

# **Resume Parser, Career Path Recommender and Salary Forecasting**

Project Work Report

Submitted in partial fulfillment of the requirements for the degree of

**BACHELOR OF TECHNOLOGY**

in

**COMPUTER SCIENCE AND ENGINEERING**

by

**S SAI GIRISH** (16CO244)

**CHETHAN** (171CO113)

**VASUDEV B M** (171CO150)



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA,

SURATHKAL, MANGALORE - 575025

April, 2021



## DECLARATION

*by the B.Tech. Student*

I/We hereby *declare* that the Project Work Report entitled  
**RESUME PARSER, CAREER PATH RECOMMENDER AND SALARY**


**FORECASTING**


which is being submitted to the **National Institute of Technology  
Karnataka, Surathkal** for the award of the Degree of **Bachelor of  
Technology** in **Computer Science and Engineering**

is a *bonafide report of the work carried out by me/us*. The material  
contained in this Project Work Report has not been submitted to any  
University or Institution for the award of any degree.

*Register Number, Name & Signature of the Student(s):*

(1) 16CO244 S Sai Girish 

(2) 17CO113 Chethan 

(3) 17CO150 Vasudev B M 

(4)

Department of **Computer Science and Engineering**

Place: Surathkal

Date: 26th April, 2021



## C E R T I F I C A T E

This is to *certify* that the B.Tech. Project Work Report entitled  
**RESUME PARSER, CAREER PATH RECOMMENDER AND SALARY**

**FORECASTING**

submitted by :

*Sl.No. Register Number & Name of Student(s)*

- |     |          |              |
|-----|----------|--------------|
| (1) | 16CO244  | S Sai Girish |
| (2) | 171CO113 | Chethan      |
| (3) | 171CO150 | Vasudev B M  |
| (4) |          |              |

as the record of the work carried out by him/her/them, is *accepted*  
*as the B.Tech. Project Work Report submission* in partial fulfillment of  
the requirements for the award of degree of **Bachelor of Technology**  
in *Computer Science and Engineering*

Guide(s)  
(Name and  
Signature with Date)

Chairman - DUGC  
(Signature with Date and Seal)



# Acknowledgment

The successful completion of any task would be incomplete without a proper suggestion, guidance and environment. Combination of these three factors act like a backbone to our project "**Resume Parser, Career Path Recommender and Salary Forecasting.**"

We would like to express our special thanks of gratitude to our project mentor Ms.Vani M, Associate Professor, Department of CSE for her able guidance and valuable advice to work on this project.

We express our sincere thanks to Dr. Shashidhar G Koolagudi, Associate Professor and Head of the CSE Department for his support to carry out the project successfully.

We would like to thank friends and respondents for their support.

Not least of all, we extend our sincere gratitude to our college National Institute of Technology Karnataka for giving us this opportunity.

Place: Surathkal

Date: 26 April, 2021.

S Sai Girish

Chethan

Vasudev B M





## Abstract

A Resume parser is a context-based software which is able to recognize the relevant information from the document and convert it to an organized structure. It does so by constructing a parse tree or an abstract input tree by taking inputs in the form of a sequence of instructions. It automatically segregates the information into various fields and parameters like contact information, educational qualification, work experience, skills, achievements, professional certifications, etc. Also, to gain more attention from the recruiters, most resumes are written in diverse formats, including varying font size, font color, and table cells. However, the diversity of format is harmful to data mining, such as resume information extraction, automatic job matching, and candidates ranking. For the Recommender part, at present, there are many job posting websites providing a huge amount of information and students need to spend hours to find jobs that match their interests. Using text mining and collaborative filtering techniques, personalized job recommendations, additional skills and learning resources required for a related job can be provided. On top of the recommendation, we also provide a salary forecast based on the current job and the career path that we will be recommending.

In this project, we divide the topic into three sub-problems, **a) Resume Parser, b) Career Path Recommender and c) Salary Forecast**. We have parsed and tested the code on a hundred resumes (including our own), all based on different profiles and industries supported. The recommendation is done using predictive Machine Learning techniques. The final salary forecast is obtained by applying regression models on vectors generated using NLP techniques on job profile skills.

**Keywords:** Resume Parser, Recommendation system, Machine Learning, Forecasting



# Contents

Chapter	Section	Subsection	Topic	Page No.
			<b>List of Figures</b>	<b>vii</b>
			<b>List of Tables</b>	<b>ix</b>
			<b>Acronyms</b>	<b>xi</b>
<b>1</b>			<b>Introduction</b>	<b>1</b>
	1.1		Motivation	<b>2</b>
	1.2		Outline of the report	<b>3</b>
<b>2</b>			<b>Literature Survey</b>	<b>4</b>
	2.1		Problem Statement	<b>6</b>
	2.2		Objectives	<b>6</b>
<b>3</b>			<b>Methodology</b>	<b>7</b>
	3.1		Resume Parser	<b>8</b>
		3.1.1	Data Extraction	<b>9</b>
		3.1.2	Data Filtering	<b>9</b>
	3.2		Career Path Recommender	<b>10</b>
		3.2.1	Data Mining	<b>10</b>
		3.2.2	Data Filtering	<b>10</b>
		3.2.3	Clustering	<b>11</b>
		3.2.4	Recommendation System	<b>16</b>
	3.3		Salary Forecasting	<b>18</b>
		3.3.1	NLP applied to skills	<b>18</b>
		3.3.2	WordCloud	<b>19</b>
		3.3.3	Binary Encoding	<b>20</b>
		3.3.4	Regression Models	<b>20</b>
	3.4		Hardware and Software Requirements	<b>23</b>



	3.5		Dataset Details	23
4			Results	24
	4.1		Resume Parser	24
	4.2		Career Path Recommender and Salary Forecasting	31
	4.3		Final Results	33
5			Conclusion	37
6			Future Work	38
			Bibliography	39



# List of Figures

3.1	Overall Architecture of the Project	7
3.2	Dataflow diagram for Resume Parser	8
3.3	Comparison of different clustering algorithms	16
3.4	Cosine Similarity	17
3.5	Wordcloud of Popular Skills	19
4.1	Sample Resume 1	27
4.2	Sample Resume 2	29
4.3	Cluster vs. No. of Job Profiles	31
4.4	Top 10 most popular skills	31
4.5	Highest paid job in each cluster	32
4.6	Average salary vs Industry	32
4.7	Average salary vs cluster	33
4.8	Han Pei Ling Resume results	34
4.9	Jai Janyani Resume results	34
4.10	Tsoi Yan (Joyce) Shum Resume results	34
4.11	Chau Gi Feng Sheron Resume results	35
4.12	Wong Chak Yu John Resume results	35
4.13	Vasudev B M Resume results	35
4.14	Executive Resume results	36
4.15	Storekeeper Resume results	36





# List of Tables

3.1	Comparison of Regression Models	21
4.1	Popular skills for each industry	23
4.2	Popular skills for each cluster	24
4.3	Sample Resume 1 Parser results	27
4.4	Sample Resume 2 Parser results	29



# Acronyms

**NLP** - Natural Language Processing

**CNN** - Convolutional Neural Network

**Bi-LSTM** - Bidirectional Long ShortTerm Memory

**CRF** - Conditional Random Field

**ML** - Machine Learning

**OPTICS** - Ordering Points To Identify the Clustering Structure

**BIRCH** - Balanced Iterative Reducing and Clustering using Hierarchies

**CF** - Clustering Feature

**TF** - Term Frequency

**IDF** - Inverse Data Frequency

**OLS** - Ordinary least squares

**LARS** - Least Angle Regression

**SOM** - Self Organising Map.

**SQL** - Structured Query Language.

**SVM** - Support Vector Machine.

**SVR** - Support Vector Regression

**DBSCAN** - Density-Based Spatial Clustering of Applications with Noise



# Chapter 1

## Introduction

The main objective of a Resume Parser project is to extract the required information about candidates without having to go through each and every resume manually. This leads to a more time and energy-efficient process owing to automation. There is no specific format or order to write a resume. Basically, resumes are a form of unstructured data. Each resume has its unique style of formatting, has its own data blocks, and has many forms of data formatting. During the placement season, the hiring managers have to go through the tedious task of reading each resume and selecting or shortlisting the right set of candidates with the appropriate skills. A resume parser aims to reduce the effort that goes into this process, making it easy to select the perfect resume.

Another aspect that our project includes is a career path recommender. Often it is a big deal to move up the ranks in an office, and it is more difficult for freshers since they are unaware of the skills required for a particular profile. So, we propose a job recommendation system which takes into account the current set of skills possessed by the user and comes up with an appropriate profile. Based on this profile, three suitable job profiles are recommended.

Also as the third objective, we aim to forecast the salary after a period of approximately five to nine years for each profile, therefore giving options for the user to choose the right profile based on skills to acquire as well as monetary compensation received.

Overall, our project serves as a one-stop destination for an user to get the recommendation of a career path based on current skills and an opportunity to choose among the recommendations based on skills to acquire and monetary compensation.

## 1.1 Motivation

The motivation for doing this project was primarily an interest in undertaking a project in an interesting area of practical application. The opportunity to learn about a mix of different areas of computing was appealing.

Automation using Natural Language Processing (NLP) is the ability for a computer to take unstructured data gathered from different resources and find meaning in it. It helps to accelerate and streamline the tasks that have predominantly been time and labor intensive to be done by hand.

Recommender systems are an important class of machine learning algorithms that offer relevant suggestions to users. Machine learning algorithms in recommender systems are typically classified into two categories – content based and collaborative filtering methods although modern recommenders combine both approaches. Collaborative methods work with the interaction matrix. The task here is to learn a function that predicts the correct job profile to each user. The main challenge we come across is that the matrix is typically huge, very sparse and most of the values are missing.

Forecasting is an important area of machine learning. It is important because there are so many prediction problems that involve a time component. However, while the time component adds additional information, it also makes time series problems more difficult to handle compared to many other prediction tasks. Salary forecast has to take the experience of the employee at a future stage and aptly predict his earnings using past data.

## **1.2 Outline of the report**

Chapter 2 focuses on various studies where Machine Learning has been used in the field of forecast and recommendation, the problem statement and its objectives. Chapter 3 describes the detailed methodology that has been followed during the course of the project that provides the base for Chapter 4 which explains the results obtained . Chapter 5 explains the conclusion that has been drawn from the above results. Chapter 6 describes the future work and the last chapter includes all the sources which we have used in the process. Chapters 3,4,5 include sub-sections according to the objectives.

# Chapter 2

## Literature Survey

Our implementation is based on the topics discussed in the following research papers (Refer bibliography for paper name and link):

[1] focuses on parsing information from the resume using natural language processing, find the keywords, cluster them onto sectors based on their keywords and lastly show the most relevant resume to the employer based on keyword matching.

[2] proposes a system for resume parsing using deep learning models such as the convolutional neural network (CNN), Bi-LSTM (Bidirectional Long ShortTerm Memory) and Conditional Random Field (CRF).

[3] uses the concept of semi-structured characteristics of the resume, information retrieval based on regular expression and text automatic classification can be applied to extract information. The research on the processing of the semi-structured document, it will not only be as a directive of the further research on the resume analysis, but also be as the reference to other forms of the semi-structured document.

