University of Pisa

Business and Project Management

Screening and selection of candidates

Student

Davide Vigna

Academic Year: 2022-2023

Contents

1 Introduction

The candidate screening and selection process is a systematic approach used by HR to identify the most qualified people for a specific job position. This process typically involves several stages and techniques for assessing candidates' skills, qualifications, experience, and suitability for the job and it varies across organizations and industries, depending on specific requirements and preferences. The entire procedure can be summarized in four main steps.

1. **Minimum requirements check** : it consists of an initial screening by reviewing resumes or CVs to shortlist candidates who meet the basic requirements of the job;

2. **Additional skills analisis**: it is about digging into the not mandatory qualities of the selected candidates, in order to find out what one candidate could bring to the role compared to another;

3. **Candidate interview**: this represents an opportunity for both the recruiter and the candidate to interact, exchange information and do some specific tests; this phase can be done in different ways and at different times;

4. **Decision Making**: based on the gathered information from the previous steps, the hiring team evaluates each candidate's qualifications, interview performance, assessment results and reference checks. At the end, it takes the final decision on which candidate/s to select for the job.

1.1 Challenges and Solutions

Nowadays, thanks also to the rapid evolution of technology, a job advertisement can attract a large number of candidates and it can be time-consuming and challenging to thoroughly review and evaluate each application. Hiring processes often have strict timelines due to business needs and organizational priorities. The challenge lies in conducting thorough assessments, interviews, and reference checks within these constraints, which could lead potentially to rushed decisions or overlooking potentially qualified candidates.

The use of artificial intelligence can bring a real benefit improving the quality of recruitments and reducing screening times, automatizing parts of the previous steps and removing biaes. In this way, the HR can spend their time on all the other "human aspects" such as taking more attention to personal details during interviews, doing more in-depth tests and of course the most important, spending more time for the final decision.

1.2 Project Goal

The aim of this project is to develop a solution that speeds up the screening phases of candidates and offers to HR a tool to choose between potentially candidates in order to organize different hiring strategies, according to the characteristics that emerge from the resumes. Two interesting and common business questions are examined using clustering techniques.

1. **Salary Segmentation**: identification of groups of people that are similar in terms of years of working experience and education level, to do an adequate salary proposal and understaning what kind of profile the business needs.

2. **Candidates Skills Profilation**: identification of distinct skill profiles among job candidates analizing known languages and additional technical skills in order to distinguish highly versatile candidates from candidates with specific domain expertise.

In this project a job posting is selected from a dataset for a computer engineer profile and all the subsequent operations about screening and selection of possible candidates are carried out on this specific announce. The full description is the following:

> Bachelor's degree in Computer Science or related field (or equivalent degree and experience). At least 3 years of experience in Software Development and Development Stack: Django, Python would be nice. Experience with design patterns and code architecture. A passion for performance and scalability. Strong programming abilities in Python. We write Python on top of Django. Know how to write and maintain unit and system tests in Python. Knowledge of databases internals. We use PostgreSQL. Experience with Memcached, Redis, MongoDB, SOLR, Riak, Celery or Rabbitmq is a plus.

The most important words in term of frequency can be easily visualized using the Word Cloud technique. The project's source code can be found at link :
`https://github.com/AlterVigna/Screening-Selection-of-Candidates`

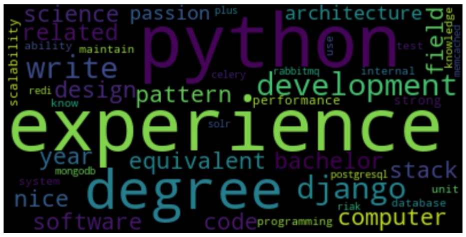


Figure 1: Word Cloud for selected job post

2 Process

2.1 Dataset

The two dataset used in this project can be found at the following links:

1. `https://nubela.co/blog/sample-data-for-linkdb`
2. `https://www.kaggle.com/code/madz2000/text-classification`

The first one contains 10 000 USA Linkedin profiles stored in JSON format. This represent the overall set of candidates that answers to the open job position. The second one is the dataset where the previous mentioned job posting is selected.

2.2 Preprocessing

The data pre-processing phase is the most challenging and time-consuming part of data science, but it's also one of the most important. If a dataset is not prepared correctly, it could compromise all the further analysis and considerations.

