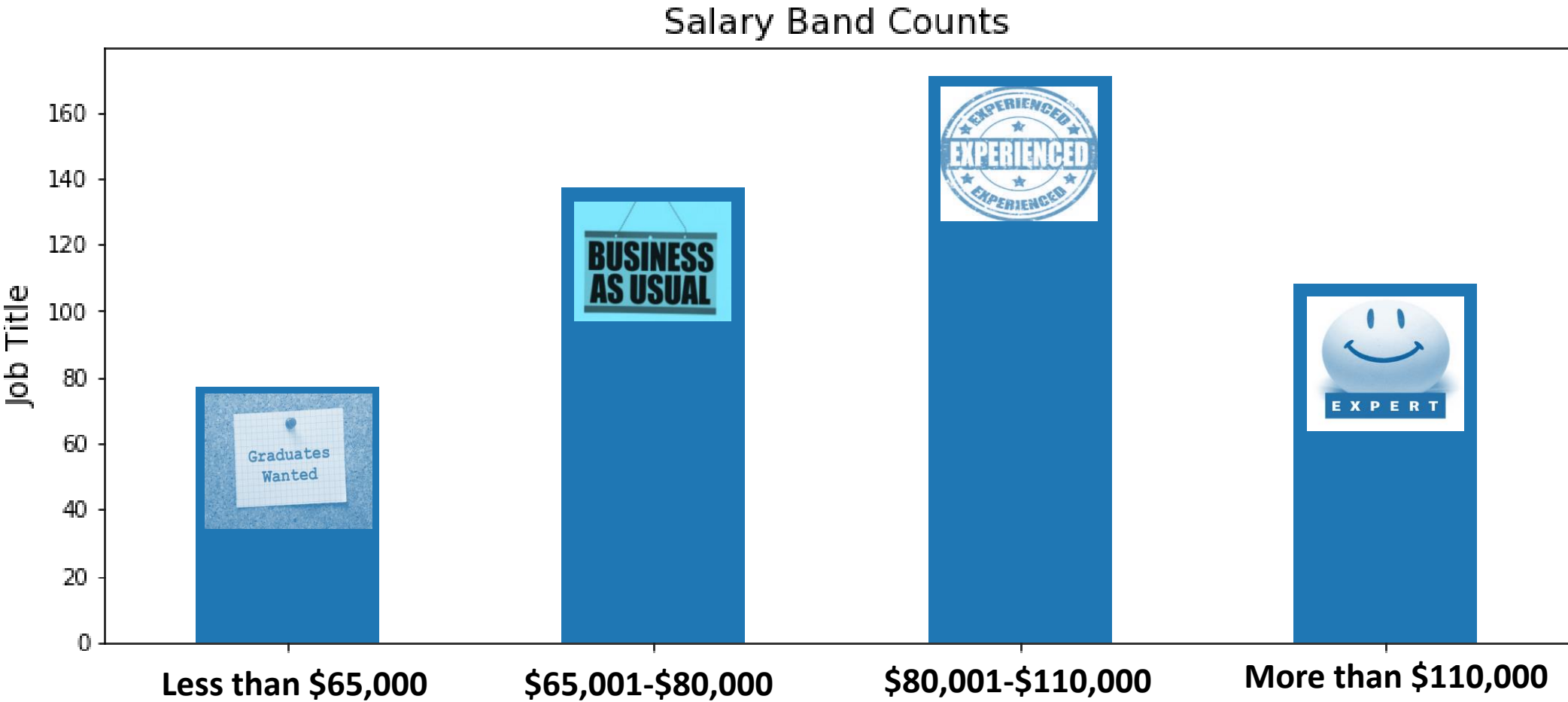


GridSearchCV best parameters: 'tfidfvectorizer__max_df': 150, 'tfidfvectorizer__max_features': 3200, 'tfidfvectorizer__min_df': 8, 'tfidfvectorizer__ngram_range': (1, 3)

