International Journal for Modern Trends in Science and Technology, 6(12): 319-322, 2020
Copyright © 2020 International Journal for Modern Trends in Science and Technology
ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST061259
Available online at: http://www.ijmtst.com/vol6issue12.html

Salary Estimator using Data Science

Winner Walecha[1] | Dr. Bhoomi Gupta[2]

[1]B. Tech Scholar, Department of Information Technology,Maharaja Agrasen Institute of Technology, Delhi, India
[2]Assitant Professor, Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi, India

To Cite this Article
Winner Walecha and Dr. Bhoomi Gupta, “Salary Estimator using Data Science”, *International Journal for Modern Trends in Science and Technology*, 6(12): 319-322, 2020.

Article Info
Received on 12-November-2020, Revised on 05-December-2020, Accepted on 11-December-2020, Published on 15-December-2020.

ABSTRACT

This paper presents a salary prediction system using the job listings from an employment website, in this case Glassdoor.com. A data mining technique is used to generate a model which will scrape number of jobs from the employment website, clean it on the basis of number of factors including the rival companies, revenue and skill required thereby predicting the salary to be expected when applying for a data science job. Techniques like linear regression, lasso regression, random forest regressors are optimised using GridsearchCV to reach the best model. The model can be further extended to build a flask API thus can be deployed on the internet for public usage.

KEYWORDS:*Data Mining, Linear Regression, Lasso Regression, Random Forest*

I. Introduction

Humans respond to incentives or motives. Any person is ready to burn the midnight oil if provided with the right incentive. Salary is one of the many incentives that a person looks while applying for any job. Employment websites are flooded with over a million job postings offering variety of posts for a particular skill. Thus it becomes difficult for one to pick out a job offering a great pay for the skill one has to offer.

Having a good pay job is what almost every fresher entering into the industry dreams of. Most of them are unaware about which job offers what salary, whether or not the skill they put their heart into is enough for them to enter into a good paying job? They confuse themselves with all of the job postings on the employment website and end up picking up job which offers less than what they are capable of. Thus it would be better for a person to have an idea about the salary which one can look for while applying for a job.

This research, therefore proposes a model to estimate the salary for a person looking for a job in the field of Data Science. In this model data mining techniques are exploited to reach the desired result. Primarily 3 techniques were used.

1. Linear Regression
2. Lasso Regression
3. Random Forest.

The conclusion was formed on the basis of the mean absolute error obtained by using each of the above mentioned techniques on the train and test sets.

This project can facilitate saving the time required to go through the website for job search. The user can have the result at one click. Searching a perfect job is of utmost important as one bad choice can ruin the whole career and the skill gets wasted.

Data Science:

What Is Data Science?

A thorough study in 2013 reported 90% of the world’s data has been created within the last two years. In just two years, we've collected and processed 9x the amount of information than the previous 92,000 years of mankind combined. And it is not slowing down. It's estimated that we've already created 2.7 zettabytes of data, and by 2020, that number will grow up to an astounding 44 zettabytes.

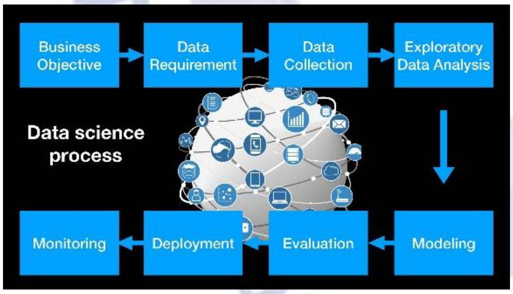
Every company says they're doing a form of data science, but what exactly does that mean? The field is emerging so rapidly, and revolutionizing so many industries, it is difficult to limit in its capabilities with a formal definition, but generally data science is related to the extraction of clean information from raw data for achieving the desired information.

Commonly referred to as the "oil of the 21st century," our digital data is of utmost importance in the field. It has thousands of benefits in business, research and our daily lives. Your route to work, your most recent Google search for the nearest coffee shop, your Instagram post about what you ate, and even the data about health from your fitbit are all important to different data scientists in numerous ways. Going through massive pool of data, searching for connections and patterns, data science is responsible for bringing us new products, delivering some great insights and help making our lives easier.

In a nutshell data science is all about the analysis of raw facts and extracting meaningful information from them.

Following picture depicts the thorough process of the data science.

Figure 1 Data Science Process



Techniques used in this research paper:

1. Linear Regression
2. Lasso regression
3. Random forest regression

Language used:

1. Python

Framework:

1. Jupyter Notebook (platform: Anaconda)

Visualisation tools:

1. MatplotLib

II. Methodology

1. Scraping:

Web Scraping , also known as web extraction, is a technique used to extract large amounts of data from websites/web pages where the data is extracted and saved to a local file in your computer or to a database in tabulated format.