2.2.1 Data Extraction and features construction

The raw dataset contains lots of information, some of them are not useful for the purpose of this project, others do. A Python routine is coded to extract interesting information and to build features on the available data. The elaborated dataset contains the following attributes:

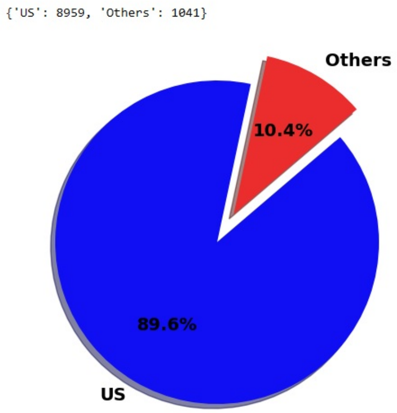
- full name description
- country
- languages spoken
- currently working
- numbers of different working experiences
- years of working experiences
- university
- number of degrees
- number of accomplishment projects
- number of accomplishment courses
- number of accomplishment patents
- full skills descriptions

2.2.2 Data Cleaning

Unfortunately not all the data contains correct values. For example, in "language spoken" some of the languages do not corresponding to a real human language or they present a correct real language but in a different translation. So a proper adaptation is needed to work in a standard format. The number of different languages spoken that emerges is quite huge. In order to work with a limited set of languages, a ranking is done based on the frequency and only the top 9 languages are considered. The other minorities are labelled as *Others*.

2.2.3 Data Reduction

Having a look to some skill descriptions, some of them are not in english language and are related to not american people althought the downloaded dataset garateed to be about USA people. This is a problem for the next NLP phase and for determining the document similarity, so all the CVs about non USA country are removed.



```
{'US': 8959, 'Others': 1041}
```

Figure 2: Dataset Reduction

2.2.4 Outlier Analysis

This dataset contains outliers and this analysis is done in order to understand how to treat these data (valid or noise). Looking at years of experience, it seems that there are some people with experiences longer than a human life. Those cases are all removed, instead other situations are considered individually:

- years of working experience > 40
- number of degrees > 10
- numbers of different working experiences > 30

2.2.5 NLP: Natural Language Processing

This is the core part of the project. NLP is used to analyze and understand human language, in this context the full skills descriptions of candidates. It is fundamental that these textual data be as clean as possible in order to achieve a better document similarity result. The operations performed are the following.

1. *Regular expression matches*: initial cleaning on special sequences of characters, symbols, metacharaters and carriage return. Removing of dates/times, percentages, links, emails and emoji.
2. *Tokenization*: segmenting text into words, punctuations marks etc.
3. *Part of Speech (PoS) tagging*: assigning word types (like verb or noun etc.) to tokens, based on its definiton and its context.
4. *Stopword removal*: removing of all the terms that do not contribute to the meaning of a sentence.
5. *Lemmatization*: reducing words to their base or root form. The purpose of lemmatization is to normalize different inflected or derived forms of a word so that they can be treated as a single common form.

The last four points are implemented using the free, open-source library *spaCy*. Furthermore, some specific technical characteristics are derived from the pre-processed text. They will be used later in the last section.