Existence of 'Manager' strongest positive contributions to variance explained in (log) salary with 'graduate' and 'entry' strongest negative contributions



Four salary groups were used for classification approach, with most salaries falling in \$80,001-\$110,000 band



Random Forest Classifier was **not** very accurate ($f1=0.76$) with classifications in the \$80,001-\$110,000 range typically misclassified in classes either side. Feature importance also somewhat less interpretable

Accuracy Score: 0.7642276422764228

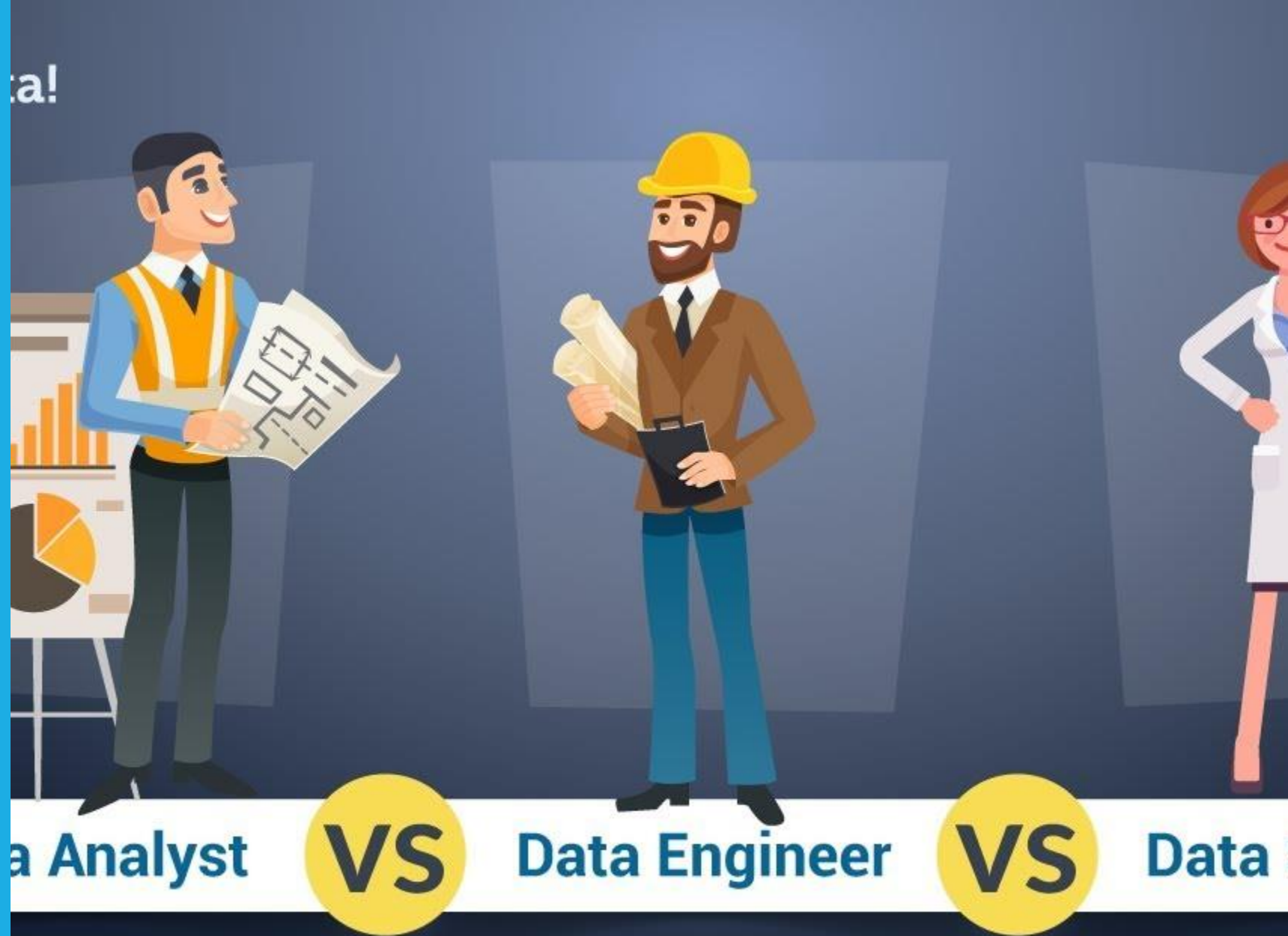
Classification report

	precision	recall	f1-score	support
\$65,001-\$80,000	0.93	0.67	0.78	39
\$80,001-\$110,000	0.58	0.90	0.71	39
Less than \$65,000	1.00	0.81	0.90	16
More than \$110,000	0.91	0.69	0.78	29
accuracy			0.76	123
macro avg	0.86	0.77	0.79	123
weighted avg	0.82	0.76	0.77	123

	Feature	rf_Importance
2141	pricing	0.004849
1823	module lead	0.003898
886	domain	0.003593
