ADZUNA TEAM

John Cusack, Steven Futter, Alyson Lamberti, Jun Long, Will Lukach, Charles Paik

1. Introduction

"Money isn't everything but it's right up there with oxygen."

- Rita Davenport

Man may not base his self-worth on the size of his wallet, but as our introductory quote suggests, it is a necessity. Therefore, he may begin on a pursuit for work. To facilitate the often arduous task of finding a job, Adzuna, the UK-based career search company, is aiming to boost and enhance a user's experience by providing predictions on salaries for posted job positions. To ensure credibility and accuracy, Adzuna is enlisting the help and support of a data science community by way of a competition.

This analysis is a proposed solution to Adzuna's request. It begins with an understanding of the data through a data quality check, and its variable relationships through an exploratory data analysis. This provides a guide into the modeling process. Due to the various variable types within the data set, multiple modeling techniques were assessed and combined to develop a single predictive solution. Each model was evaluated on the training data by goodness-of-fit measures. The predictive results of final models were then evaluated on the test data for accuracy. The best resulting model was ...

2. The Modeling Problem

Adzuna requires a prediction engine that predicts the average salary for any given UK job ad. The data provided is already split into a training data set, of which we utilize for model creation, and a test data set, used to analyze the accuracy of the model. The data consists of both structured and unstructured variables taken from their database of UK job ads that already are associated with a salary value.

The open source R software is the platform of choice for this analysis. Multiple R packages including ... will be used to develop the solution. The primary challenge facing this analysis lies in the the unstructured textual nature of much of the data set. In addition, while it is a popular, if not the de facto and go-to language for most data science teams, R is not particularly well-suited for language processing. Despite this challenge, the … and … packages will be used to aid in exploring and predicting this data type.

Literature Review

The job salary prediction competition was won by Vlad Mnih shortly after he had just completed his PhD in Machine Learning at the University of Toronto in 2013. In a Kaggle article, "Q&A With Job Salary Prediction First Prize Winner Vlad Mnih", he explains that relatively little preprocessing and feature engineering was needed to produce optimal results. By using a separate bags of words for the job title, description, and the raw location, and stemming the words in the title and description using a Porter stemmer technique he was able to train and

use a neural network to achieve a mean absolute error (MAE) of 3435. Mnih did not combine neural networks with any other learning methods.

"Predicting Job Salaries from Job Descriptions", by Shaun Jackman and Graham Reid, provides an alternative set of approaches to the ADZUNA salary prediction problem. Jackman and Reid tested a variety of regression methods, including maximum-likelihood regression, lasso regression, artificial neural networks and random forests. Using each of these methods Jackman and Reid optimized parameters and validated the performance of each model using cross validation. A log-transform of the output variable (salary) was used since "without this transform, for a linear model each word of the description would contribute a fixed amount to the salary, such as £5,000 for the word 'manager', or –£10,000 for the word 'intern'." Using the log of the output variable Jackman and Reid explain that each word adds to the salary via a multiplicative factor, where "the word 'manager' may cause the salary to be multiplied by 1.25, whereas the word 'intern' may cause the salary to be multiplied by 0.50."

3. The Data

The Adzuna data set consists of 244,768 observations across 12 variables. All the predictors are of text type and three are missing a large number of values. The Job ID variable was converted in R to a Text variable due to its definition. We provide a quick survey of the data that lies before us in Table 1.