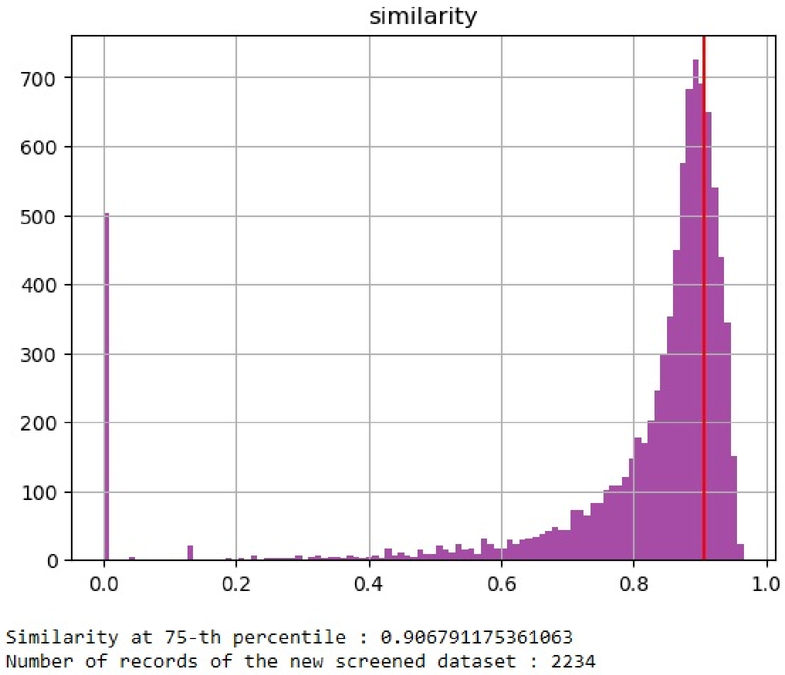
2.3 Screening of candidates

2.3.1 Document similarity

The spaCy library offers also the possibility to compare the degree of similarity of two distinct documents based on semantic similarity. This is determined using word embeddings, multi-dimensional representations of meanings of words. In this project the metric used to compare vectors is the cosine similarity that's the default used by spaCy and can be expressed in the next formula:

$$\text{cosine similarity} = \frac{A \cdot B}{\|A\|\|B\|}$$

where A and B are two distinct vectors of documents. When the documents are similar in meaning the angle between them is small so this ratio is close to 1, otherwise the two vector tends to be orthogonal and the formula returns 0. The *en_core_web_lg* pre-trained English language model is used in order to achieve a better result. It has been trained on a large set of words, respect to standard default models. The similarity is computed between each pre-processed skills description and the selected pre-processed job post. Candidate screening consists of selecting all CVs from the corresponding 75^{th} percentile on the similarity score obtained onwards.



```
Similarity at 75-th percentile : 0.906791175361063
Number of records of the new screened dataset : 2234
```

Figure 3: Histogram about similarty score

2.3.2 Additional "manual" screening

The number of CVs is drastically reduced, however some of those remaining may still not contain mandatory skills. The word vector-based similarity calculation does not explicitly model long-term dependencies or capture the interplay of phrases or sentence structure, which can limit its ability to interpret a specific context. A solution could be training a deep neural network from scratch but it requires large amounts of labeled training data, substantial computational resources and time. An alternative solution could be to leverage the latest generations of Generative-AI techniques, such as OpenAI's GPT-3 or similar models. The use of this service has increased a lot in recent times and tariff plans have been introduced that regulate its use through API calls.

To overcome these limitations, a last screening is done removing all CVs which NOT:

- attended university and to have at least a degree
- at least 3 years of experience
- knowledge of python and programming

The final filtered number of CVs is 483.

2.4 Clustering

Unsupervised learning techniques are used to discover interesting patterns among candidates. Before using any method, a *standard scalarization* is applied on all numerical data in order to work on the same range with all the variables.

2.4.1 Hopkins Statistic Test

A clustering algorithm is a procedure that returns always a result even the points in the dataset do not have a cluster tendency. This method quantifies numerically how the dataset asses to be a cluster. Considering a number of n data points sampled without replacement from the CVs dataset X and generating n other data point from a uniform distribution Y. The formula is:

$$H = \frac{\sum_{i=1}^{n} d_i}{\sum_{i=1}^{n} d_i + \sum_{i=1}^{n} u_i}$$

where n is the number of data points, d_i is the minimum distance between the i^{th} point of Y and its nearest neighbor among the data points of X and u_i is the minimum distance between the i^{th} point of X and its nearest neighbor among the data points of X, different from it self. A good result is when the output is near to 1.

2.4.2 Algorithms and evaluation metric

In this project two different typologies of clustering algorithms are used:

- K-Means: belonging to partitioning methods category; it groups data points into K clusters based on their proximity to the cluster centroids with the aim to minimize the within-cluster sum of squared distances. The *elbow method* is used to determine a proper value of K.
- DBScan: belonging to density-based methods category; it groups data points based on their density and connectivity in the feature space. The choice of *eps* (maximum distance between two points for them to be considered neighbors) and *minPoints* (minimum number of points to consider as cluster) values is done with the help of *knee method* and *grid-search*.