1312	health objective	0.003567
2146	principles	0.003491
1754	matrix mapping requirements	0.003353
1622	learning methodologies	0.003329
2369	requests	0.003292
1365	https ey	0.003285
106	advanced	0.003206
350	broader	0.003175
1311	health care	0.003150
1881	non financial	0.003082
2113	pre employment	0.002893
2492	run execute	0.002790
2493	run execute developed	0.002706
313	basic	0.002698

Distinguishing Job Types/Titles

Question 2



Approach to predict a simple **binary classification** of **data science** vs. **non-data science** roles, with target defined by whether “data scientist” in job description (not title)

Job Title Column

Number of jobs with 'data scientist' in job_title column: 242
This represents 41.3 % of originally scrapped data

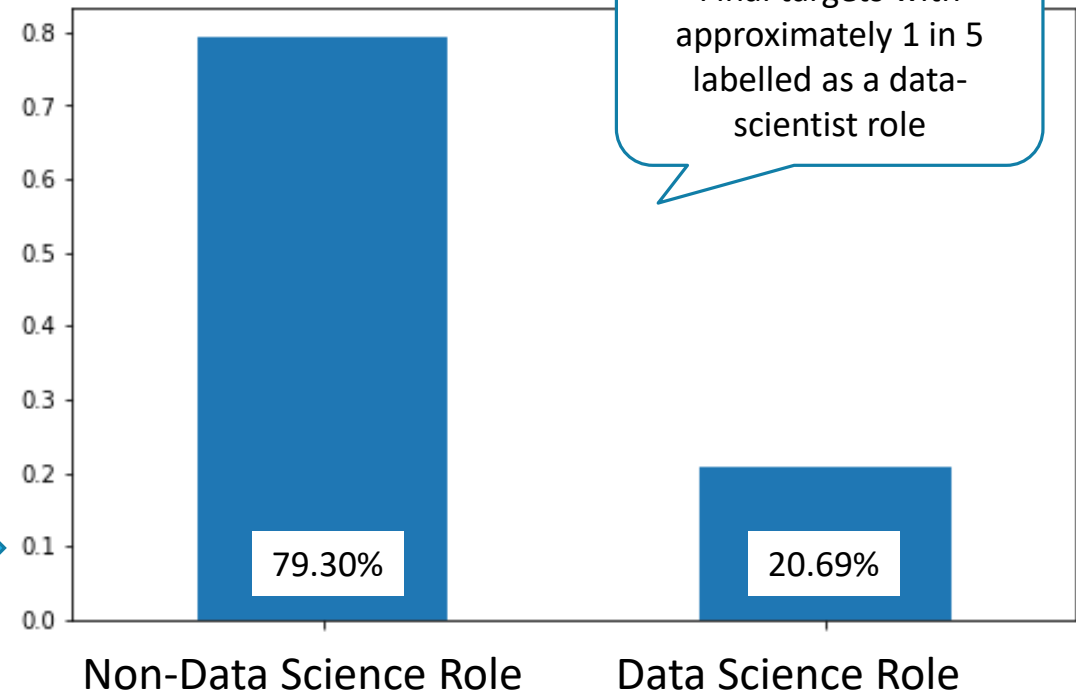
Job Description Column

Number of jobs with 'data scientist' in description: 481
This represents 82.1 % of originally scrapped data