Data displayed by most websites can only be viewed through a web browser. They do not offer the service to save a copy of this data for personal use. The only option remains is to manually copy and paste the data which is a very tedious job and can take many hours or sometimes days to complete. Web Scraping is the method to shorten this process in a way that instead of manually selecting and copying the data, the Web Scraping tool will perform the same task in much less time as compared to when it is done manually.

The web scraping tool used here is selenium along with the chromedriver.

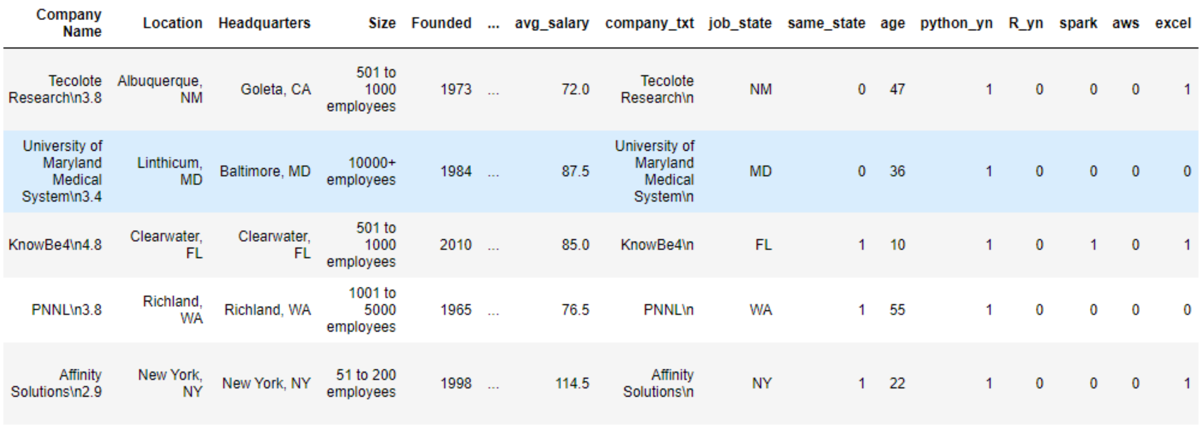
The url is entered as the parameter and the command is executed for 1000 jobs from the Glassdoor.com. The jobs are visited one by one and the scraper picks up the following attributes from each of them:

1. Job Title
2. Salary Estimate
3. Job Description
4. Rating
5. Location
6. Competitors
7. Industry
8. Revenue
9. Type of Ownership

The scraper returns a tabulated collection of 1000 jobs in a form of data frame (as called in python) and this data frame is exported to a csv file.

2. Data Cleaning

Data cleaning is the process of refining incorrect, corrupt, faulty formatted, repetitive or incomplete data within a dataset.



Company Name	Location	Headquarters	Size	Founded	...	avg_salary	company_txt	job_state	same_state	age	python_yn	R_yn	spark	aws	excel
Tecolote Research\n3.8	Albuquerque, NM	Goleta, CA	501 to 1000 employees	1973	...	72.0	Tecolote Research\n	NM	0	47	1	0	0	0	1
University of Maryland Medical System\n3.4	Linthicum, MD	Baltimore, MD	10000+ employees	1984	...	87.5	University of Maryland Medical System\n	MD	0	36	1	0	0	0	0
KnowBe4\n4.8	Clearwater, FL	Clearwater, FL	501 to 1000 employees	2010	...	85.0	KnowBe4\n	FL	1	10	1	0	1	0	1
PNNL\n3.8	Richland, WA	Richland, WA	1001 to 5000 employees	1965	...	76.5	PNNL\n	WA	1	55	1	0	0	0	0
Affinity Solutions\n2.9	New York, NY	New York, NY	51 to 200 employees	1998	...	114.5	Affinity Solutions\n	NY	1	22	1	0	0	0	1

Figure 2 Salary Data Cleaned

When combining multiple data sources, there are many ways for data to be replicated or differently labelled. If data is incorrect, outcomes and algorithms could not be trusted, even though they may look accurate. There is no absolute way to prescribe the precise steps in the data cleaning process because the processes vary from dataset to dataset. But it is of utmost importance to establish a template for the data cleaning process, so we know we are doing it the right way every time.

Here after collecting the data features were engineered from the text of each job description to quantify the value companies put on python, excel, aws and spark.

The numeric salary was parsed out and the rows without effective efficient data were removed.

Also, a column was added if the job was at the company's headquarters. The age of the company was also taken into account considering the year of foundation of company.