The ground truth is not available so the intrinsic method of Silhouette Index is used to evaluate the goodness of clustering results. It quantifies how well each data point fits within its own cluster compared to other clusters. The formula is:

$$\text{Silhouette Index} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{b_i - a_i}{\max(a_i, b_i)} \right)$$

where N represents the total number of data points, a_i represents the average distance between the i^{th} data point and all other data points within the same cluster, b_i represents the average distance between the i^{th} data point and all other data points in the nearest neighboring cluster. A good result is achieved when the output tends to 1.

3 Results

This part analyses the results obtained concerning the two business questions.

3.1 Salary Segmentation

The two features examined are *years working experience* and *education score*, a feature obtained by combining different education attributes:

$$education_{score} = nr.of degrees + \frac{acc.courses}{100} + \frac{certifications}{500}$$

The clustering tendency is equal to 0.89, so the dataset tend to be a good cluster. The k-means algorithm is applied to this dataset because the density of the points is quite similar and in this situation a partitioning algorithm works better. The elbow method suggests to use 3 as value of K. The silhouette index

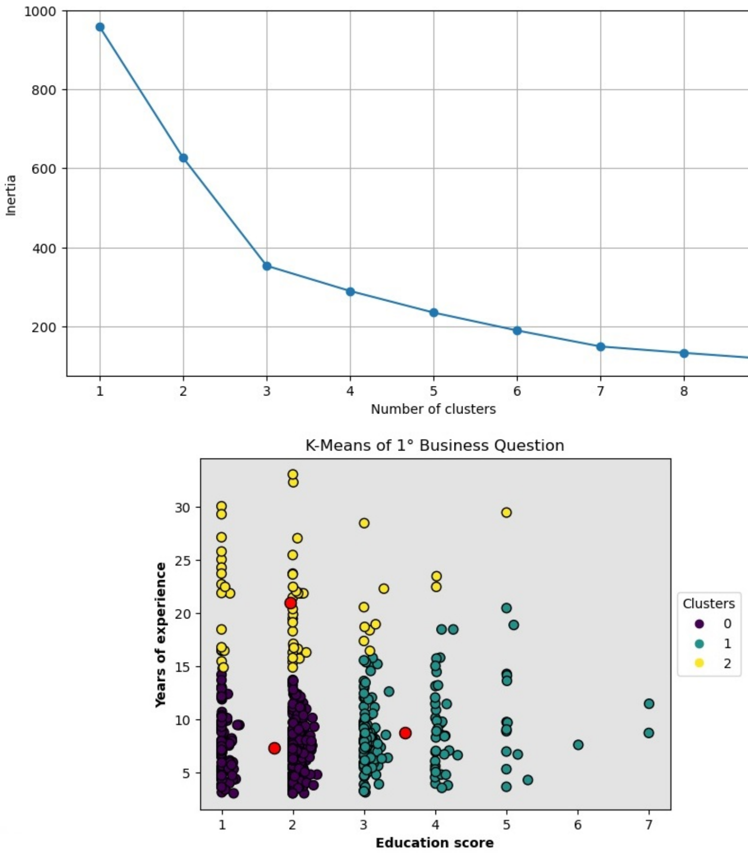


Figure 4: Elbow method and K-Means result

obtained is about 0.47 that is not a very high but suggests that points inside a cluster are relatively well-separated from the points in other clusters, indicating some level of distinctiveness and separation between the clusters. The result obtained can be interpret from a business level point of view labelling the three clusters as in this table.

Nr.Cluster	Nr. of CVs	Name	Year of Exp.	Ed. Score
0	277	Junior	3-15	1-2
1	149	Intermediate	3-20	3-7
2	57	Expert	15+	1-7

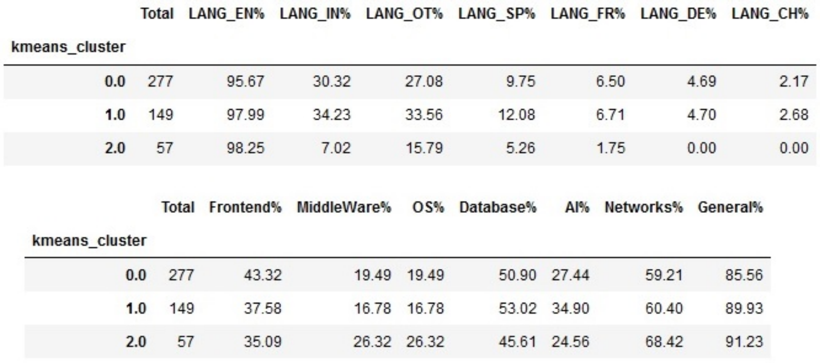
Table 1: Salary segmentation via clustering analysis

- The cluster labelled as *junior* consists of CVs of people who are at a beginner level. They have finished the university obtaining (Banchelor and/or Master Degree) and they have a limited working experience. This is the most populate.