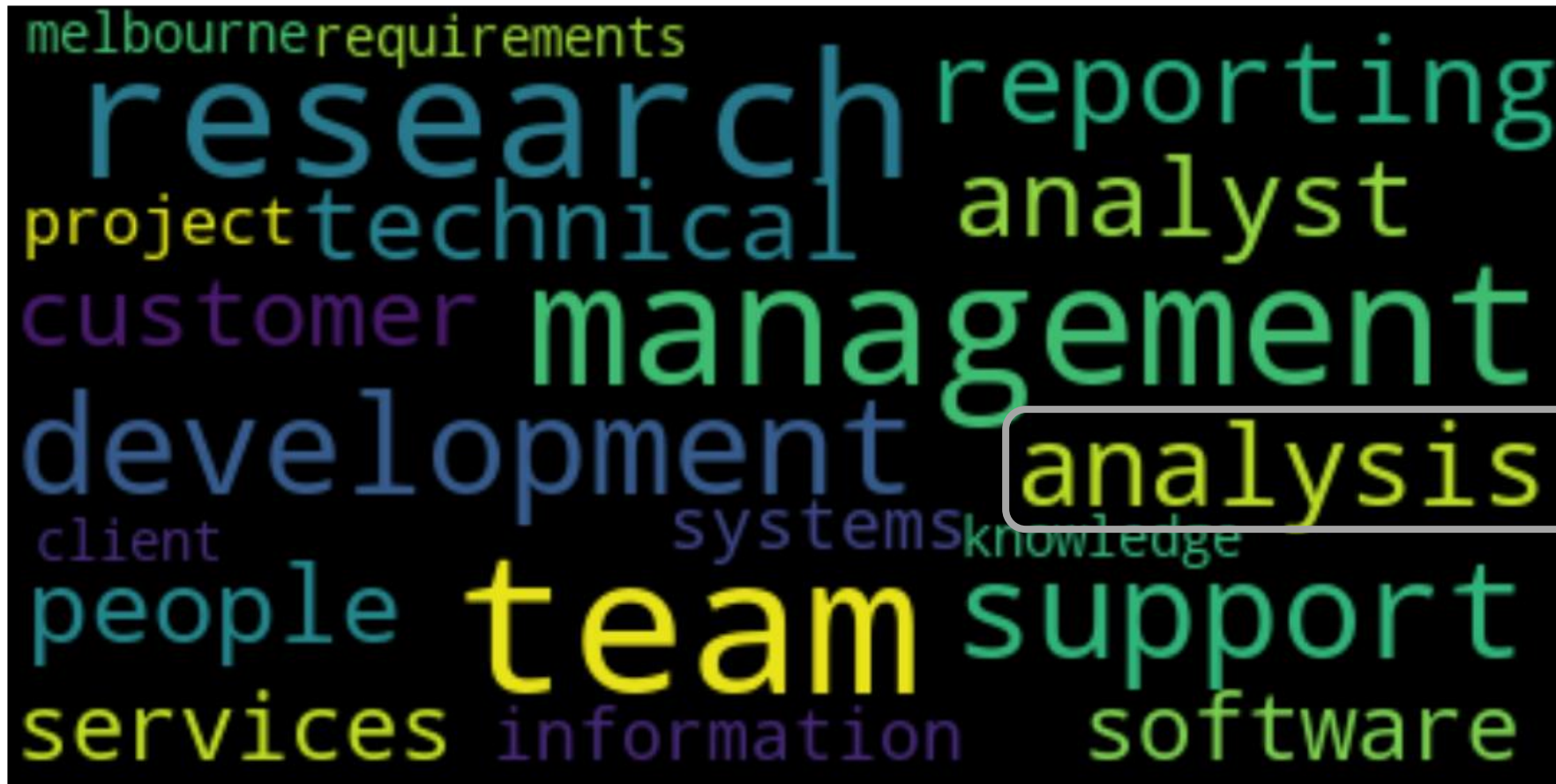
Use of description
when searching results
in larger number of
roles identified, which
will represent target