The job description was checked for the skills such as :

1. Python
2. R
3. Excel
4. AWS
5. Spark

And separate columns were made for the same. Job seniority and title for the job offered were also taken into the account.

Below is the snippet of the refined/cleaned dataset.

3. Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) is used by data scientists to review the data sets and summarize their main properties, often using data visualization methods. It helps in determining patterns, spotting anomalies, testing any hypothesis, or checking assumptions.

EDA is primarily used to see what data can disclose beyond the formal modeling or hypothesis testing task and provides a better view of data set variables and the relationships between them. It can also help in determining if the statistical techniques considered for data analysis are appropriate. EDA techniques are widely used method in the data discovery process today.

The main objective of EDA is to help look at data before reaching any conclusion. It can help in identifying obvious errors, as well as better understanding of patterns in the data, detect anomalies, and finding fascinating relations among the data variables.

Data scientists can use exploratory analysis to make sure the results they produce are correct and applicable to any desired goals. EDA also helps stakeholders by verifying they are asking the right questions. EDA can help in answering questions about categorical variables, confidence intervals and standard deviations. Once EDA is finished and insights are studied, its aspects can then be used for more reliable data analysis or modelling, including ML (Machine Learning).

EDA was performed on the dataset and libraries such as Matplotlib and seaborn were used to obtain a graphical representation of the data. Dependencies among various variables were checked and the results were obtained.

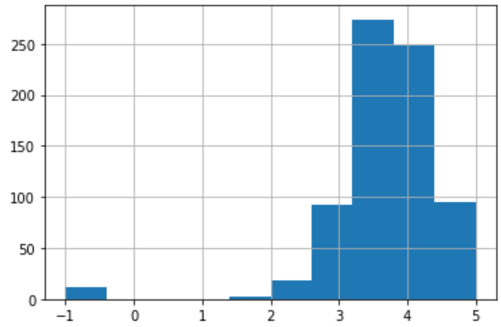


Figure 3 Rating Histogram

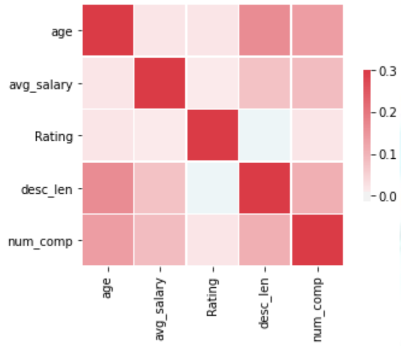


Figure 4 Heat Map

4. Model Building

In this process team needs to create data sets for training, testing, and production basis. These data sets make it possible for data scientist to formulate an analytical method and train it, while keeping some of data for testing the model.

Team create datasets for testing, training, and production purposes. In addition, in this phase, the team formulate and carry out models based on work done in the model planning process. The team also considers whether its existing mechanism will serve the purpose for running the models, or if it needs more robust environment for executing models and workflows.

For this process the categorical variables were converted into dummy variables. Then the data is split into train and tests sets having test size of 20%.

Then 3 different models were exploited namely, Multiple Linear Regression, Lasso Regression and Random Forest. These models were further evaluated using Mean Absolute Error.

III. RESULTS

Method Used	**Mean Absolute Error**
Random Forest	11.22
Linear Regression	18.86
Ridge Regression	19.67

Table 1 Comparison of Methods

The Random Forest model stood apart when compared to the other approaches on the test and validation sets.

IV.CONCLUSION

Regression analysis is a statistical process for predicting the relationship among variables and random forest is machine learning method capable of performing both regression and classification tasks .In Regression analysis we fit a curve to the data points, in a manner that the error between the value of fitted regression model at each point and the actual value is minimized .

Doing variable selection with Random Forest isn't insignificant. LASSO and its variants are exactly there for variable selection. So they are the evident superior choice for it. Even the linear LASSO's objective function is carefully designed to return a variable selection that explains an observation.

At a higher level, the aim of machine learning is prediction/estimation, whereas the aim of statistics is interpretation. Random Forest will certainly give superior predictions i.e. classifications.

REFERENCES

[1] https://dl.acm.org/doi/abs/10.1145/2500499

[2] https://onlinelibrary.wiley.com/doi/abs/10.1111/jbl.12010

[3] https://www.researchgate.net/publication/332276688_Data_science_big_data_and_statistics

[4] R Weihs, C., Ickstadt, K. Data Science: the impact of statistics. Int J Data Sci Anal 6, 189–194 (2018).

[5] Wu, X., Kumar, V., Ross Quinlan, J. *et al.* Top 10 algorithms in data mining. *KnowlInfSyst* **14,** 1–37 (2008). https://doi.org/10.1007/s10115-007-0114-2

[6] Annals of Data Science Yong Shi Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences, Beijing 100190, China.