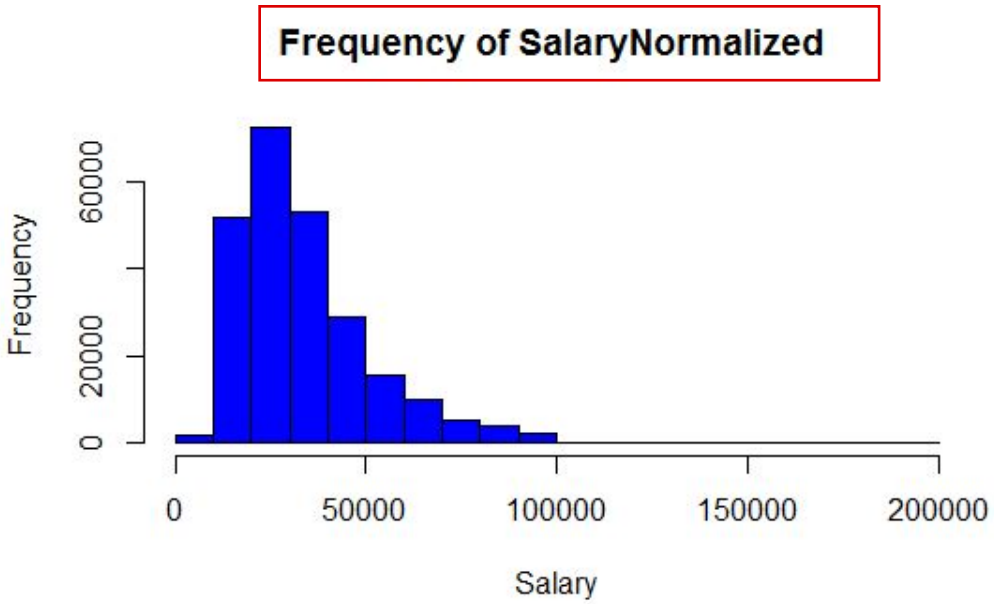
#	Variable	Description	Type	Missing
1	**Id**	Job ID	Text	
2	**Title**	Job Title	Text	
3	**FullDescription**	Long Job Description	Text	
4	**LocationRaw**	Job Location (User-Entered)	Text	
5	**LocationNormalized**	Job Location (Extracted)	Text	
6	**ContractType**	Position Type	Text	179326
7	**ContractTime**	Position Length	Text	63905
8	**Company**	Company Name	Text	32430
9	**Category**	Job Category	Text	
10	**SalaryRaw**	Job Salary Range (User-Entered)	Text	
11	**SalaryNormalized (RESPONSE)**	Job Salary (Extracted)	Numeric	
12	**SourceName**	Source of Job Description	Text	

(Table 1)

It is observed in these definitions that some of the variables contain the same information in slightly different ways, signifying the need to filter out one of the alike variables for modeling.

Since the SalaryNormalized variable is both the response as well as the only numeric variable, it is explored first. The mean salary of £34,123 is only slightly higher than the median

salary of £30,000 indicating the possibility of skewed data, an observation confirmed by the histogram shown in Figure 1.



(Figure 1)

With a positively-skewed distribution of the response variable, further investigation is warranted in the proper handling of the response variable for predictive models. Often such a skewness leads to a log transformation of the variable.

As for the predictors, Table 2 shows these nine variables and the number of unique values in each.

Variable	Numer of Unique Values
LocationRaw	20986
LocationNormalized	2732
ContractType	2
ContractTime	2
Company	20812
Category	29
SalaryRaw	97286
SourceName	167
Title	135435

(Table 2)

Those variables with a high number of unique values indicate the possible need to create new dummy variable groups. The ID variable was left out of this list, as it is exists to only uniquely identify each observation.

The FullDescription variable was also left out since each observation will also have a unique, or no description. Since it is also is a variable composed of a block of text, a world cloud was created to get a better understanding of the type of context the majority of the descriptions entail. The word cloud in Figure 2, shows the most common text as the largest and going from the top down in the image. It leads with Experience, Role and Work.