- The cluster labelled as *intermediete* consists of CVs of people who have an important academic experience. They probably have a solid theoretical foundation of the subject but a less field experience even there are some clear outliers.

- The cluster labelled as *expert* consists of CVs of people who have lots of year of working experience. These people have probably mastered the subject matter and they have lots of pratical job experience and could be selected for example to lead a work team. This is the less populate.

Additional statistical information about these clusters are shown in the next images. Note that the percentages are referred to the total number of CVs belowns to a cluster.



kmeans_cluster	Total	LANG_EN%	LANG_IN%	LANG_OT%	LANG_SP%	LANG_FR%	LANG_DE%	LANG_CH%
0.0	277	95.67	30.32	27.08	9.75	6.50	4.69	2.17
1.0	149	97.99	34.23	33.56	12.08	6.71	4.70	2.68
2.0	57	98.25	7.02	15.79	5.26	1.75	0.00	0.00

kmeans_cluster	Total	Frontend%	MiddleWare%	OS%	Database%	AI%	Networks%	General%
0.0	277	43.32	19.49	19.49	50.90	27.44	59.21	85.56
1.0	149	37.58	16.78	16.78	53.02	34.90	60.40	89.93
2.0	57	35.09	26.32	26.32	45.61	24.56	68.42	91.23

Figure 5: Languages and Skills % per cluster

3.2 Candidates Skills Profilation

Some weights are assigned to each languages and skills inversely proportional to their frequency on the screened dataset. The two features examined are *language score* and *skill score*. This analyis is focused on just two languages, *English* and *Spanish* but the others still contribute to the language score.

	EN	IN	OT	ES	FR	DE	CH	RU	JP	IT
Weights	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

Table 2: Skills weights assignment based on occurency

	General	Networks	Database	Frontend	AI	MiddleWare
Weights	0.14	0.29	0.43	0.57	0.71	1.0

Table 3: Languages weights assignment based on occurency

$$language_{score} = nr.of.lang.spoken + \frac{\sum weights.of.lang.spoken}{5.5}$$

$$skills_{score} = nr.of.skills.owned * \sum weights.of.skill.owned$$

The clustering tendency is equal to 0.84. The DBScan is used for this second task because there are several outliers and a different concentration of data points in the dataset. In this case a density based method is able to identify more interesting clusters then a partitioning algorithm. The *eps* and *minPoints*

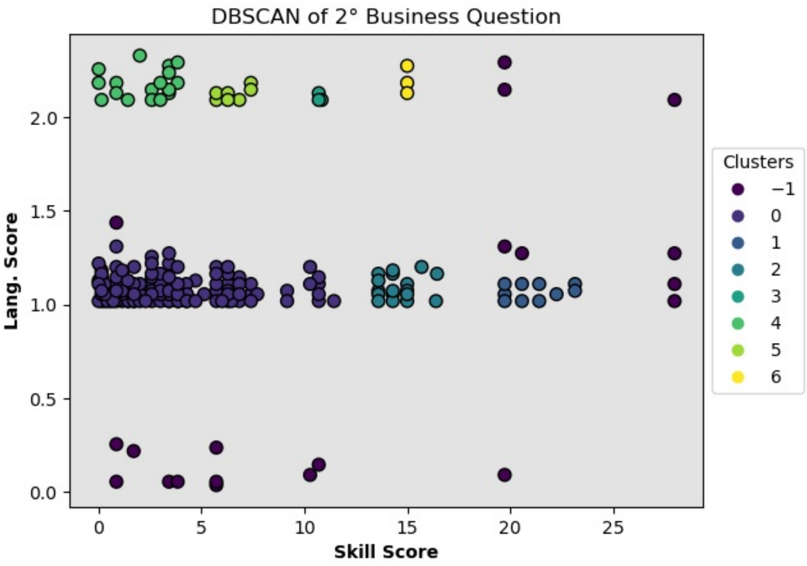


Figure 6: DBScan result

values are chosen empirically with a grid-search and they are equal respectively to 0.29 and 3. The silhouette index obtained is about 0.50. This indicates a moderately good clustering outcome. The DBScan is able to detect all the outliers and treat them as a separate clusters, labelled as -1.