[4] proposed implementations of career recommender systems introduce features of security, reliability and transparency in the process of career recommendation using ML. The student's profile can be handled in a more secure way by providing data encryption.

[5] uses text mining and collaborative filtering techniques the system first scans the user's profile and resume, identifies the key skills of the candidate and generates personalized job recommendations.



[6] introduces a novel machine learning model which uses the candidates' job preference over time to incorporate the dynamics associated with a highly volatile job market. Also shown the use of latent competency groups helps in capturing the hidden skill domains for the candidates and the jobs.

[7] uses a data mining technique which is applied to generate a model to predict a salary for individual students who have similar attributes to the training data. Compared Decision trees, Naive Bayes, K-Nearest neighbor, Support vector machines, and Neural networks. Results showed that K-Nearest neighbor provided the best efficiency to be used as a model for salary prediction.

[8] focused on the problem of predicting salary for job advertisements in which salary is not mentioned and also tried to help fresher to predict possible salary for different companies in different locations.

[9] uses Deep Learning techniques to construct a model which predicts the monthly salary of job seekers in Thailand solving a regression problem which is a numerical outcome is effective. Deep Learning techniques are used to automate and formulate a proposed model for salary prediction.

## 2.1 Problem Statement

*To build a “Resume parser and career path recommender along with salary forecasting”.*

## 2.2 Objectives

- a. Obtain resumes belonging to different supported industries from online sources
- b. Parse each resume for relevant information, such as contact details, academic background and skill set.
- c. Scrape the Job Profile data from Payscale website.
- d. Using extracted skills, obtain the current job profile of the user
- e. Recommend the future career path and required skills each of the job profiles in the career path.
- f. Forecast the expected salary after a period of five to nine years for each suggested job profile

# Chapter 3

## Methodology

### Overall Architecture of Project:

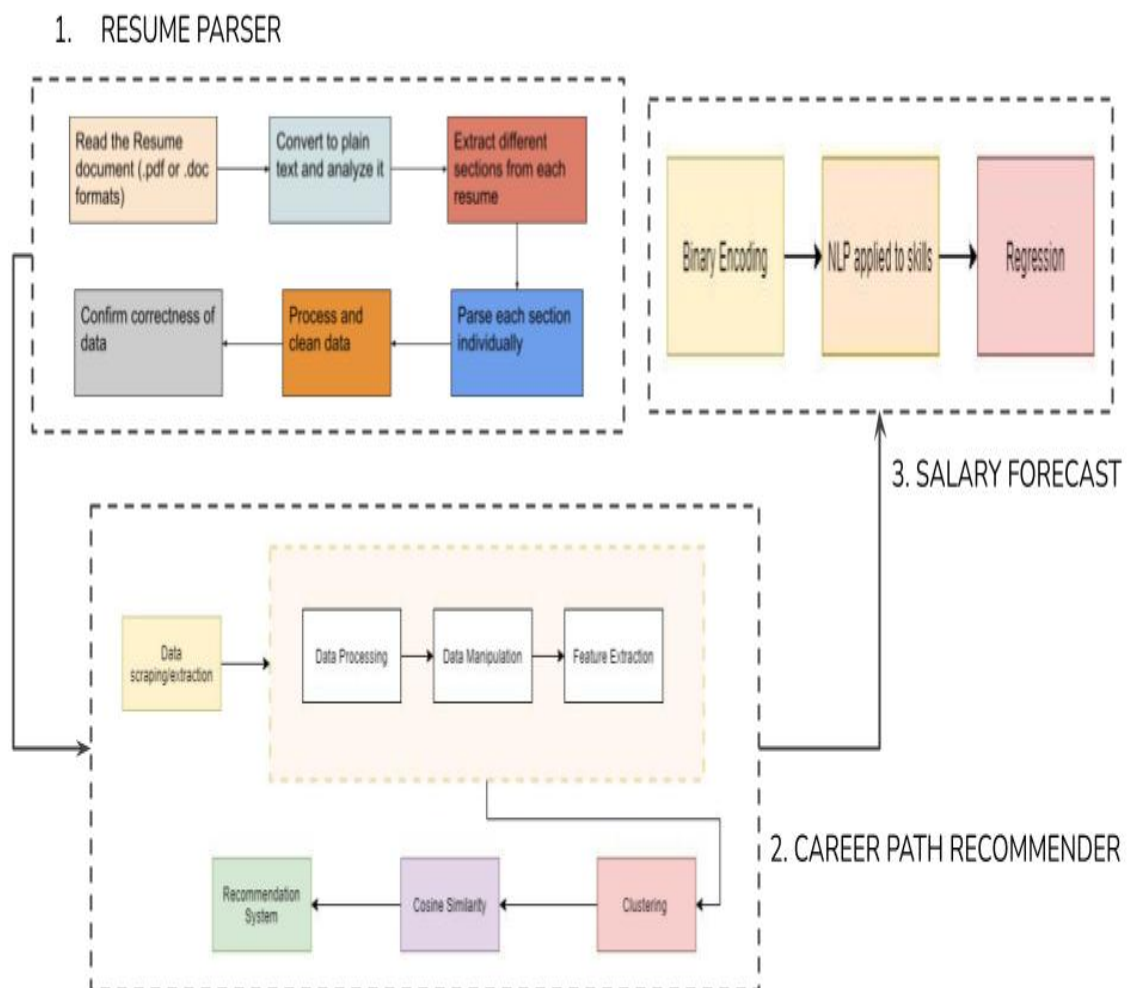


Figure 3.1: Overall Architecture of the Project

The methodology includes 3 sections: Section 3.1 describes the methodology of Resume Parser, Section 3.2 describes the methodology of Job Recommender and Section 3.3 describes the methodology of Salary Forecasting.

## 3.1 Resume Parser

The stages followed to parse a resume document is depicted in the following data flow diagram:

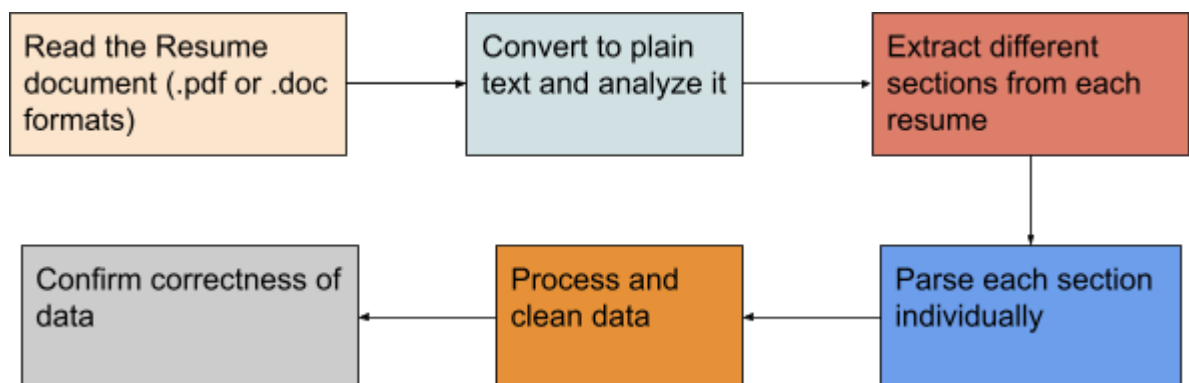


Figure 3.2: Dataflow diagram for Resume Parser

The main process of resume parser, which is extraction and processing of data is explained below:

### 3.1.1 Data extraction

Resumes are commonly presented in PDF or MS word format, And there is no particular structured format to present/create a resume. So, we can say that each individual would have created a different structure while preparing their resumes. For parsing, we have obtained various resumes from different sites in both pdf and word formats. We have used PDFMiner to parse the resume and convert text in pdf format to string format. We are extracting information such as names, contact details including phone number and email, academic/educational details such as degree, year of passing and university and finally the skill set of the user.

### 3.1.2 Data filtering

**Incorrect Data :** As we have extracted data using regex and NLP, we need to ensure correctness since there are always chances of error due to incorrect format being followed. So handling such data gives us accurate results . After parsing, we give an option to the user to check and enter the correct contact details and also to enter skills manually. The contact details of the user will be crucial during the latter parts of the recruiting stage and has to be correct. We are entering skills manually even though we have parsed the skills, because the skills dataset that we are using takes into account all types of skills whereas we are limiting the type of industries to six. The dataset has a very broad range of skills while we have to narrow it down. So we made a subset of the popular skills required for these six industries and numbered them. The user can manually select the appropriate skill by entering the skill's number. We have handled the above problem by manual work , sorting the data and matching the names etc. . .

**Incorrect values :** The data collected after parsing are corrected and is being assumed to be completely right while passing to the recommendation system.

## 3.2 Career Path Recommender

This stage consists of 4 stages, a) Data mining/extraction, b) Data filtering, c) Clustering and d) Recommendation system.

### 3.2.1 Data Mining

As mentioned earlier, we have limited the job profiles to the top hundred jobs across the following six industries:

- 1) Accounting and Finance
- 2) Architecture and Engineering
- 3) Business Operations
- 4) General Managers and Executives
- 5) Information Technology
- 6) Marketing and Advertising

We have extracted the most recent data across the payscale website for these industries. The data consists of the salary data and its associated skills. Initially we mined the data using scrapy but came across an obstacle in the form of lazysection\_wrapper. Some parts of the data were rendered inaccessible when mining resulting in incomplete data. Later we came across Octoparse tool for data mining and have used it for further processes. It outputs the mined data to JSON format. We then combine all the data into a single csv file called merged\_data.

### 3.2.2 Data filtering

**Missing Data :** Not all job profiles have the data pertaining to different fields or attributes that we require. There were many cells with NaN values, and we have filtered such data out from most attributes which resulted in unusable data in further processes.

**Formatting Data:** Another problem at hand was to format all the data the way we require, for instance the job profiles had characters such as ‘\_’ in place of spaces and

we also had to format all the salary related components into integer values from strings. We made these operations into functions since they had to be repeated for each profile we had. It was also necessary since for further comparisons the data had to be of compatible types. Also all the fields were regarded as strings after data mining, so we had to validate the data type of each column and change it accordingly.

### 3.2.3 Clustering

Clustering algorithms are used to analyse data to predict its type of class, we have tested various cluster algorithms to analyze its performance. K-means analysis algorithm is one of the most common methods, but it requires pre-processing steps for mislaid values. Hierarchical method does not cluster all objects in a single step, it takes more time and iterations to cluster objects when the data set is large. The DBSCAN method has distance functions which are cosine and Euclidean distance functions. In a similar way, we have tested ten clustering algorithms.

a) K-means [16] - It is an unsupervised method used for clustering. It is largely used because of its efficiency and simplicity. If compared with hierarchical clustering, k-means clustering has a time complexity of  $O(n)$  and hierarchical clustering has a time complexity of  $O(n^2)$ . Hence, k-means clustering works better for large dataset.