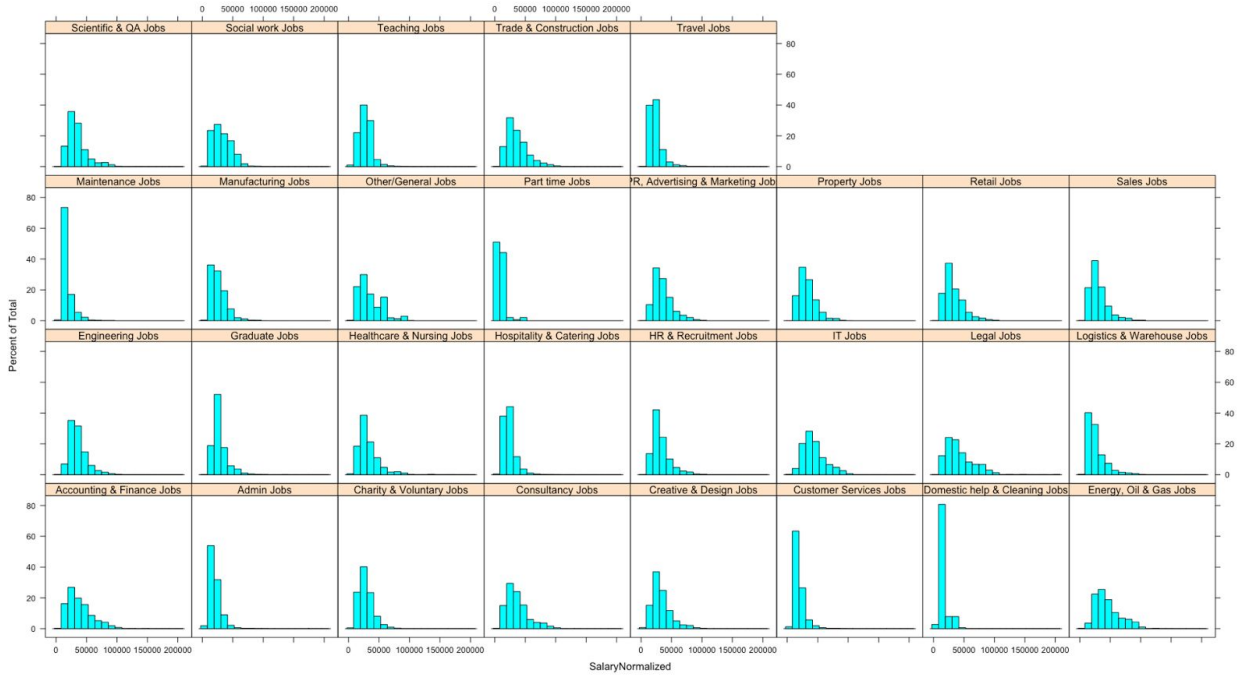

EXPERIENCE ROLE WORK
TEAM BUSINESS SKILLS WORKING JOB
SALES CLIENT MANAGEMENT MANAGER
COMPANY DEVELOPMENT UK SUPPORT EXCELLENT
SERVICE REQUIRED OPPORTUNITY RECRUITMENT
SUCCESSFUL KNOWLEDGE APPLY Â CUSTOMER BASED
SERVICES ABILITY STRONG ENSURE CANDIDATE PROJECT HIGH JOIN
SALARY ENVIRONMENT DESIGN TRAINING GOOD INCLUDING LEADING
CARE CLIENTS WWW CV TECHNICAL POSITION KEY CANDIDATES
EMPLOYMENT QUALITY PROVIDE ESSENTIAL CONTACT LEVEL FULL
OPPORTUNITIES CAREER SYSTEMS TIME REQUIREMENTS INFORMATION
BENEFITS SENIOR ENGINEER EXPERIENCED AGENCY SOFTWARE PROJECTS
STAFF POSTED RESPONSIBLE JOBSEEKING ENGINEERING DEVELOP ORIGINALLY
PART MARKETING LONDON COMMUNICATION PEOPLE APPLICATIONS INCLUDE

(Figure 2)

4. Exploratory Data Analysis

We begin by taking a look at the distribution of salaries across job categories. As can be seen in the lattice histogram plot of job categories below the distribution of salaries are positively skewed. ~~From the summary() function~~ the mean salary of 34,123 is slightly higher than the median salary of 30,000 and the density plot of normalized salary is positively skewed. To ensure our models produce optimal results we will therefore use log-transformed SalaryNormalized as the response variable.

Histograms of SalaryNormalized by Job Category



(Figure 3)

The table below presents the mean salary by job category in descending order of magnitude. We see that Energy, Oil & Gas Jobs are the highest mean salary earners with a mean salary of £45,653 followed by IT Jobs (£43,983), and Legal (£42,649). Not surprisingly, the lowest paid job category by mean salary is Part time Jobs with a mean salary of £10,030.

