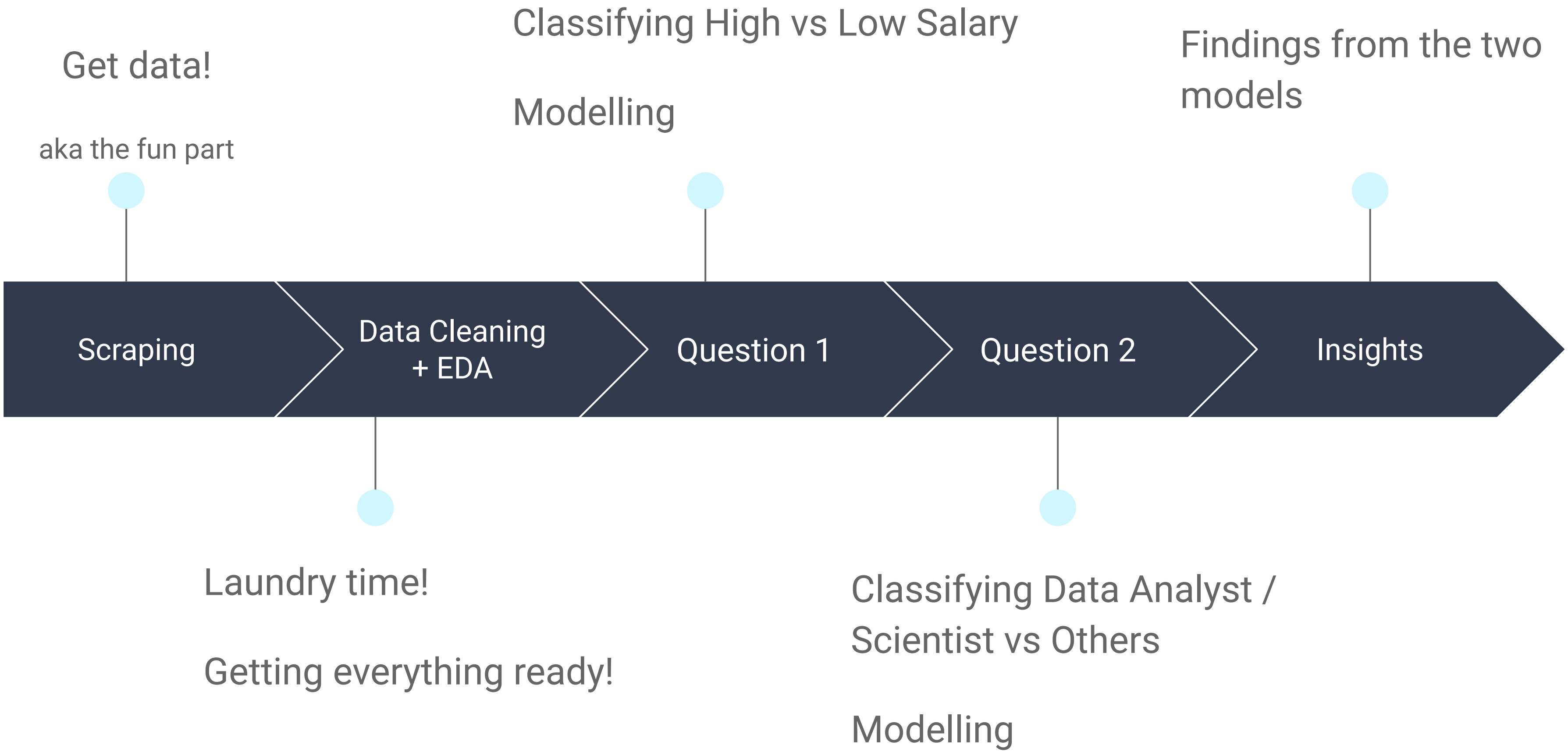
DSI-7 Project 4

by Vincent Kwan

Web Scraping
Perdicting & Classifying Data-related Jobs
NLP Techniques



Scraping

Get data!

aka the fun part

Data Cleaning + EDA

Laundry time!