The approach k-means follows to solve the problem is called Expectation Maximization. The E-step is assigning the data points to the closest cluster. The M-step is computing the centroid of each cluster. The objective function is:

$$J = \sum_{i=1}^m \sum_{k=1}^K w_{ik} \|x^i - \mu_k\|^2$$

, where  $w_{ik}=1$  for data point  $x_i$  if it belongs to cluster  $k$ ; otherwise,  $w_{ik}=0$ . Also,  $\mu_k$  is the centroid of  $x_i$ 's cluster. We first minimize  $J$  w.r.t.  $w_{ik}$  and treat  $\mu_k$  fixed. Then we minimize  $J$  w.r.t.  $\mu_k$  and treat  $w_{ik}$  fixed. Technically speaking, we differentiate  $J$  w.r.t.  $w_{ik}$  first and update cluster assignments (E-step). Then we differentiate  $J$  w.r.t.  $\mu_k$  and recompute the centroids after the cluster assignments from the previous step (M-step).

Therefore, E-step is:

$$\frac{\partial J}{\partial w_{ik}} = \sum_{i=1}^m \sum_{k=1}^K w_{ik} \|x^i - \mu_k\|^2 \Rightarrow w_{ik} = \begin{cases} 1, & \text{if } k = \arg \min_j \|x^i - \mu_j\|^2 \\ 0, & \text{otherwise} \end{cases}$$

And M-step is:

$$\frac{\partial J}{\partial \mu_k} = 2 \sum_{i=1}^m w_{ik} (x_i - \mu_k) = 0 \Rightarrow \mu_k = \frac{\sum_{i=1}^m w_{ik} x_i}{\sum_{i=1}^m w_{ik}}$$

b) Affinity propagation [12] - It takes as input measures of similarity between pairs of data points. Real-valued messages are exchanged between data points until a high-quality set of exemplars and corresponding clusters gradually emerges. The main configuration to tune is the damping set between 0.5 and 1, and the preference.

c) DBSCAN [15] - It stands for Density-Based Spatial Clustering of Applications with Noise. It is designed to discover clusters of arbitrary shape, by involving finding high-density areas in the domain and expanding those areas of the feature space around them as clusters.. DBSCAN requires only one input parameter and supports the user in determining an appropriate value for it.

d) Mini batch K-means [17]- Mini-Batch K-Means is a modified version of k-means that makes updates to the cluster centroids using mini-batches of samples rather than the entire dataset, which can make it faster for large datasets, and perhaps more robust to statistical noise.

e) Mean shift [18]- Mean shift clustering involves finding and adapting centroids based on the density of examples in the feature space. For discrete data, the recursive mean shift procedure converges to the nearest stationary point of the underlying density function. It has only one hyperparameter called bandwidth.

f) OPTICS [19]- It stands for Ordering Points To Identify the Clustering Structure, and it is a modified version of DBSCAN clustering. It creates an augmented ordering of the database representing its density-based clustering structure, instead of producing a clustering of a data set explicitly.



g) Spectral Clustering [20]- It is based on the fundamentals of linear algebra. It uses the top eigenvectors of a matrix derived from the distance between points.

h) Gaussian Mixture Model [21]- A Gaussian mixture model summarizes a multivariate probability density function with a mixture of Gaussian probability distributions

i) Agglomerative clustering [13]- It is a part of the broader class of hierarchical clustering methods. It works in a “bottom-up” manner. Agglomerative clustering involves merging examples until the desired number of clusters is achieved. The algorithm starts by treating each object as a singleton cluster. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects. The result is a tree-based representation of the objects, named dendrogram.

We can use a dendrogram to visualize the history of groupings and figure out the optimal number of clusters.

- i) Determine the largest vertical distance that doesn't intersect any of the other clusters
- ii) Draw a horizontal line at both extremities
- iii) The optimal number of clusters is equal to the number of vertical lines going through the horizontal line

In our case, Agglomerative clustering has resulted in almost even distribution of profiles and most job titles within a cluster appear to be similar when manually checked.

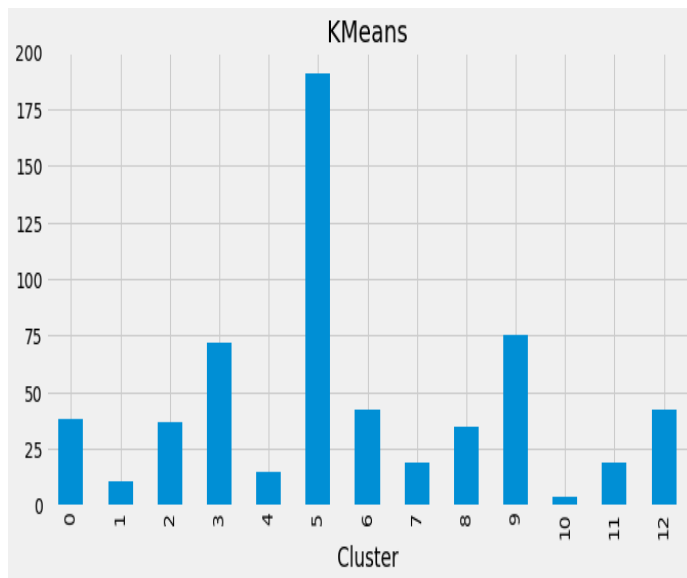
j) BIRCH [14]- It stands for Balanced Iterative Reducing and Clustering using Hierarchies. It incrementally and dynamically clusters the incoming multi-dimensional metric data points to try to produce the best quality clustering. The threshold and n\_clusters hyperparameters need to be tuned in this approach. It does so by first generating a small and compact summary of the large dataset that retains as much information as possible.

It summarizes large datasets into smaller, dense regions called Clustering Feature (CF) entries. Formally, a Clustering Feature entry is defined as an ordered triple,  $(N, LS, SS)$  where 'N' is the number of data points in the cluster, 'LS' is the linear sum of the data points and 'SS' is the squared sum of the data points in the cluster.

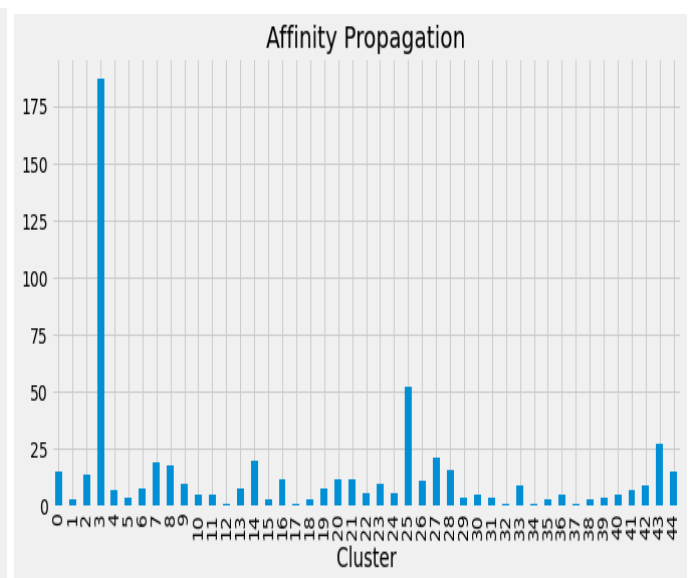
A CF tree is then generated, which is a tree where each leaf node contains a sub-cluster. Every entry in a CF tree contains a pointer to a child node and a CF entry made up of the sum of CF entries in the child nodes. The threshold parameter is the maximum number of data points a sub-cluster in the leaf node of the CF tree can hold.

BIRCH also provides an optimal number of clusters and the job titles in each cluster all appear to be very similar w.r.t the title as well as the skills.

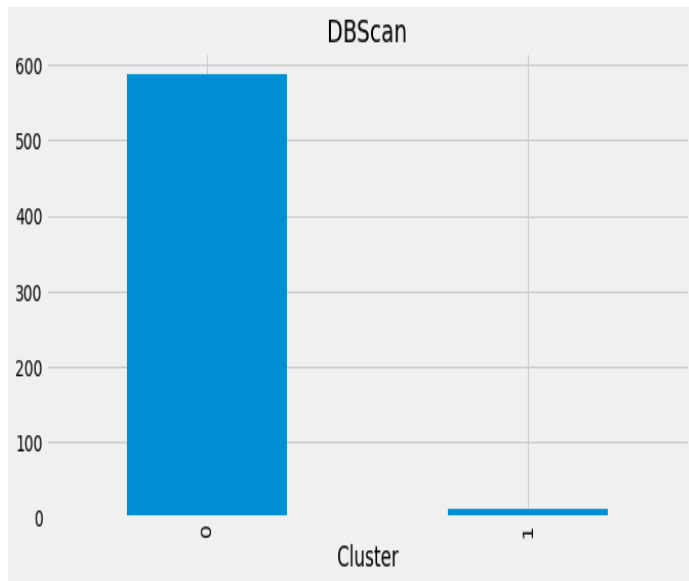
### Comparison of the Clustering Algorithms:



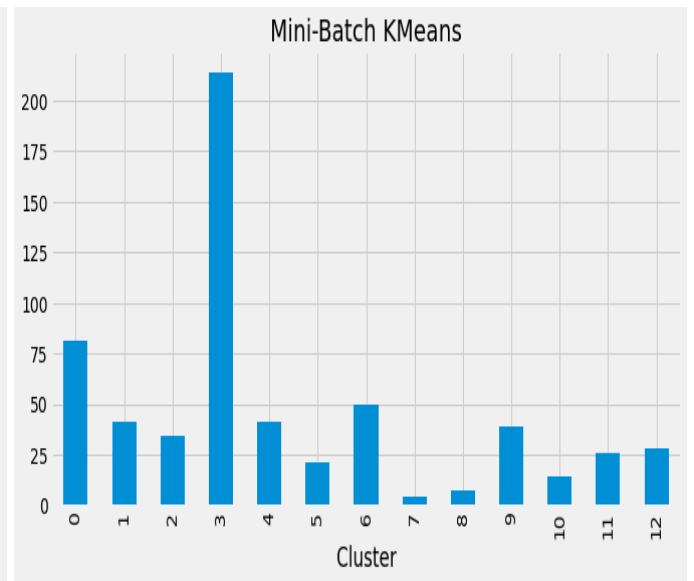
(i)



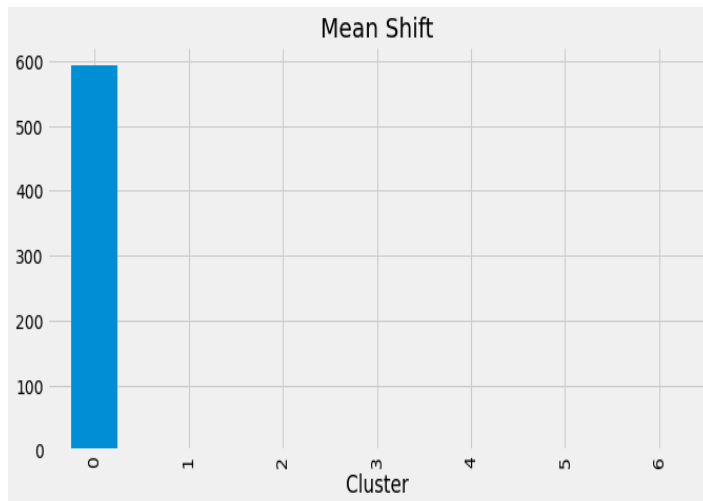
(ii)