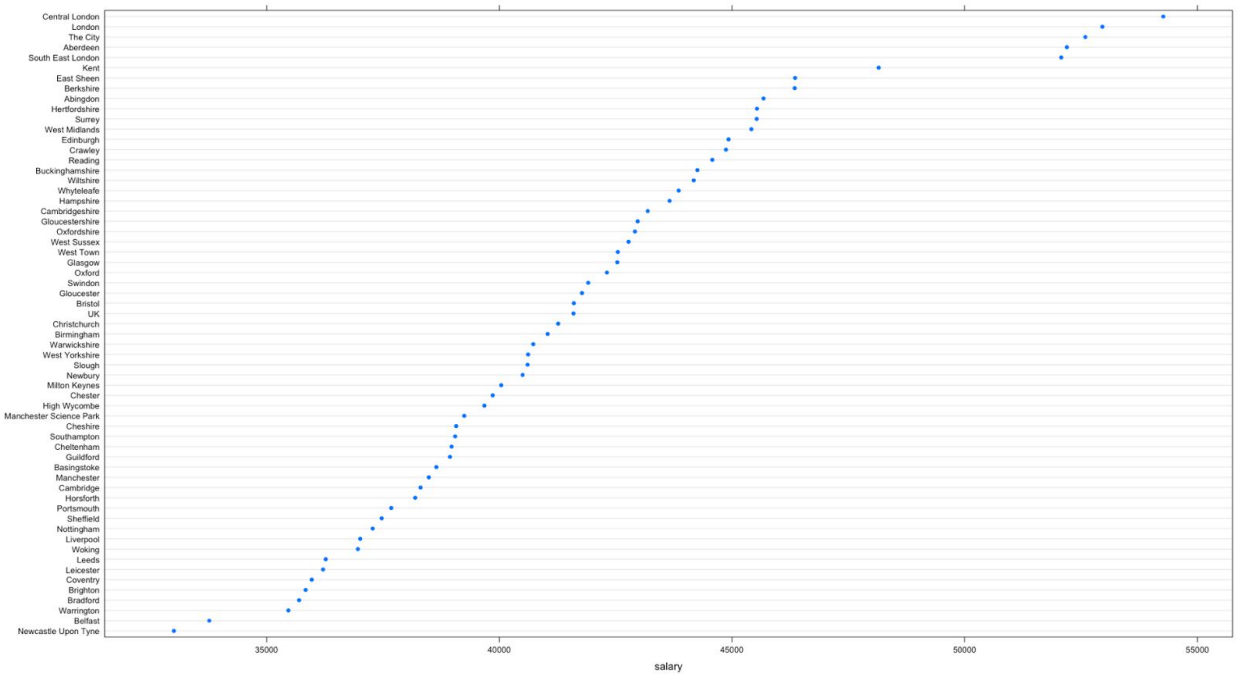
Mean Salary by Job Category (Table 3)

Category <fctr>	salary <dbl>	n <int>
Energy, Oil & Gas Jobs	45653.09	2255
IT Jobs	43983.91	38483
Legal Jobs	42649.00	3939
Accounting & Finance Jobs	38751.22	21846
Consultancy Jobs	37028.87	3263
Trade & Construction Jobs	36406.66	8837
Engineering Jobs	35838.27	25174
PR, Advertising & Marketing Jobs	35593.79	8854
Other/General Jobs	35346.43	17055
Scientific & QA Jobs	34436.93	2489
Creative & Design Jobs	33173.54	1605
Retail Jobs	32955.73	6584
HR & Recruitment Jobs	32589.86	7713
Healthcare & Nursing Jobs	32589.24	21076
Property Jobs	32512.81	1038
Social work Jobs	32381.34	3455
Sales Jobs	30814.88	17272
Charity & Voluntary Jobs	28272.92	2332
Graduate Jobs	28107.51	1331
Teaching Jobs	27671.02	12637
Manufacturing Jobs	26497.90	3765
Logistics & Warehouse Jobs	26497.80	3633
Travel Jobs	23838.97	3126
Hospitality & Catering Jobs	23702.74	11351
Admin Jobs	21053.66	7614
Customer Services Jobs	19861.44	6063
Maintenance Jobs	17726.06	1542
Domestic help & Cleaning Jobs	17553.62	291
Part time Jobs	10030.07	145