Getting everything ready!

Question 1

Classifying High vs Low Salary

Modelling

Question 2

Classifying Data Analyst / Scientist vs Others

Modelling

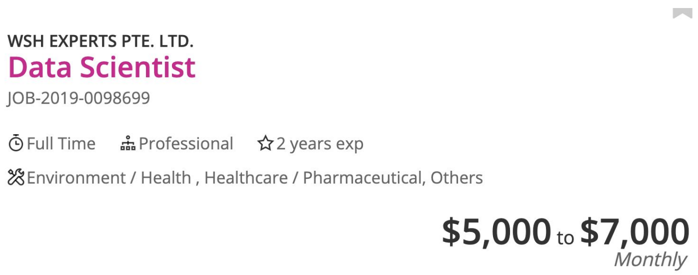
Insights

Findings from the two models

Web Scraping

- Careersfuture.sg
- BeautifulSoup + Selenium
- **Search Terms:** [data scientist, data analyst, business analyst, business intelligence, data engineer, data architect, database engineer, research scientist, data governance, data manager, python developer]

- **Features:** Role & Responsibilities + Job Requirements +



WSH EXPERTS PTE. LTD.

Data Scientist

JOB-2019-0098699

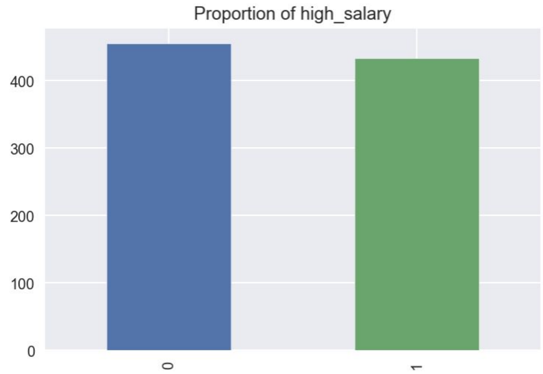
Full Time Professional 2 years exp

Environment / Health , Healthcare / Pharmaceutical, Others

$5,000 to $7,000
Monthly

Data Cleaning & EDA

- Remove leading & trailing whitespaces
- Split salary range to min & max
- Convert salaries to monthly frequency
- Get median between min & max salary
- Drop Role & Responsibilities. Job requirements more meaningful
- Create target variable: >= $6750 = high_salary



```
0    455
1    433
Name: high_salary, dtype: int64
```

Qn 1: High vs Low Salary

Preprocessing:

CountVectorizer (Binary=True):

- Seniority, Industry, Job Requirements

Pick out key skills from Job Requirements:

- Sql, python, scala, hadoop… masters, phd...

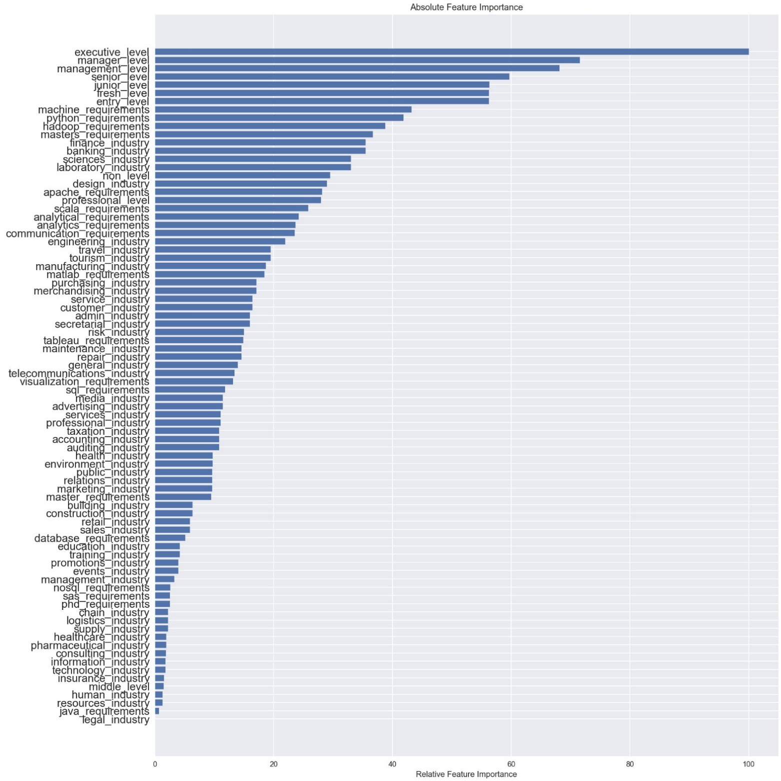
Modelling:

- Baseline = 51%
- StandardScaler
- Vanilla Logistic Regression
- **Logistic Regression with GridSearch**
- Bernoulli Naive Bayes
- Random Forest Classifier with Gridsearch

Qn 1: High vs Low Salary

Insights:

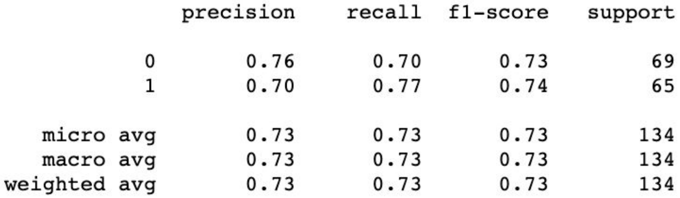
- **Seniority levels** are most influential (would be interesting to see no. of years of experience but data is incomplete)
- Being **skilled** in machine learning, hadoop, apache, matlab, scala and visualisation are very important
- Jobs in **banking and finance** could potentially fetch higher salaries as compared to other industries
- Holding a **masters degree** is important but having a phd is not as important as we think



Qn 1: High vs Low Salary

Tradeoffs between decision thresholds:

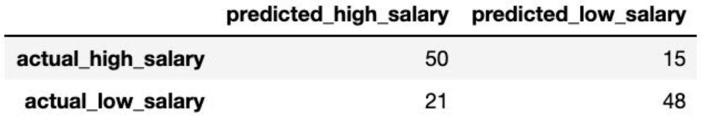
- Completely removing False Positives comes at the expense of overall precision
- Better to lower threshold that determines high/low salary



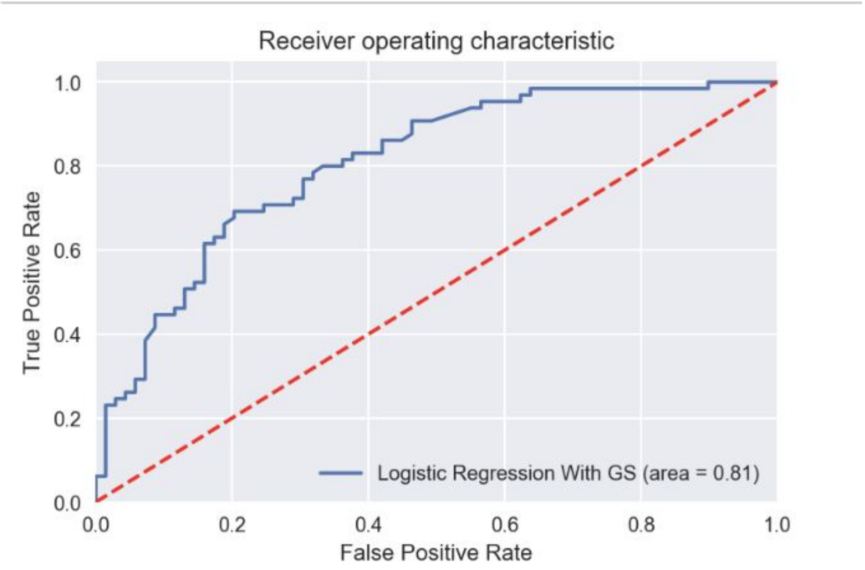
```
              precision    recall  f1-score   support

           0       0.76      0.70      0.73        69
           1       0.70      0.77      0.74        65

   micro avg       0.73      0.73      0.73       134
   macro avg       0.73      0.73      0.73       134
weighted avg       0.73      0.73      0.73       134
```



	predicted_high_salary	predicted_low_salary
actual_high_salary	50	15
actual_low_salary	21	48