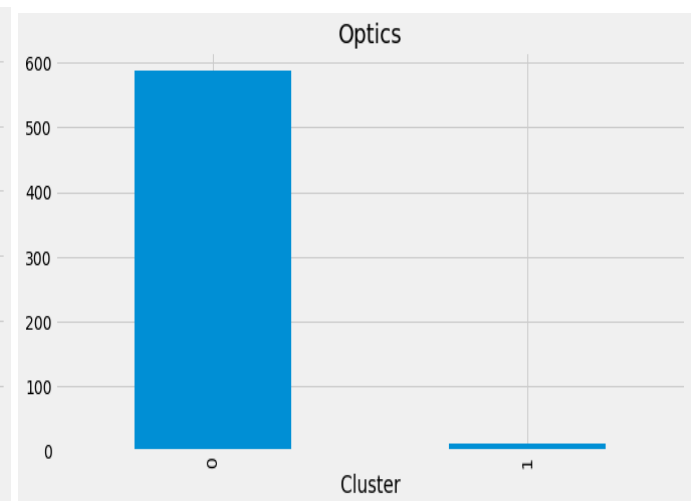
(iii)



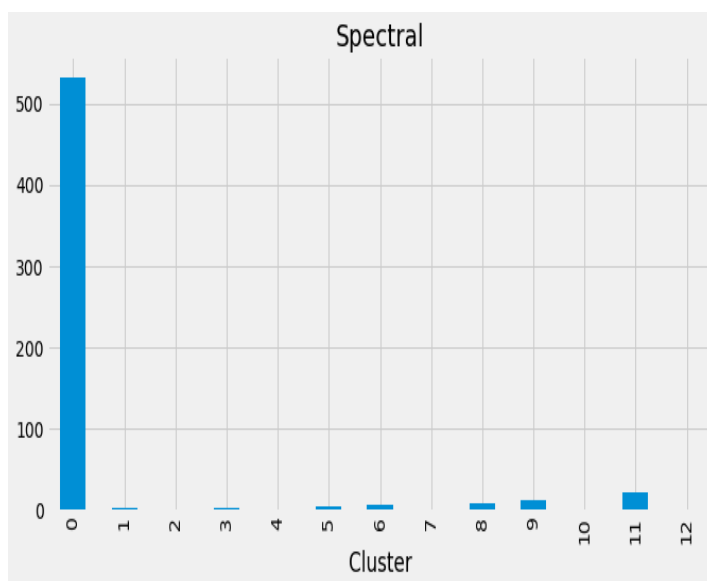
(iv)



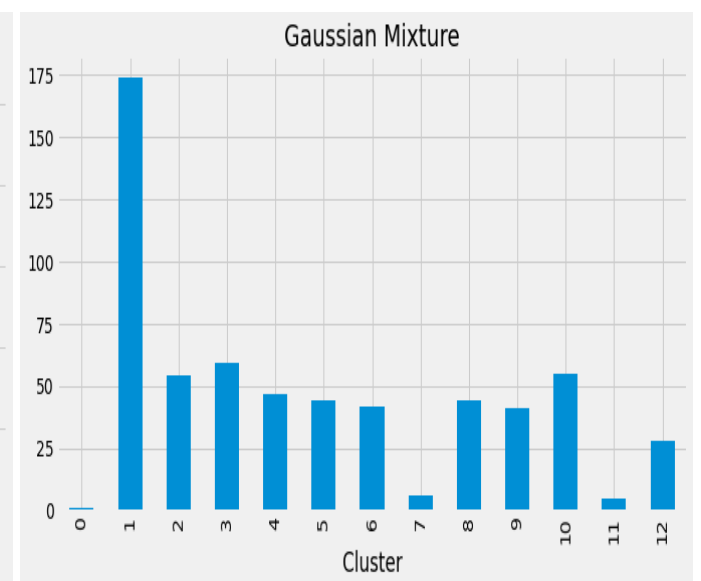
(v)



(vi)



(vii)



(viii)

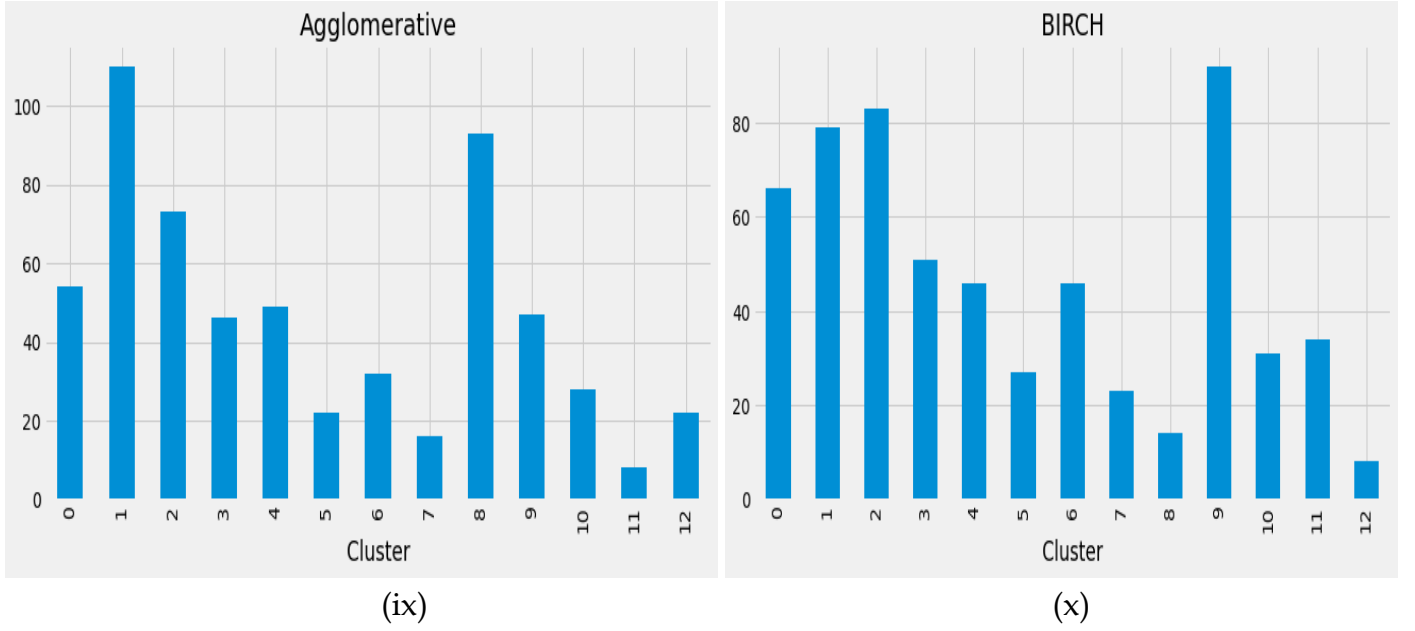


Figure 3.3: Comparison of different clustering algorithms

Considering the clustering of the profiles and by manually cross verifying the job titles in each cluster, we have concluded that BIRCH clustering is the best approach for our data.

### 3.2.4 Recommendation system

For recommending the next best job profile, and also an appropriate career path, we have used cosine similarity function and also a predictive function which works on a particular cluster and chooses the next best available profile.

In Data Mining, similarity measure refers to distance with dimensions representing features of the data object, in a dataset. If this distance is less, there will be a high degree of similarity, but when the distance is large, there will be a low degree of similarity.

- 1) Cosine similarity: Cosine similarity is the cosine of the angle between two n-dimensional vectors in an n-dimensional space. It is the dot product of the two vectors divided by the product of the two vectors' lengths (or magnitudes).

$$\cos(x, y) = \frac{x \cdot y}{\|x\| \times \|y\|}$$

$$\text{similarity}(x, y) = \frac{\sum_{i=1}^n x_i \times y_i}{\sqrt{\sum_{i=1}^n x_i^2} \times \sqrt{\sum_{i=1}^n y_i^2}}$$

$x \cdot y$  = product (dot) of the vectors 'x' and 'y'  
 $\|x\|$  and  $\|y\|$  = length of the two vectors 'x' and 'y'  
 $\|x\| \times \|y\|$  = cross product of the two vectors 'x' and 'y'

We can use the Cosine Similarity algorithm to work out the similarity between two things. We might then use the computed similarity as part of a recommendation query. For example, to get movie recommendations based on the preferences of users who have given similar ratings to other movies that you've seen. We use the similarity obtained to rank the job profiles within a cluster and return the one with highest similarity.

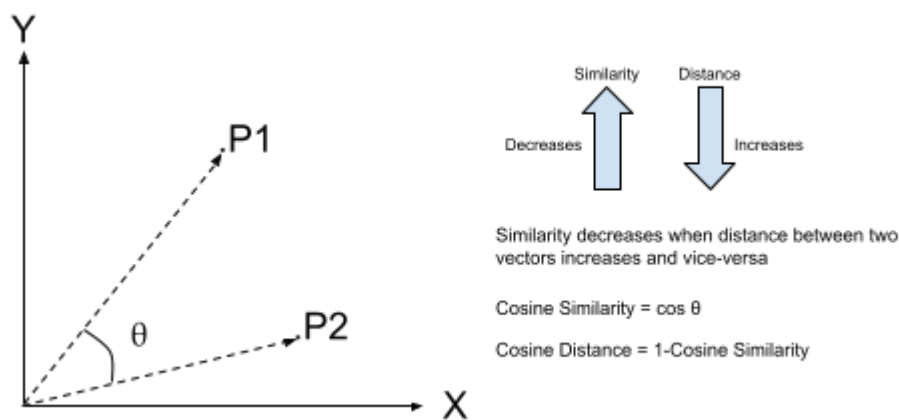


Figure 3.4: Cosine Similarity

The working principle is as follows. Assume there are two points  $P_1$  and  $P_2$  in a two dimensional space, and a vector from origin is drawn to both points. Let the angle that  $P_1 O P_2$  makes be  $\theta$ . If both points are along the same direction, that means  $\theta = 0^\circ$ , so  $\cos \theta = 1$ . Similarly, if the vectors formed by both points are perpendicular to each other then  $\theta = 90^\circ$ , so  $\cos \theta = 0$ . The cosine distance is then found from this cosine angle, cosine distance =  $1 - \cos \theta$ . The distance measured will be minimum when both points lie in the same direction since, cosine distance =  $1 - \cos \theta = 1 - 1 = 0$ .

2) Job recommender function - We sort the cosine similarities obtained in the previous step. Our objective now is to find the best fit profile for the user, with the main constraint being salary. So this function takes into account the similarities of the profile and sorts it according to the average salary for each profile.

After sorting with respect to both the similarity and the salary, we obtain a new profile with a similar skill set and also a higher salary. By including both factors into consideration, we achieve a better recommendation compared to solely using the similarity function.

To build a career path as a recommendation to the user, we suggest three job profiles in order of increasing salary, using the above function. The user can use this as a baseline to build his career, and learn the applicable skills that are required for each profile. We then prompt the required skills for each profile, thereby giving the user options to choose a profile based on what skills the user is interested to learn and his expected salary.