29 rows

There is a significant jump in average salary for jobs based in and around London. Central London, London, The City, and South East London are all in the top five highest paying regions of the UK. Apparent from the dotplot is the large decrease in average salary as you leave the capital city. With the exception of Aberdeen in Scotland, which has the fourth highest average salary, average salaries decrease by approximately £5,000 for regions outside of London.

Average Salary By Region Dotplot (Figure 4)



5. Predictive Modeling: Methods and Results

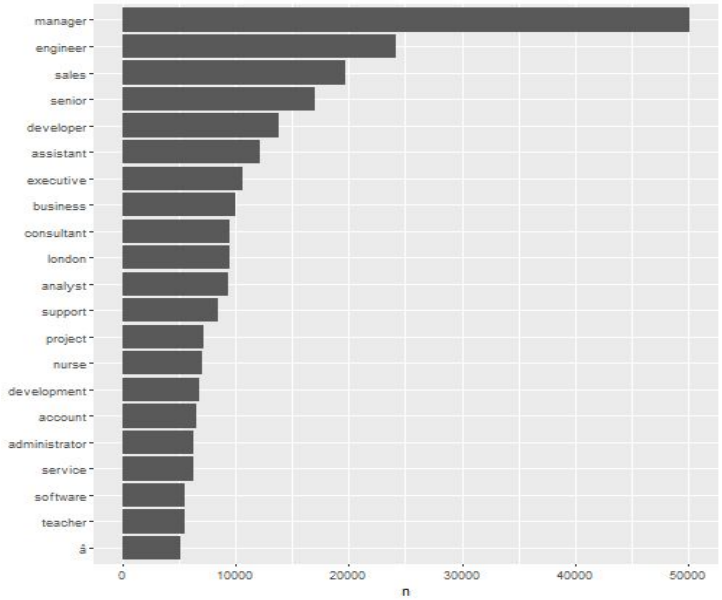
6. Comparison of Results

7. Conclusions

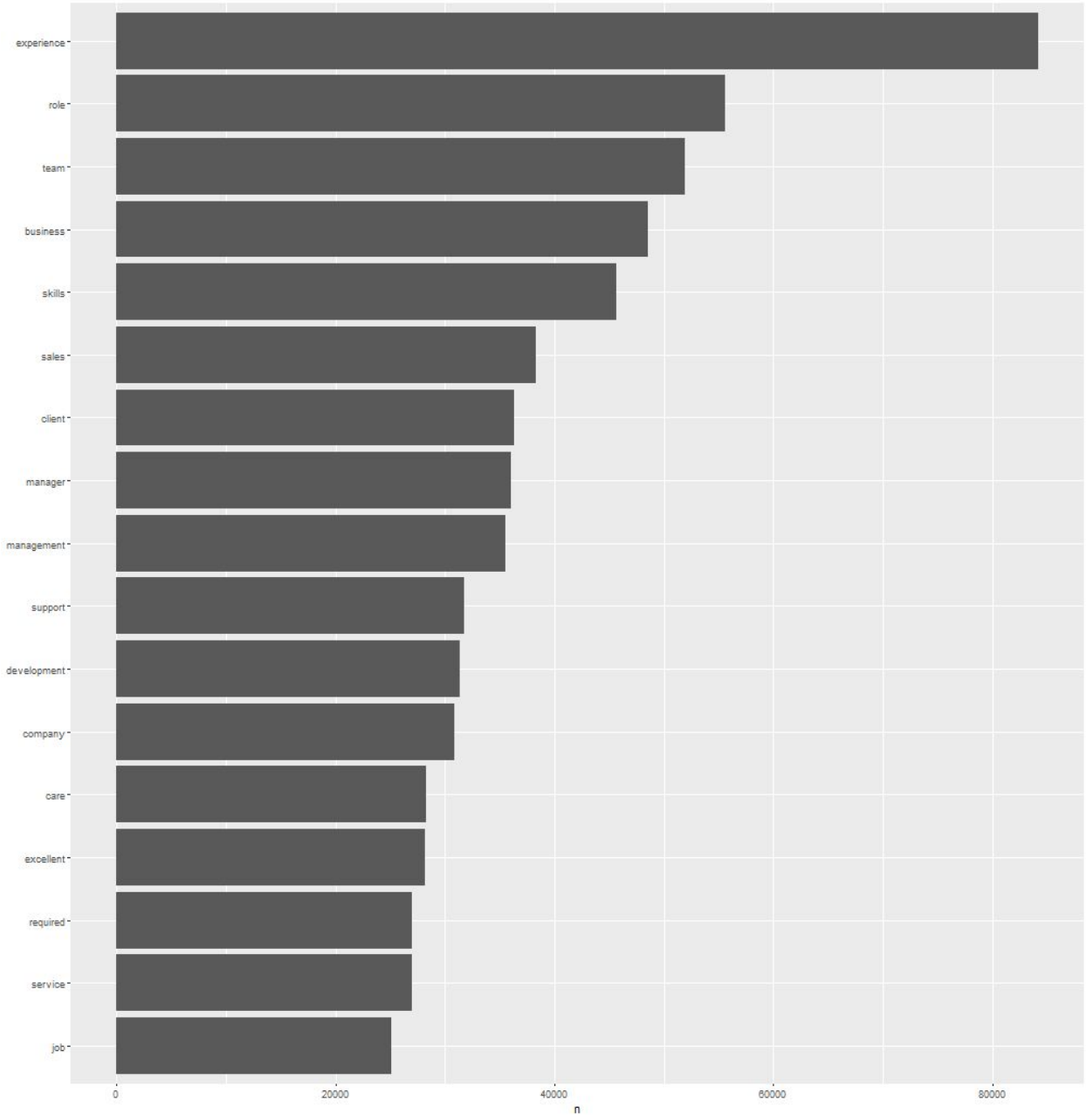
8. Bibliography

X. Appendix

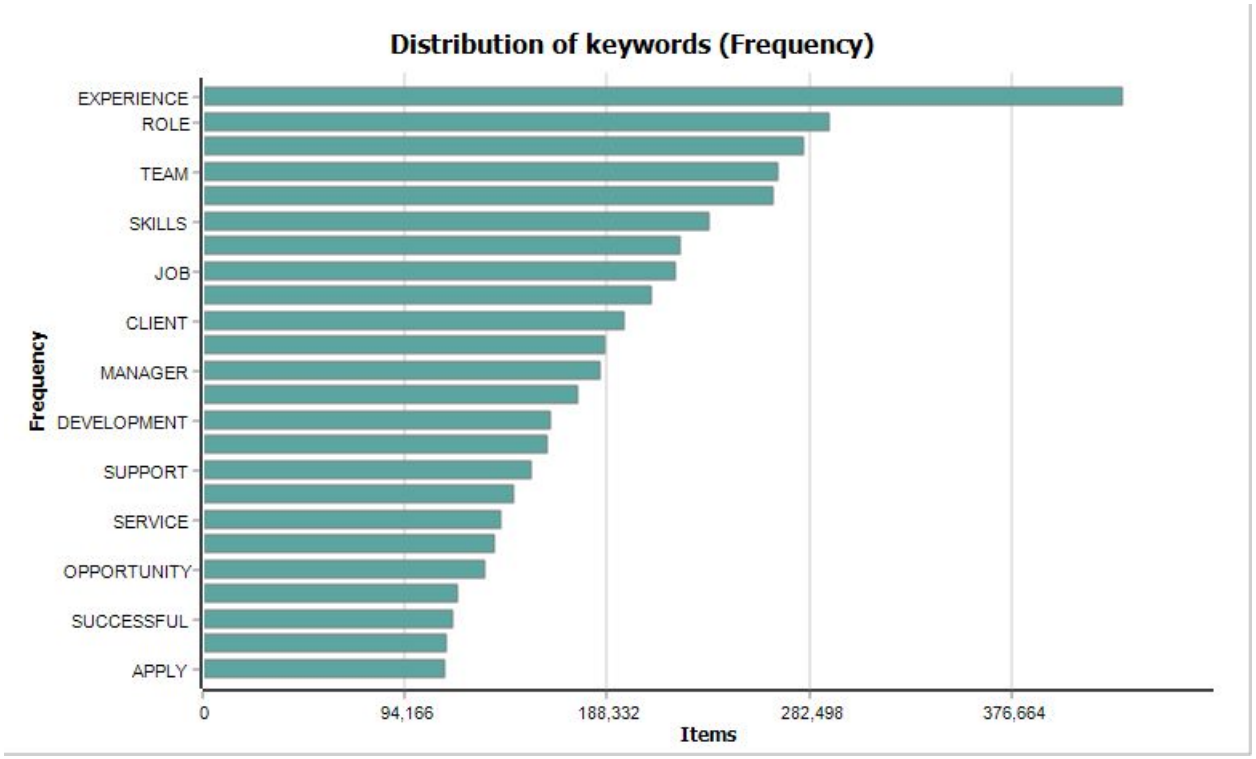
Title - Frequency Plot of Terms

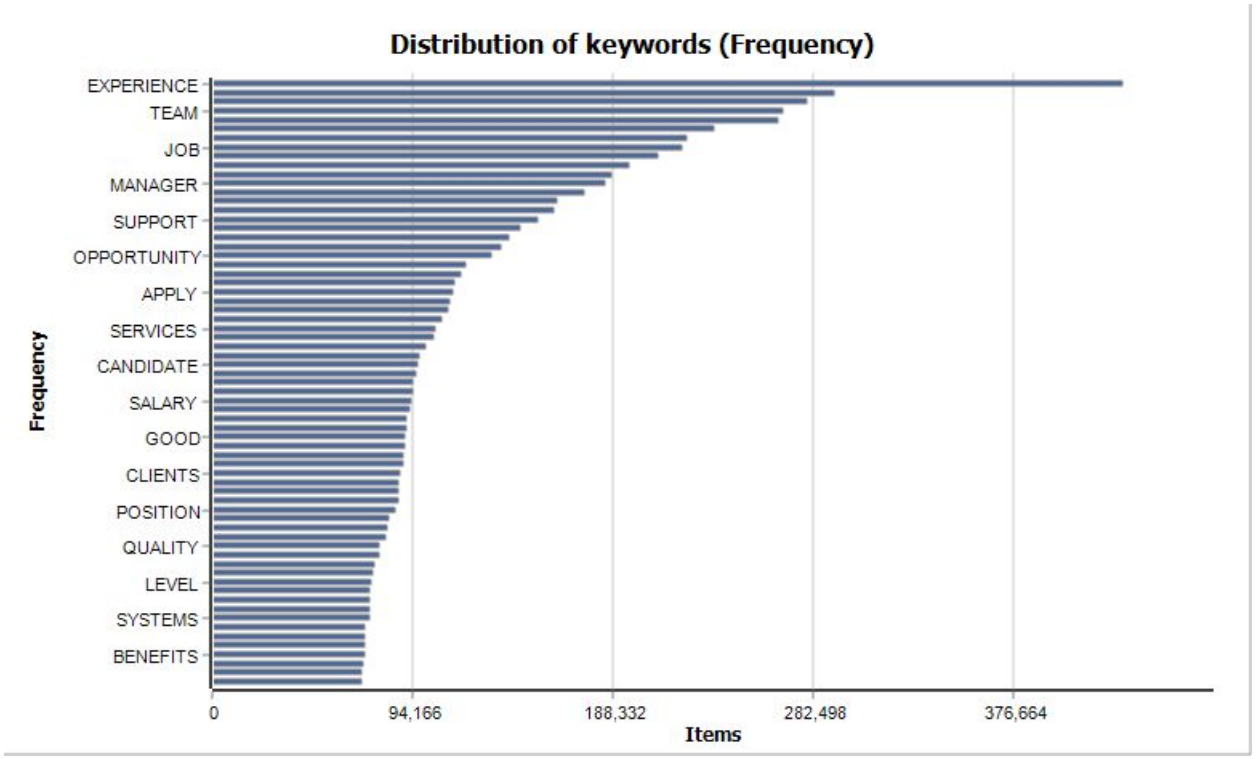

manager
engineer
sales
senior
developer
assistant
executive
business
consultant
london
analyst
support
project
nurse
development
account
administrator
service
software
teacher
á
0
10000
20000
30000
40000
50000
n

FullDescription - Frequency Plot of Terms (first 50,000 rows)