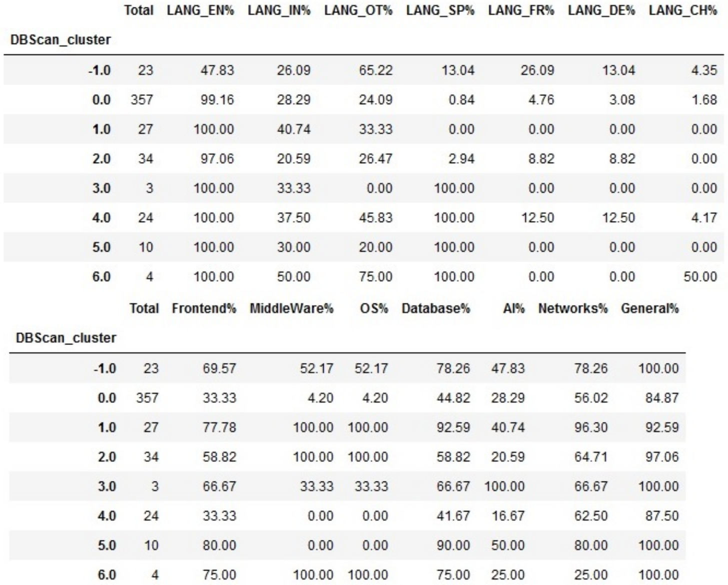
- Cluster 0 can be labelled as *standard domain expert*. It is composed by CVs with few common skills and speaking one of the two languages considered. It is the most populated.

- Clusters 1,2 can be labelled as *standard versatility* (*medium-high*). They are composed by CVs with less common skills, speaking one of the two languages considered. They are relatively limited.

- Clusters 4,5 can be labelled as *pro-domain expert* (*lv.1-2*). They are composed by CVs with common and specific skills, speaking at least the two languages considered.

- Cluster 3,6 can be labelled as *pro versatility* (*medium-high*). They are composed by CVs with rarest skills, speaking at least the two languages considered. There are few CVs with these characteristics.

Additional statistical information about these clusters are shown in the next images. Note that the percentages are referred to the total number of CVs belowns to a cluster.



	Total	LANG_EN%	LANG_IN%	LANG_OT%	LANG_SP%	LANG_FR%	LANG_DE%	LANG_CH%
DBScan_cluster								
-1.0	23	47.83	26.09	65.22	13.04	26.09	13.04	4.35
0.0	357	99.16	28.29	24.09	0.84	4.76	3.08	1.68
1.0	27	100.00	40.74	33.33	0.00	0.00	0.00	0.00
2.0	34	97.06	20.59	26.47	2.94	8.82	8.82	0.00
3.0	3	100.00	33.33	0.00	100.00	0.00	0.00	0.00
4.0	24	100.00	37.50	45.83	100.00	12.50	12.50	4.17
5.0	10	100.00	30.00	20.00	100.00	0.00	0.00	0.00
6.0	4	100.00	50.00	75.00	100.00	0.00	0.00	50.00

	Total	Frontend%	MiddleWare%	OS%	Database%	AI%	Networks%	General%
DBScan_cluster								
-1.0	23	69.57	52.17	52.17	78.26	47.83	78.26	100.00
0.0	357	33.33	4.20	4.20	44.82	28.29	56.02	84.87
1.0	27	77.78	100.00	100.00	92.59	40.74	96.30	92.59
2.0	34	58.82	100.00	100.00	58.82	20.59	64.71	97.06
3.0	3	66.67	33.33	33.33	66.67	100.00	66.67	100.00
4.0	24	33.33	0.00	0.00	41.67	16.67	62.50	87.50
5.0	10	80.00	0.00	0.00	90.00	50.00	80.00	100.00
6.0	4	75.00	100.00	100.00	75.00	25.00	25.00	100.00

Figure 7: Languages and Skills % per cluster