### 3.3 Salary Forecasting

After recommending a career path, we aim to forecast the salary that one can expect after having about five to nine years of experience. We achieve this by forecasting the average salary for each recommended profile after a period of five to nine years of experience, based on the current average salary. This value is then prompted along with the skill set required to be eligible for that particular profile.

#### 3.3.1 NLP applied to skills

In NLP, the words which are useless are called stop words. Python's natural language toolkit library has a list of english stop words. The first step when processing textual data is the removal of such stop words. The next strategy is to score the relative importance of words using TF-IDF.

**Term Frequency (TF)** - It is the number of times a particular word appears in the overall document, divided by the total number of words in the document. Every document has its own term frequency.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$



### 3.3.3 Binary Encoding

It automatically takes categorical variables and dummy codes them whilst reducing the number of output columns equal to the  $\log_2$  of the length of unique values. For binary encoding, each unique label in the categorical vector is associated randomly to a number between zero and  $N-1$ , where  $N$  is the number of unique labels. Now, this number is encoded in base 2 and "transcript" the previous number in 0 and 1 through the newly created columns.

### 3.3.4 Regression Models

We have used NLP techniques described in the previous section over the skillset with average salary as the target variable, we have tried multiple models to find the best prediction model for our data. Here is a list of these models, with its description. We have used the  $R^2$  value and the regression score(`reg_score`) as metrics to judge the accuracy of the models.

a) Linear Regression : LinearRegression fits a linear model with coefficients  $w = (w_1, \dots, w_p)$  to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation. Here, we are using Ordinary least squares Linear Regression. In statistics, ordinary least squares (OLS) is a type of linear least squares method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being observed) in the given dataset and those predicted by the linear function of the independent variable.

b) Support Vector Regression (SVR) - Support Vector Machine(SVM) is commonly used for classification purposes, and SVR uses the same principle as SVM. Instead of finding a hyperplane which separates the different classes, SVR finds the hyperplane on which the majority of the points lie on. The test results were sub par, with most profiles getting a forecasted value ranging between ₹700423 to ₹700461.

c) Ridge Regression - Ridge regression is a model tuning method that is used



to analyse any data that suffers from multicollinearity. This method performs L2 regularization i.e. it adds a penalty equivalent to square of the magnitude of coefficients. Its cost function is:

$$\min(||Y - X(\theta)||^2 + \lambda ||\theta||^2)$$

$\lambda$  here denotes the penalty term, which is the parameter alpha in the ridge function. It follows the assumptions of linear regression except that as ridge regression does not provide confidence limits, the distribution of errors to be normal need not be assumed.

d) Ridge complexity - RidgeCV implements ridge regression with built-in cross-validation of the alpha parameter. This method has the same order of complexity as Ordinary Least Squares.

d)Lasso Regression - Lasso stands for least absolute shrinkage and selection operator. It not only helps in reducing overfitting but it can help us in feature selection, as it uses L1 regularization (adds penalty equivalent to absolute value of the magnitude of coefficients). Lasso regression can lead to zero coefficients because some coefficients are completely neglected to evaluate the output.

e) LassoLarsCV Regression - LARS stands for Least Angle Regression, it provides an alternate and efficient way of fitting a Lasso regularized regression model that does not require any hyperparameters. At each step, it finds the feature most correlated with the target. When there are multiple features having equal correlation, instead of continuing along the same feature, it proceeds in a direction equiangular between the features. It is similar to ridge complexity, in the sense it is lasso regression with built in cross validation of the alpha parameter.

f) LassoLars Regression - LassoLars is a lasso model implemented using the LARS algorithm, and unlike the implementation based on coordinate descent, this yields the exact solution, which is piecewise linear as a function of the norm of its coefficients.

g) Bayesian Ridge Regression - In the Bayesian viewpoint, we formulate

linear regression using probability distributions rather than point estimates. The response,  $y$ , is not estimated as a single value, but is assumed to be drawn from a probability distribution. The aim of Bayesian Linear Regression is not to find the single “best” value of the model parameters, but rather to determine the posterior distribution for the model parameters.

h) Tweedie Regressor - It is most applicable where there is a mixture of zeros and non-negative data points. Basically, if you see a histogram with a spike at zero, it's a possible candidate to be fitted to a Tweedie model.

i) Polynomial Regressor - In statistics, polynomial regression is a form of regression analysis in which the relationship between the independent variable  $x$  and the dependent variable  $y$  is modelled as an  $n$ th degree polynomial in  $x$ . Polynomial regression fits a nonlinear relationship between the value of  $x$  and the corresponding conditional mean of  $y$ . Although polynomial regression fits a nonlinear model to the data, as a statistical estimation problem it is linear.

Regression Type	Regression Score	$R^2$ value
LinearRegression	0.7921219732980164	0.5985742547476479
SVR	0.7921219732980164	-0.37483292130042156
Ridge	0.7410420320857125	0.5152265957710986
Ridge complexity(Ridge CV)	0.7126925698903903	0.4661590645334571
Lasso	0.7921379027390649	0.5990990164719194
LassoLarsCV	0.40277705153277576	0.04895445830414491
LassoLars	0.7226533721394824	0.457687202524912
Bayesian Ridge	6.098976879087559e-11	-0.013280267631390785
Tweedie Regressor	0.851383669509586	-0.013280267631390785
Polynomial Regressor	0.18833333333333332	0.2243280126648045

Table 3.1: Comparison of Regression Models

## 3.4 Hardware and Software Requirements

### Hardware

The machine used to run the above model requires a minimum of 4GB RAM, a multiprocessor with a minimum of 2 cores and a hard drive enough to store the dataset accordingly.

### Software

The above model can be run on Jupyter Notebook or on Google Colaboratory. For this model, we used a machine with Windows 10 operating system and Jupyter Notebook installed on it. The same can also be executed on any Linux platform as well.

### Experiment Details

Programming Language : Python

Platform : Jupyter Notebook, Google Colab

Packages used : matplotlib, pdfminer, pandas, spacy, numpy, nltk, sklearn, wordcloud, BinaryEncoder, scrapy

Tools: OctoParse

## 3.5 Dataset Details

The resumes have been sourced from random sites and are taken as per the industries upon which we are basing our project. Further, a linkedin skill dataset is being used as reference to extract the skills present in the resume. Finally we obtain data such as the skills required for each profile and the salary details pertaining to it by mining the payscale website. We obtained the data for the top hundred jobs across six different industries, totalling to six hundred job profiles.

Resumes dataset - Obtained resumes from the following link-

<https://drive.google.com/file/d/17M9oDPip5JFFFNJhDCBQKy8BMqoyxajU/view>

LinkedIn skills dataset - 36945 skills in single vector form

Mined data (job profiles) - 600 profiles, top 100 profiles from six industries. The job profiles data is split as 500 profiles for train and 100 profiles for testing.

# Chapter 4

## Results

### 4.1 Resume Parser

Popular skills for each industry:

Job Industry	Popular Skills
Accounting and Finance	<ol style="list-style-type: none"><li>1. Microsoft Excel</li><li>2. Financial Analysis</li><li>3. Financial Reporting</li><li>4. Risk Management / Risk Control</li><li>5. Data Analysis</li></ol>
Architecture and Engineering	<ol style="list-style-type: none"><li>1. Engineering Design</li><li>2. Project Management</li><li>3. Autodesk AutoCAD</li><li>4. Microsoft Office</li><li>5. Microsoft Excel</li></ol>
Business Operations	<ol style="list-style-type: none"><li>1. Project Management</li><li>2. Data Analysis</li><li>3. Microsoft Excel</li><li>4. Business Analysis</li><li>5. Operations Management</li></ol>
General Managers and Executive	<ol style="list-style-type: none"><li>1. Operations Management</li><li>2. People Management</li><li>3. Leadership</li><li>4. Project Management</li><li>5. Strategic Planning</li></ol>
Information Technology	<ol style="list-style-type: none"><li>1. Java</li><li>2. SQL</li><li>3. JavaScript</li><li>4. C# Programming Language</li><li>5. Project Management</li></ol>
Marketing and Advertising	<ol style="list-style-type: none"><li>1. Marketing Communications</li><li>2. Strategic Marketing</li><li>3. Marketing Management</li><li>4. Social Media Marketing</li><li>5. Project Management</li></ol>

Table 4.1: Popular skills for each industry

## Popular skills for each Cluster:

Cluster	Popular Skills
0	<ol style="list-style-type: none"> <li>1. Risk Management / Risk Control</li> <li>2. Financial Analysis</li> <li>3. Microsoft Excel</li> <li>4. Data Analysis</li> <li>5. Internal Audit</li> </ol>
1	<ol style="list-style-type: none"> <li>1. Engineering Design</li> <li>2. Project Management</li> <li>3. Microsoft Office</li> <li>4. Microsoft Excel</li> <li>5. Autodesk AutoCAD</li> </ol>
2	<ol style="list-style-type: none"> <li>1. Operations Management</li> <li>2. People Management</li> <li>3. Project Management</li> <li>4. Leadership</li> <li>5. Strategic Planning</li> </ol>
3	<ol style="list-style-type: none"> <li>1. Java</li> <li>2. SQL</li> <li>3. JavaScript</li> <li>4. C# Programming Language</li> <li>5. .NET</li> </ol>
4	<ol style="list-style-type: none"> <li>1. Microsoft Excel</li> <li>2. Financial Reporting</li> <li>3. Financial Analysis</li> <li>4. Accounting</li> <li>5. General Ledger Accounting</li> </ol>
5	<ol style="list-style-type: none"> <li>1. Project Management</li> <li>2. Product Development</li> <li>3. Product Management</li> <li>4. Agile Software Development</li> <li>5. Leadership</li> </ol>
6	<ol style="list-style-type: none"> <li>1. Project Management</li> <li>2. Data Analysis</li> <li>3. Business Analysis</li> <li>4. Microsoft Excel</li> <li>5. Requirements Analysis</li> </ol>
7	<ol style="list-style-type: none"> <li>1. Negotiation</li> <li>2. Vendor Management</li> <li>3. Contract Negotiation</li> <li>4. Procurement</li> <li>5. Sourcing</li> </ol>
8	<ol style="list-style-type: none"> <li>1. Test Automation</li> <li>2. Java</li> <li>3. System Testing</li> <li>4. Automation Scripting</li> <li>5. Selenium Automated Test Tool</li> </ol>

9	<ol style="list-style-type: none"> <li>1. Project Management</li> <li>2. Account Management</li> <li>3. Business Development</li> <li>4. Customer Relationship Management (CRM)</li> <li>5. Social Media Marketing</li> </ol>
10	<ol style="list-style-type: none"> <li>1. Linux</li> <li>2. Troubleshooting</li> <li>3. Microsoft Active Directory</li> <li>4. Systems Troubleshooting</li> <li>5. System Administration</li> </ol>
11	<ol style="list-style-type: none"> <li>1. Marketing Communications</li> <li>2. Strategic Marketing</li> <li>3. Marketing Management</li> <li>4. Branding</li> <li>5. Digital Marketing</li> </ol>
12	<ol style="list-style-type: none"> <li>1. IT Security &amp; Infrastructure</li> <li>2. Security Testing and Auditing</li> <li>3. Cyber Security</li> <li>4. Security Policies and Procedures</li> <li>5. Security Risk Management</li> </ol>