Final Target

Data Science Roles n = 481
Total roles n = 2324



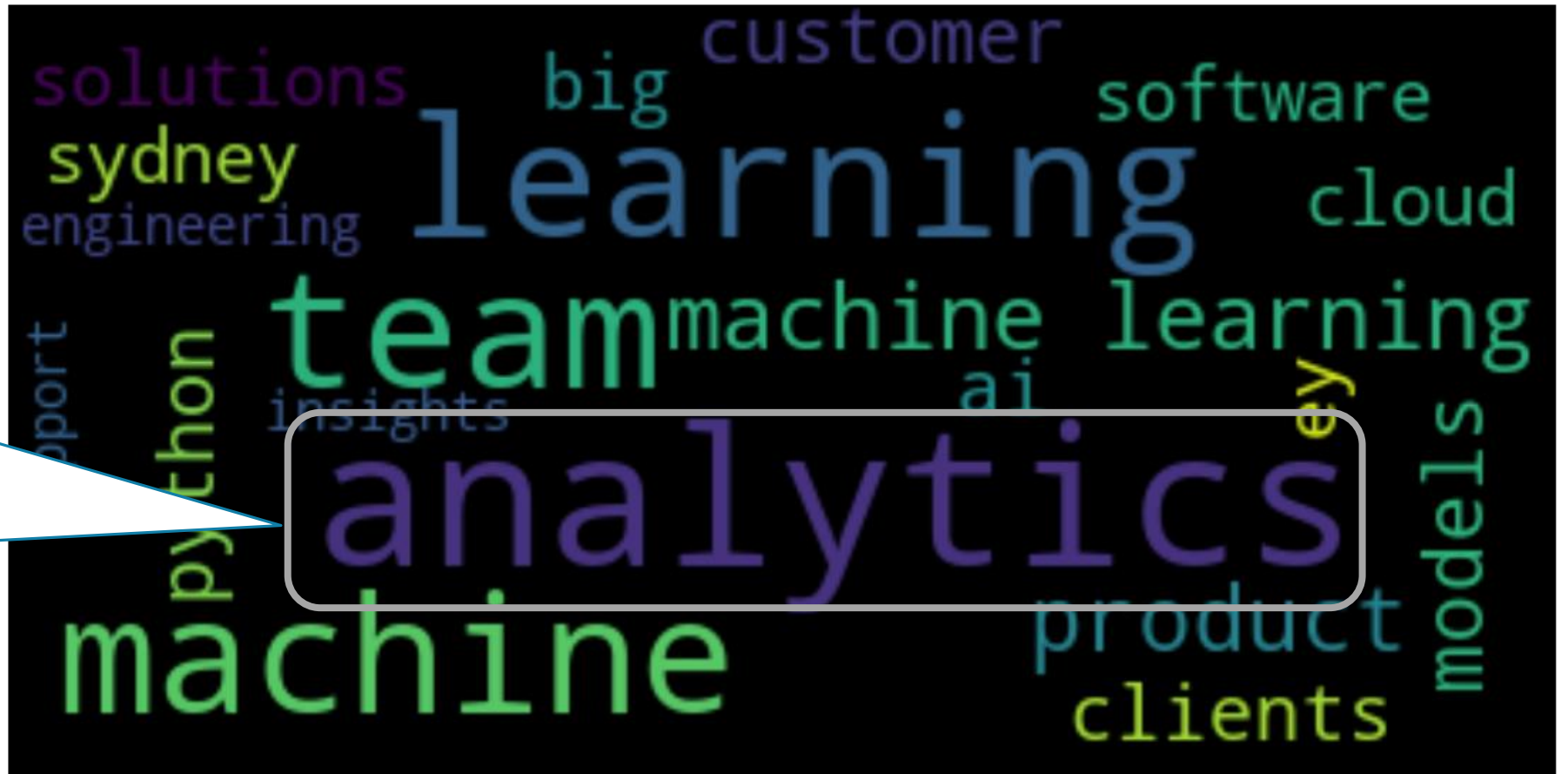
The top 20 most common top words (relative frequency) for **non-data science roles** included 'team', 'research' and 'management'