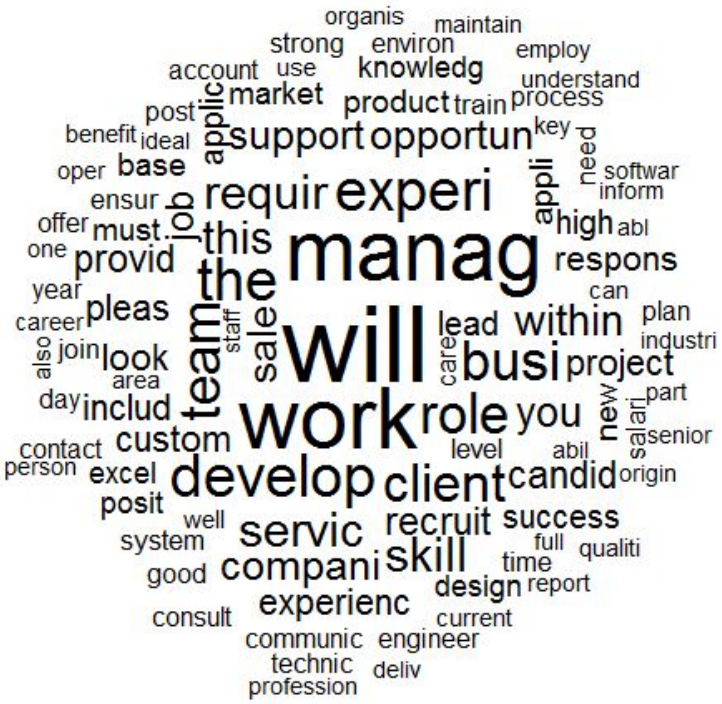


Word Frequency for Training Set:


Distribution of keywords (Frequency)
EXPERIENCE
ROLE
TEAM
SKILLS
JOB
CLIENT
MANAGER
DEVELOPMENT
SUPPORT
SERVICE
OPPORTUNITY
SUCCESSFUL
APPLY
Frequency
0
94,166
188,332
282,498
376,664
Items


Distribution of keywords (Frequency)
EXPERIENCE
TEAM
JOB
MANAGER
SUPPORT
OPPORTUNITY
APPLY
SERVICES
CANDIDATE
SALARY
GOOD
CLIENTS
POSITION
QUALITY
LEVEL
SYSTEMS
BENEFITS
Frequency
0
94,166
188,332
282,498
376,664
Items

Word Cloud


organis
maintain
strong
environ
employ
account
use
knowledg
understand
market
product
train
process
post
applic
support
opportun
key
benefit
ideal
need
oper
base
softwar
inform
requir
experi
appli
ensur
job
high
abl
offer
must
this
manag
one
provid
respons
the
can
year
pleas
will
lead
within
plan
career
team
staff
sale
busi
industri
also
join
look
care
project
area
work
role
part
day
includ
you
new
sal
ari
senior
contact
custom
level
abil
person
excel
develop
client
candid
origin
posit
well
recruit
success
servic
system
skill
full
qualiti
time
good
compani
design
report
experienc
current
consult
communic
engineer
technic
deliv
profession