Table 4.2: Popular skills for each cluster

## Result for Abhishak\_resume-

# ABHISHAK VARSHNEY

+91-8433489919

live:abhishakvarshney

abhishakvarshney@gmail.com

in linkedin.com/in/abhishakvarshney

github.com/abhishakvarshney

## EDUCATION

NIT Jaipur(MNIT/MREC)

B.Tech - Metallurgical & Materials Engineering. CGPA: 7.5

2014 - 2018

B.B.S.S.M. Inter College

Intermediate/+2; Uttar-Pradesh Board Result: 91.00%

2012 - 2013

B.B.S.S.M. Inter College

High-School; Uttar-Pradesh Board Result: 72.83%

2010 - 2011

## EXPERIENCE

Software Engineer - Analytics

Skilrock Technologies | Sugul & Damani Group

June 2018 - Present

Gurugram, India

NLP - ChatBot (ARENA)

- Developed an NLP based chatbot in Python from that users can play games, purchase tickets & can chat small talk with user.
- Deployed on Skilrock Technologies Client Gaming and Lottery Engine website, Android and iOS Apps and on Facebook Messenger.
- Technology: Python | Rasa, Microsoft Bot Framework

Trainee | INTERNSHIP

Tata Steel

May 2017 - July 2017

Jamshedpur, India

Heat & Mass Balance in BOF Vessel

- Analyzed and Balanced Heat and Mass data from Raw Material to liquid steel making process i.e. from raw to production.
- Technology: MS - Excel

## CERTIFICATION

- Machine Learning: Coursera
- R Basics - R Programming Language Introduction: Udemy
- Introduction to Python Programming: Udemy
- MongoDB Basics: MongoDB Inc. - MongoDB University
- SQL Fundamental Course: SOLO Learn

## ACHIEVEMENTS

- Placed in top 21% in Housing Price Prediction Kaggle challenge
- Secured II rank in 'International Robotics Challenge - Sixth Sense Robotics - 2014' organized by 'ROBOTech Labs and IIT-Bombay'.
- Participated in '58th National School Skating Championship 2012-13' organized by 'Indian Olympic Association'
- Secured I rank in 'State Skating Championship-2012' organized by 'Vidya-Bharti' at Meerut.

## SKILLS

- Development & Machine Learning: Python
- Analytics & Visualisation Tools: R - Language, MS-Excel, Google Analytics
- Database Languages: MySQL, MongoDB
- Tools & Frameworks: Git, Nginx, Supervisor

## TECHNICAL PROFILES

- Rasa Community | Chatbot : @abhishakskilrock - Rank: 15/2992
- Kaggle | Data Science : @abhishakvarshney - Rank: Top 25%
- HackerRank | Programming : @abhishakvarshney

## PROJECTS

Stanford Open Policing Project- California

- Prediction of traffic stop rates, their times and places that reliably generate stops.
- Technology: Time-Series Analysis | ARIMA Model | R Language

Housing Price Prediction

- Prediction of Sale Price of Houses in USA based on various features.
- Technology: Random Forest | R Language

Twitter Text Mining

- Extract data from twitter and Predict the sentiments of four pharma companies: Bayer, Pfizer, Roche and Novartis.
- Technology: Naive Bayes Theorem | R Language

Image Processing: Object Detection

- A Python based application which can detect different objects. Detected racoon, horses, dogs and cat in various images and videos using trained data/images.
- Technology: CNN, YOLO | Python - TensorFlow, Keras, OpenCV

Linux Path Traversal

- Created virtual linux terminal that can execute various commands: md, cd, cd., ls, pwd, dir, rm, cp, mv, session\_clean.
- Technology: Python

## HOBBIES

Skating | Travelling

Figure 4.1: Sample Resume 1

Feature extracted	Values
Education	'NIT', ('BTech', '2014'), ('MS', '2017')
Email	abhishakvarshney@gmail.com
Mobile Number	8433489919
Skills	1: Image processing 2: Mysql 3: Mongoddb 4: Analytics 5: Machine learning 6: R 7: Python 8: Object detection 9: Android 10: Nginx 11: Opencv 12: Git 13: Excel

Table 4.3: Sample Resume 1 Parser results

## Results for Vasudev Resume

### VASUDEV B M

Roll No.: 171CO150  
V Semester B.Tech  
Department of Computer Science and Engineering  
National Institute of Technology Karnataka  
Phone: +91 9035435225  
Email: [vasubm.171co150@nitk.edu.in](mailto:vasubm.171co150@nitk.edu.in)



#### ACADEMIC PROJECTS

- STUDENTS FOR STUDENTS | ANDROID APPLICATION**  
Built an Android application using which students are able to advertise books they wish to sell and find a prospective buyer. The application serves as a source of contact between the buyer and seller. Currently working on it to expand for greater reach.
- OS SIMULATOR | WEBAPP**  
Built a web application that simulated the major concepts of an OS. Specifically worked on process scheduling and the UI for the same including a wiki page.
- LINUX SIMULATION IN NS-3 DCE**  
The DCE NS-3 module provides facilities to execute existing implementations of userspace and kernelspace network protocols within NS-3. Our task was to demonstrate the Linux example present in it.
- HOME SERVICE PROVIDER | WEBAPP**  
(I)DBMS – Home Service Provider system is an application which can be used by people who face difficulties in searching for people for their household repairs. Instead of going out and searching for them, they can sit at home and get their work done. The system maintains details about all the workers and their masters termed as authorizers and the transactions done by them for reference.  
(II)IS – As a part of Information Science (IS) Course we used the same structure of the existing application and included the relevant concepts of Discretionary Access Control (DAC), Mandatory Access Control (MAC) and RBAC (Role Based Access Control) and some key security features.
- ENABLING THE SUPPORT OF GENTLE RED (GRED) IN DPDK**  
Data Plane Development Kit (DPDK) is a fast packet processing library originally developed by Intel. It bypasses the network stack in Linux kernel. Random Early Detection (RED) is a popular Active Queue Management (AQM) algorithm to keep queuing delays within specified bounds. Gentle RED (GRED) is a minor enhancement to RED. The aim of the project was to implement GRED in DPDK.
- EVALUATING THE IMPACT OF DQ\_THRESHOLD PARAMETER IN PIE ALGORITHM**  
Proportional Integral controller Enhanced (PIE) is a popular AQM algorithm deployed in DOCSIS 3.0 modems of CableLabs, USA. In this project, the main idea was to evaluate the impact of setting DQ\_THRESHOLD on the performance of PIE algorithm.

1

- COLORIZING BLACK AND WHITE IMAGES | WEBAPP**  
In Black and White Images, each pixel has a value (0-255) that corresponds to its brightness. First, we changed the color channels, from RGB to Lab. We used the L as input and train a CNN to predict the values of a and b. Finally we performed color rebalancing to make the image appear more natural. We thus generate our version of a colored image for the provided black and white image.

- MINI C COMPILER**  
This was a project in which we had to undertake a sequence of experiments aimed at the design and implementation of the various phases of a compiler for the C programming language. This subset included a sufficiently rich collection of data types and control structures.

#### ADDITIONAL PROJECTS

- SHAZAM | WEBAPP**  
Created a web-based application using Django which can recognize songs and display the title. The programming was done in python. Used a file as database which had the list of songs. A hash function was used to map the songs to its title. Worked mainly on the front end and also helped in coding phase.
- SOLAR TRACKER | DEVICE**  
Built a solar tracking device which was showcased as part of the Astro Committee in Engineer 2018. It was done using two Light Dependent Resistors and connected to a DC motor. The device was made to rotate towards the light source using an Arduino.
- KWD -FOREST DEPARTMENT |STATUS PROVIDER | WEBSITE**  
As there are actually 32 steps involved in any paper transfer, it is difficult to track down the status. Hence it helps in tracking down the status of the paper/document under verification. Worked on the database design.
- SHOPPING ASSISTANT | ANDROID APPLICATION**  
As part of Lowe's Campus Hackathon, the goal was to build a solution to help the customers find products in the store and help them navigate to the corresponding aisle/shelf. If there is a shopping list, the best shopping trip to complete the purchases was to be provided.

#### AREAS OF INTEREST

- Android Application Development
- Web Development
- Machine Learning
- Computer Networks

2



## SKILLS

- Programming Languages: C, C++, Python, Java
  - Web Technologies: HTML, CSS
  - Frameworks: Django, Android
  - Operating Systems: Windows, Linux
  - DBMS: MySQL, Firebase
- 

## WORK EXPERIENCE

- **Intern at Publicis Sapient – June 2020**  
As part of my internship, the project I was assigned was to build a plugin to JIRA application. Native Gadgets for JIRA Dashboard do not support any time / estimate related fields. The Objective is to build Dashboard Gadgets that are similar to native JIRA gadgets that supports these Time / Estimate fields.
- 

## EDUCATION

Degree	Institution	University/Board	Year of Passing	Percentage/CGPA
10th Grade	Bethesda International School	ICSE	2015	92.80
II PUC (12 <sup>th</sup> Equivalent)	Surana Independent PU College	Karnataka State PU Board	2017	95.66
Under - Graduate	National Institute of Technology Karnataka	NIT	2021 Expected	7.57 As of August, 2020

---

## ACTIVITIES

- Executive Member of Film's Club and part of organizing team in 4th edition
  - Website Coordinator in Amateur's Astronomy Club
  - Astro Committee Member in Engineer 2018
  - Incident Office Team Member in Incident 2019
  - Publicity Team Member for Incident 2019
- 

## ACHIEVEMENTS

- NTSE Scholar from 2015.
- 3<sup>rd</sup> place in State Level Abacus Competition from 2009

Figure 4.2: Sample Resume 2

Feature extracted	Values
Education	'BTech', ('ICSE', '2015'), ('NIT', '2021')
Email	vasubm.171co150@nitk.edu.in
Mobile Number	+919035435225
Skills	1: Web development 2: C++ 3: Machine learning 4: Arduino 5: Web 6: C 7: Css 8: Astronomy 9: Python 10: Web technologies 11: Html 12: Operating systems 13: Access control 14: Android 15: Django 16: Information science 17: Jira 18: Firebase 19: Mysql 20: Java