'Analysis' implies emphasis of analysis on PAST events...


Among the most common top words (relative frequency) for data science roles included 'analytics', 'learning' and also 'team'. Other important features included 'ai' and 'python'

... 'Analytics' though implies emphasis of analysis on FUTURE events and so the difference do make sense in terms of relative importance to both roles

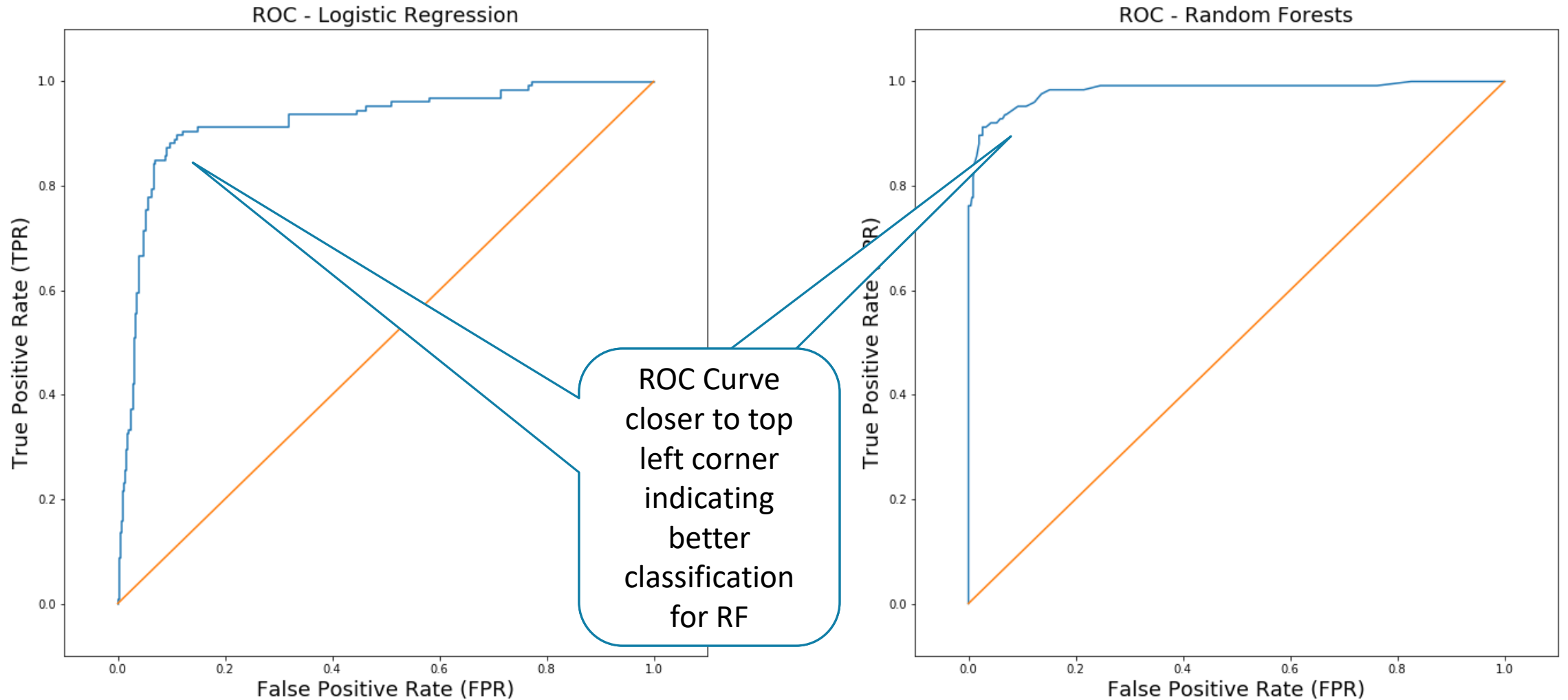


Among logistical regression and Random Forest, the latter showed higher model performance, with an accuracy of 93.8% (vs. baseline 79.3%) and ROC score (.983)