Qn 2: Data Analyst/Scientist vs Others

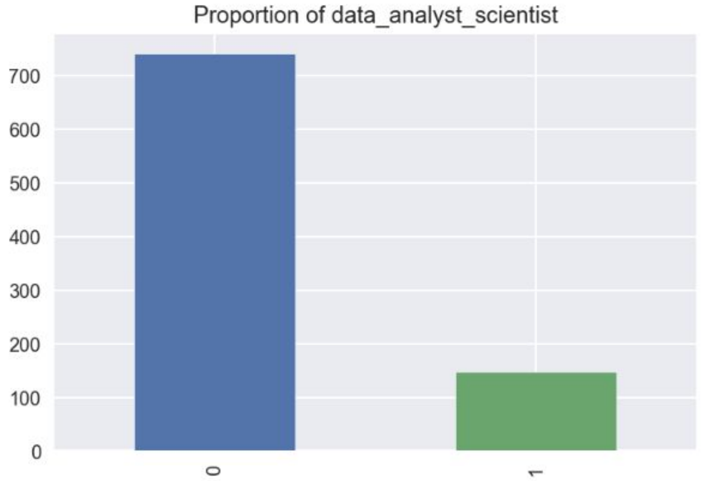
Preprocessing:

TfidfVectorizer on Job Requirements:

- ngram_range=(1,2)
- max_df=0.95
- min_df=0.1

Target variable: if data/scientist/analyst exist in Job_Title, 1. Else 0

SMOTE to overcome class imbalance



```
0     740
1     148
Name: data_analyst_scientist, dtype: int64
```

Qn 2: Data Analyst/Scientist vs Others

Modelling:

Multinomial Naive Bayes

Random Forest Classifier

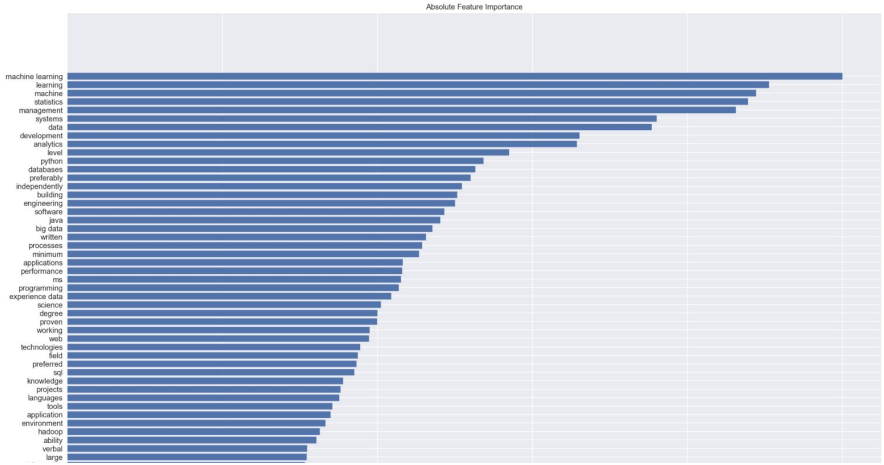
Logistic Regression

Qn 2: Data Analyst/Scientist vs Others

Distinguishing Features of Data Analyst/Scientist:

They should possess knowledge of **machine learning** and **statistics**.

They do not necessary require knowledge of engineering, software, systems or applications.


Absolute Feature Importance
machine learning
learning
machine
statistics
management
systems
data
development
analytics
level
python
databases
preferably
independently
building
engineering
software
java
big data
written
processes
minimum
applications
performance
ms
programming
experience data
science
degree
proven
working
web
technologies
field
preferred
sql
knowledge
projects
languages
tools
application
environment
hadoop
ability
verbal
large