Table 4.4: Sample Resume 2 Parser results

## 4.2 Career Path Recommendation and Salary Forecasting

### Cluster vs No. of Job Profiles

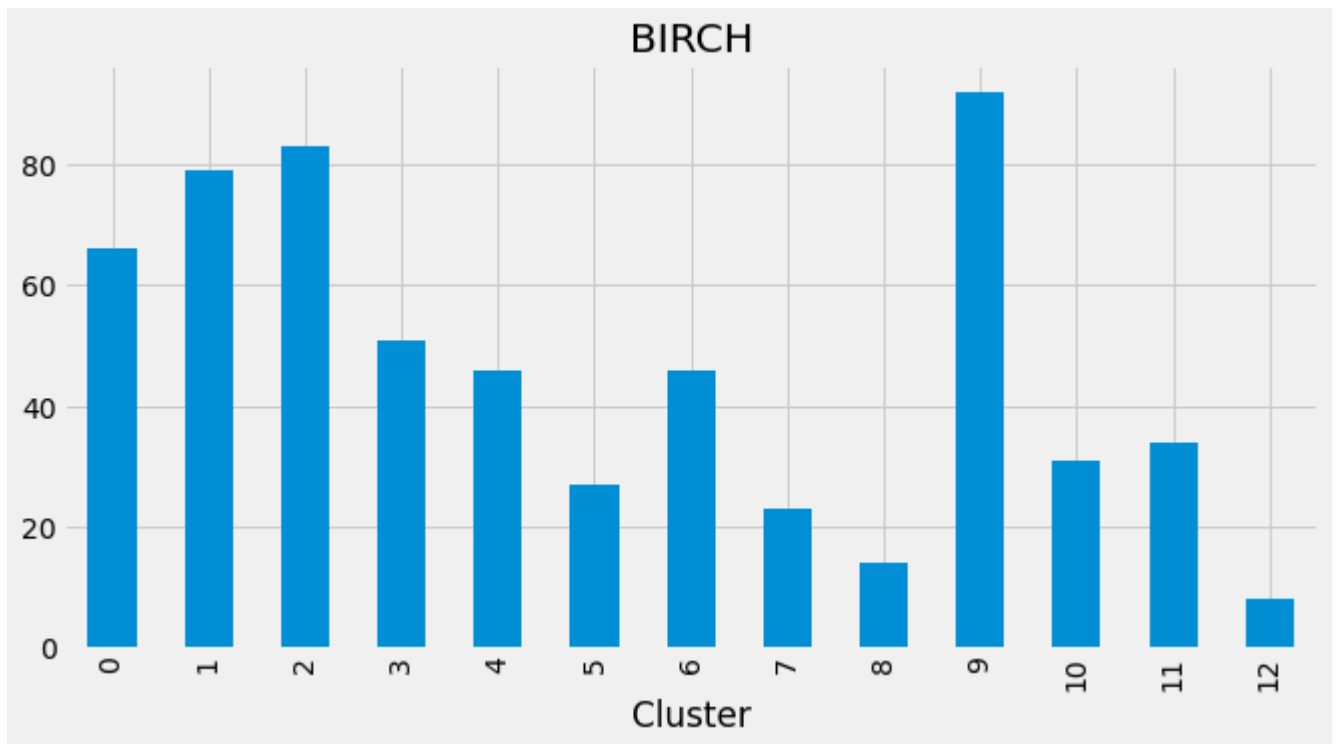


Figure 4.3 : Cluster vs. No. of Job Profiles

### Top 10 most popular skills



Figure 4.4 : Top 10 most popular skills

## Highest paid job in each cluster

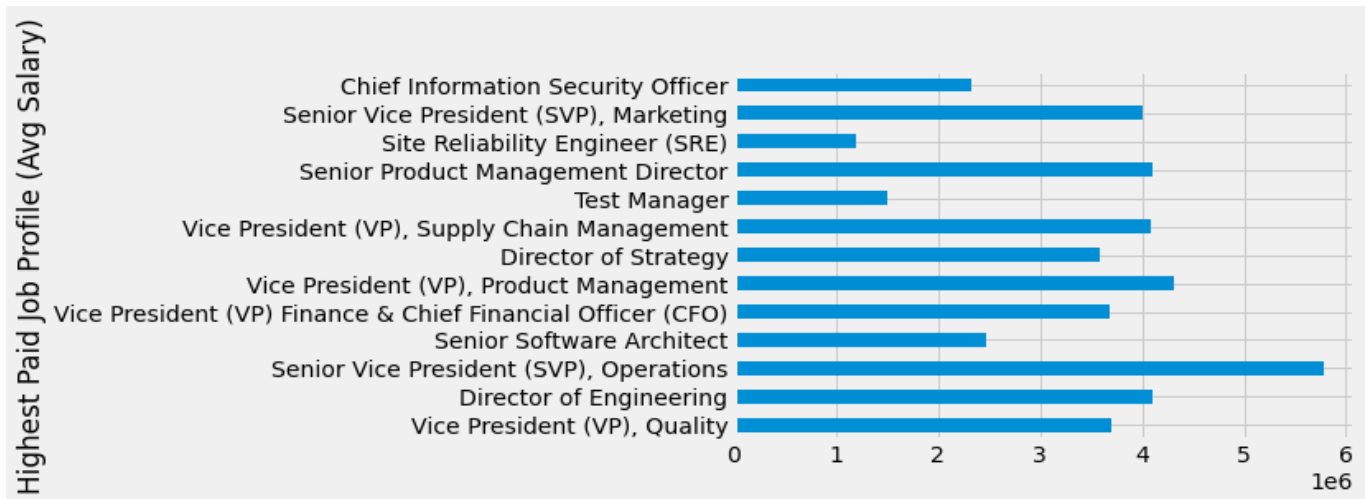


Figure 4.5 : Highest paid job in each cluster

## Average salary vs industry

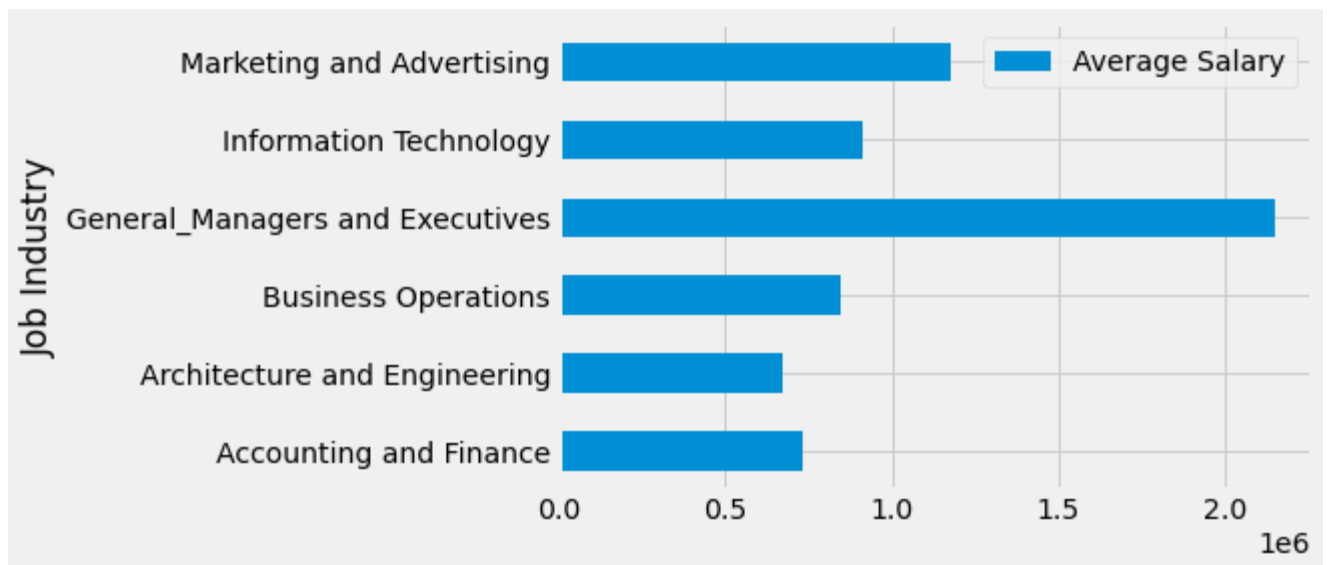


Figure 4.6 : Average salary vs Industry

## Average salary vs cluster

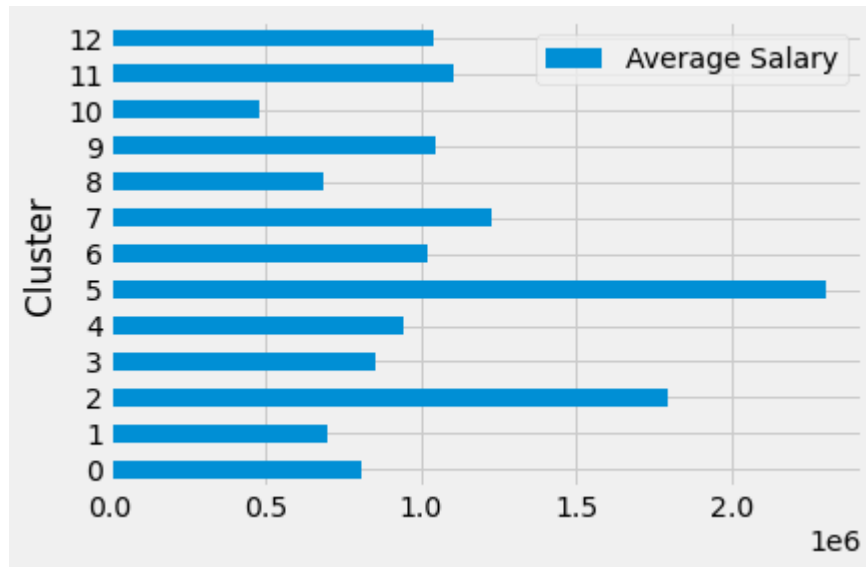


Figure 4.7 : Average salary vs cluster

## 4.3 Final Results

To improve final performance in forecasting further, we've tried different models in our data set.

In [9], the authors were able to achieve an  $R^2$  value of 0.462 when using Random Forest and Gradient Boost Trees. In [22], the results have yielded an adjusted  $R^2$  value in the range of 0.1493 to 0.1496 for the three models they proposed. One fact to note here is that, while [22] have used Adjusted  $R^2$  as the metric, Adjusted  $R^2$  value will always be less than or equal to  $R^2$  value.

Meanwhile our results are far more promising, with an  $R^2$  value of 0.5990 when using the Lasso regression model and 0.5985 for Linear Regression model.