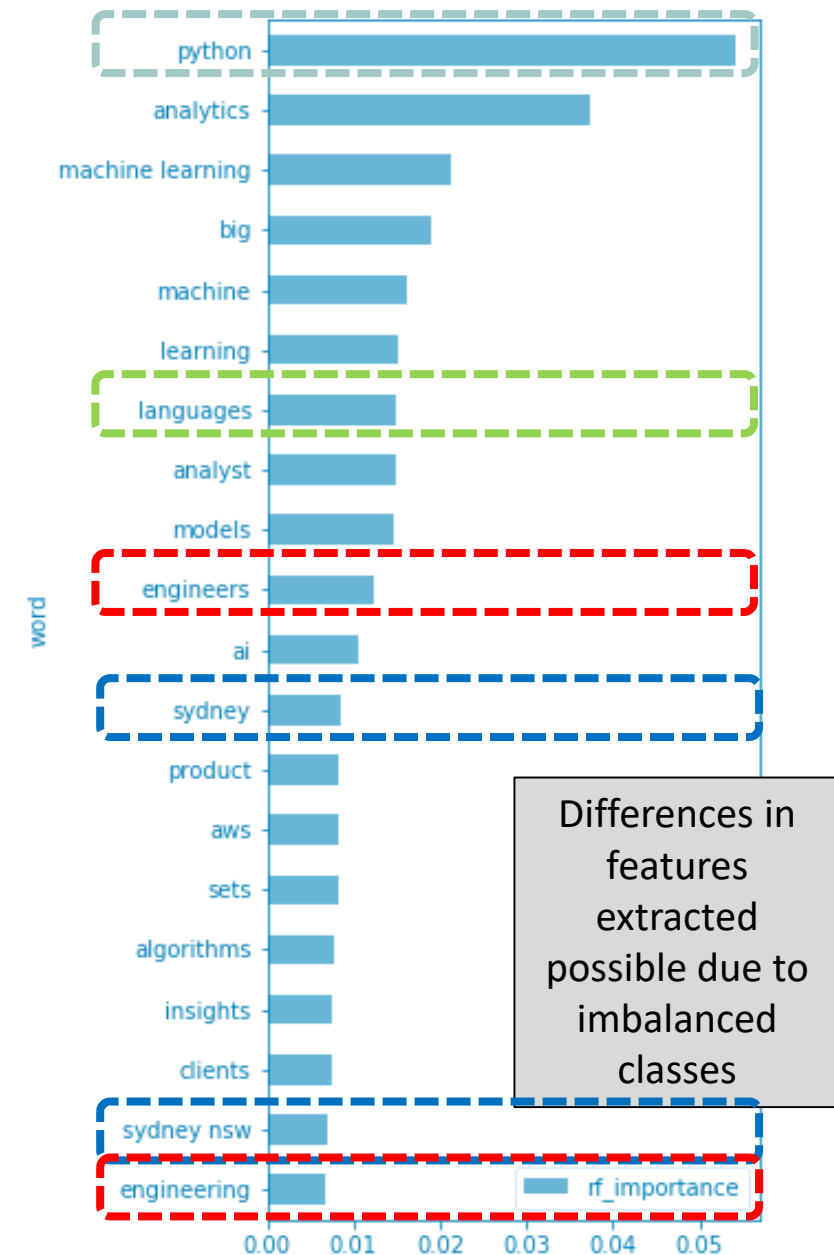
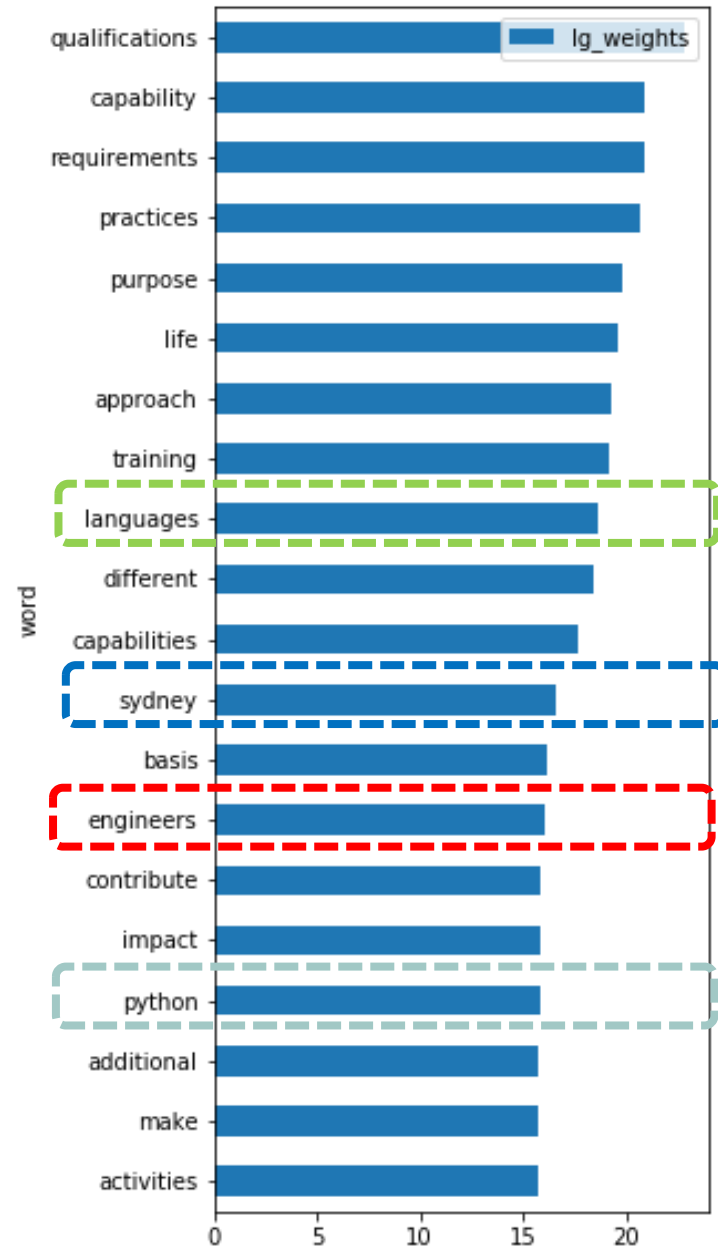
All modelling used Cross Validation with CV=10; RandomState=42

	LogisticRegression()	RandomForestClassifier()
Best parameters (via gridsearch)	'logisticregression__C': 20 'logisticregression__penalty': 'l1'	'criterion': 'entropy', 'max_depth': 20, 'max_features': 50
Highest Accuracy (<i>Std</i>)	.901 (.032)	.938 (.020)
ROC Score	.921	.983
		

Plotting of ROC curves shows Random Forests as superior in classification of Data-science vs. Non-Data science roles



Top 20 features extracted by **Random Forests** are more interpretable than those identified by **logistical regression** with key skills and competencies (e.g. Python; machine learning; AI) relevant to data scientists being identified





**Thank
you!**