Han Pei Ling Resume:

	Current Job Title	Current Average Salary	Recommended Job 1	Salary Forecast 1	Recommended Skills 1	Recommended Job 2	Salary Forecast 2	Recommended Skills 2	Recommended Job 3	Salary Forecast 3	Recommended Skills 3
0	Senior Auditor	₹222743	Auditing Manager	3.819e+06	[Accounting, Regulatory Compliance, Microsoft Excel, Risk Management / Risk Control, Project Management, Sarbanes-Oxley (SOX) Compliance Audit]	Risk Management Consultant	2.28868e+06	[Risk Consulting, Regulatory Compliance, Risk Management / Risk Control, Project Management]	Risk Management Manager	3.83948e+06	[Regulatory Compliance, Microsoft Office, Risk Management / Risk Control, Project Management]

Figure 4.8 : Han Pei Ling Resume results

Jai Janyani Resume:

	Current Job Title	Current Average Salary	Recommended Job 1	Salary Forecast 1	Recommended Skills 1	Recommended Job 2	Salary Forecast 2	Recommended Skills 2	Recommended Job 3	Salary Forecast 3	Recommended Skills 3
0	PHP Developer	₹274728	Junior Software Engineer	515894	[Amazon Web Services (AWS), Test Automation, Software Test, C# Programming Language, Java, iOS, .NET, SQL, Web Development]	Web Developer	2.74297e+06	[Amazon Web Services (AWS), Java/J2EE, Adobe Illustrator, Hibernate, Web Development, Graphic Design]	Web Designer & Developer	2.57017e+06	[e-Commerce, jQuery, Java, Adobe Photoshop, Angular.js, Bootstrap, Web Design]

Figure 4.9 : Jai Janyani Resume results

Tsoi Yan (Joyce) Shum Resume:

	Current Job Title	Current Average Salary	Recommended Job 1	Salary Forecast 1	Recommended Skills 1	Recommended Job 2	Salary Forecast 2	Recommended Skills 2	Recommended Job 3	Salary Forecast 3	Recommended Skills 3
0	Auditor	₹484550	Underwriter	1.64303e+06	[Mortgage Loans, Loan Underwriting, Risk Management / Risk Control, Insurance, Data Analysis, Customer Service]	Compliance Officer	1.44295e+06	[Anti-Money Laundering (AML), Financial Compliance, Risk Management / Risk Control, Writing Procedures & Documentation, Internal Audit, Legal Compliance, Regulatory Compliance]	Assistant Branch Manager, Banking	1.20741e+06	[Team Leadership, Mortgage Loans, Consumer Loans, Credit Control, Risk Management / Risk Control, Leadership, Sales, Operations Management, Banking, People Management, Customer Service, Loan Processing]

Figure 4.10 : Tsoi Yan (Joyce) Shum Resume results

### Chau Gi Feng Sheron Resume:

Current Job Title	Current Average Salary	Recommended Job 1	Salary Forecast 1	Recommended Skills 1	Recommended Job 2	Salary Forecast 2	Recommended Skills 2	Recommended Job 3	Salary Forecast 3	Recommended Skills 3
0 Tax Consultant	₹511657	Cost Accountant	839733	[SAP Financial Accounting and Controlling (SAP FICO), Cost Accounting, Budgeting, Month-End Close, General Ledger Accounting, Microsoft Word, Financial Analysis]	Financial Analyst, Corporate	750479	[Financial Modeling, Financial Reporting, Data Analysis, Financial Analysis]	Senior Financial Analyst	740278	[Budgeting, Financial Modeling, Research Analysis, Risk Management / Risk Control, Financial Reporting, Python, Data Analysis, SQL, Financial Analysis, Forecasting]

Figure 4.11 : Chau Gi Feng Sheron Resume results

### Wong Chak Yu John Resume:

Current Job Title	Current Average Salary	Recommended Job 1	Salary Forecast 1	Recommended Skills 1	Recommended Job 2	Salary Forecast 2	Recommended Skills 2	Recommended Job 3	Salary Forecast 3	Recommended Skills 3
0 Senior Financial Analyst	₹599646	Financial Consultant	5.27562e+06	[Project Management]	Senior Tax Accountant	1.8269e+06	[Tax Consulting, Tax Preparation, Tax Compliance]	Senior Audit Associate	1.87074e+06	[Generally Accepted Accounting Principles (US GAAP), Accounting]

Figure 4.12 : Wong Chak Yu John Resume results

### Vasudev B M:

Current Job Title	Current Average Salary	Recommended Job 1	Salary Forecast 1	Recommended Skills 1	Recommended Job 2	Salary Forecast 2	Recommended Skills 2	Recommended Job 3	Salary Forecast 3	Recommended Skills 3
0 Web Designer & Developer	₹349698	Android Software Developer	1.89624e+06	[Software Development, iOS, JavaScript, Object Oriented Programming (OOP), Mobile applications development, SQLite, Android Operating System Development]	Systems Engineer, IT	851881	[System Administration, Perl, Computer Hardware Technician, SQL, Windows Server 2008 R2, Cloud Computing, Linux, Windows Operating System General Use]	.NET Software Developer / Programmer	1.56131e+06	[C# Programming Language, .NET, Bootstrap, ASP.NET Framework, RESTful Web Services, SQL Server Integration Services (SSIS), Model-view-controller (MVC), Microsoft SQL Server, Angular.js]

Figure 4.13 :Vasudev B M Resume results

## Executive Resume:

	job_title	Recommended Job 1	Recommended Job 2	Recommended Job 3	salary_forecast_basic 1	salary_forecast_basic 2	salary_forecast_basic 3	avg_salary
359	Vice President (VP), Product Management	Vice President (VP), Product Management	NaN	NaN	3400000	NaN	NaN	4488869

Figure 4.14 : Executive Resume results

## StoreKeeper Resume:

	Current Job Title	Current Average Salary	Recommended Job 1	Salary Forecast 1	Recommended Skills 1	Recommended Job 2	Salary Forecast 2	Recommended Skills 2	Recommended Job 3	Salary Forecast 3	Recommended Skills 3
0	Storekeeper	₹283944	Client Service Executive	503731	[Microsoft Excel, Client Interaction, Customer Service, Customer Relationship Management (CRM), Account Management, Project Management, Oral / Verbal Communication]	Lead Generation Manager	355092	[Lead Generation]	Business Development Officer	451391	[New Business Development, Sales, Microsoft Office, Microsoft Excel, Business to Business (B2B) Sales, Business Development, Sales Management]

Figure 4.15 : Storekeeper Resume results



# Chapter 5

## Conclusions

Presenting a novel career path framework for personalized job and skills recommendations, which focuses on students and young professionals. The proposed model consists of three sections namely Resume parser, Job Recommendation and Salary forecasting. The resumes will be written in many file formats and may differ in their fonts or structure. Hence resume parsing is an intricate task in automatic job recruitment tools. The proposed system extracts the required data from the resumes by matching the regular expressions for contact and academic details and using NLP techniques for extracting name and skills. After extracting the data using regex and NLP, to ensure the correctness of the data an option is provided to the user to check and update their contact details and skills as it is important for the later part of the process. Next, the system exploits an integrated skills knowledge base for carrying out the recommendation task. This consists of four stages namely Data mining, Data filtering, BIRCH Clustering and Recommendation system. Combination of these four stages recommends an appropriate career path. BIRCH clustering is used because of its efficiency and simplicity. BIRCH clustering scales well in terms of running time and quality as the size of the dataset increases. We have used cosine similarity function and also a predictive function which works on a particular cluster and chooses the three next best available profiles and recommends it as a career path. We aim to forecast the salary that one can expect after having about five to nine years of experience. The results are promising.

# Chapter 6

## Future Work

Future work suggested was to take up an enlarged dataset and improve the performance of the proposed system. Also, the work can be extended by including industry name and getting the skills related to that particular industry. Can try to integrate parser and recommender more precisely, i.e. by extracting the skills and directly passing it to the recommender. To do so, we will need an appropriate dataset of skills grouped by the industry so that we can filter out the skills which occur in that particular industry.

Another idea can be to include a job description, and rank the resumes in the order in which it best fits the job description.

## Bibliography

- [1] Satyaki Sanyal, Souvik Hazra, Neelanjan Ghosh and Soumyashree Adhikary, "Resume Parser with Natural Language Processing", IJESC, February 2017.
- [2] C. H. Ayishathahira, C. Sreejith and C. Raseek, "Combination of Neural Networks and Conditional Random Fields for Efficient Resume Parsing," *2018 International CET Conference on Control, Communication, and Computing (IC4)*, 2018, pp. 388-393, doi: 10.1109/CETIC4.2018.8530883.
- [3] Z. Chuang, W. Ming, L. C. Guang, X. Bo and L. Zhi-qing, "Resume Parser: Semi-structured Chinese Document Analysis," *2009 WRI World Congress on Computer Science and Information Engineering*, 2009, pp. 12-16, doi: 10.1109/CSIE.2009.562.
- [4] T. V. Yadalam, V. M. Gowda, V. S. Kumar, D. Girish and N. M., "Career Recommendation Systems using Content based Filtering," *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, 2020, pp. 660-665, doi: 10.1109/ICCES48766.2020.9137992.
- [5] B. Patel, V. Kakuste and M. Eirinaki, "CaPaR: A Career Path Recommendation Framework," *2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService)*, 2017, pp. 23-30, doi: 10.1109/BigDataService.2017.31.
- [6] A. Nigam, A. Roy, H. Singh and H. Waila, "Job Recommendation through Progression of Job Selection," *2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems (CCIS)*, 2019, pp. 212-216, doi: 10.1109/CCIS48116.2019.9073723.
- [7] P. Khongchai and P. Songmuang, "Implement of salary prediction system to improve student motivation using data mining technique," *2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS)*, 2016, pp. 1-6, doi: 10.1109/KICSS.2016.7951419.
- [8] S. Dutta, A. Halder and K. Dasgupta, "Design of a novel Prediction Engine for predicting suitable salary for a job," *2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, 2018, pp. 275-279, doi: 10.1109/ICRCICN.2018.8718711.
- [9] P. Viroonluecha and T. Kaewkiriya, "Salary Predictor System for Thailand Labour Workforce using Deep Learning," *2018 18th International Symposium on Communications and Information Technologies (ISCIT)*, 2018, pp. 473-478, doi: 10.1109/ISCIT.2018.8587998.
- [10] Y. -C. Chou and H. -Y. Yu, "Based on the application of AI technology in resume analysis and job recommendation," *2020 IEEE International Conference on Computational Electromagnetics (ICCEM)*, 2020, pp. 291-296, doi: 10.1109/ICCEM47450.2020.9219491.

- [11] A. Nigam, A. Roy, H. Singh and H. Waila, "Job Recommendation through Progression of Job Selection," *2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems (CCIS)*, 2019, pp. 212-216, doi: 10.1109/CCIS48116.2019.907372
- [12] Frey, B. J. & Dueck, D. Clustering by passing messages between data points. *Science* 315, 972–976 (2007).
- [13] S. Patel, S. Sihmar and A. Jatain, "A study of hierarchical clustering algorithms," 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), 2015, pp. 537-541.
- [14] Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases. *ACM Sigmod Record*, 1996, 25(2): 103–11.
- [15] Ester, M, Kriegel, H P, Sander, J, and Xiaowei, Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. United States: N. p., 1996. Web.
- [16] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations", *Proc. of the Fifth Berkeley Symposium on Math. Stat and Prob.*, vol. 1, pp. 281-296, 1967.
- [17] Sculley, D.: Web-scale k-means clustering. In: *Proceedings of the 19th International Conference on World Wide Web (WWW 2010)*, pp. 1177–1178. Association for Computing Machinery, New York (2010).
- [18] G. Jones and B. Bhanu, "Recognition of Articulated and Occluded Objects" in *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 24, no. 07, pp. 603-613, 1999.
- [19] M. Ankerst, M.M. Breunig, H.P. Kriegel, J. Sander, Optics: Ordering Points to Identify the Clustering Structure. in *Proceedings ACM SIGMOD'99 International Conference on Management of Data (Philadelphia, PA, 1999)*
- [20] Ng, Andrew & Jordan, Michael & Weiss, Yair. (2002). On Spectral Clustering: Analysis and an algorithm. *Adv. Neural Inf. Process. Syst.* 14.
- [21] H. Wan, H. Wang, B. Scotney and J. Liu, "A Novel Gaussian Mixture Model for Classification," *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 2019, pp. 3298-3303, doi: 10.1109/SMC.2019.8914215
- [22] Singh, R. (2016), "A Regression Study of Salary Determinants in Indian Job Markets for Entry Level Engineering Graduates", Masters Dissertation. Dublin Institute of Technology.