INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 06 ISSUE: 05 | MAY - 2022 IMPACT FACTOR: 7.185 ISSN: 2582-3930

Salary Estimator using Big Data

Dr. Jayashree Rahul Pansare
Assistant Professor, Computer Engineering Department
Modern Education Society's College Of Engineering
Pune, India

Dr. Shubhangi R Khade
Assistant Professor, Computer Engineering Department
Modern Education Society's College Of Engineering
Pune, India

Sakshi Lunawat
Student, Computer Engineering Department
Modern Education Society's College Of Engineering
Pune, India

Janhavi Shivatare
Student, Computer Engineering Department
Modern Education Society's College Of Engineering
Pune, India

Jhanavi Oswal
Student, Computer Engineering Department
Modern Education Society's College Of Engineering
Pune, India

Krupa Mehta
Student, Computer Engineering Department
Modern Education Society's College Of Engineering
Pune, India

***Abstract*— This paper aims to present a salary prediction system. This system use the economic liberalization of Indian markets in early 90s boosted the economic growth of the nation in various sectors over the next two decades. One such sector that has seen a massive growth in this time is Information Technology (IT). The IT industry has played a very crucial role in transforming India from a slow moving economy to one of the largest exporters of IT services. This growth created a huge demand in the labour markets for skilled labour, which in turn made engineering one of the top choices of study after high school over the years. In addition, the earning potential and an opportunity to contribute to technology advancements after engineering, makes it a popular choice of study. These growth dynamics along with the diversified education and labor markets demands gives insight into the factors affecting the employment outcomes of engineering students job postings from job recruiting websites. The growth dynamics along with the employment outcomes of engineering Using big data techniques .Techniques like lasso and random forest are made in use to optimize the API**

Keywords*— *Feature Selection, Hypothesis Testing, Job Salary, Prediction engine, Random Forest, Regression, Salary Predictors, Support Vector Machines, Salary Predictor

I. INTRODUCTION

The growing enthusiasm for data-driven decision-making has created the importance of accurate and precise prediction over the previous years. The rapid growth of the data drives new opportunity for business and the process of analyzing the data quickly become more essential. Big data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size. A Job is nothing but a piece of work, especially a specific task one as part of the routine of one's occupation or for an agreed price. The estimates of the salary are on basis of various attributes. In recent time people are unable to overcome challenges of jobs and find a job with appropriate salary.

Most sites that offer salary information cover a limited range of jobs and industries. Hence, a Salary estimator comes to the aid of job-seekers to get easily accessible salary information.

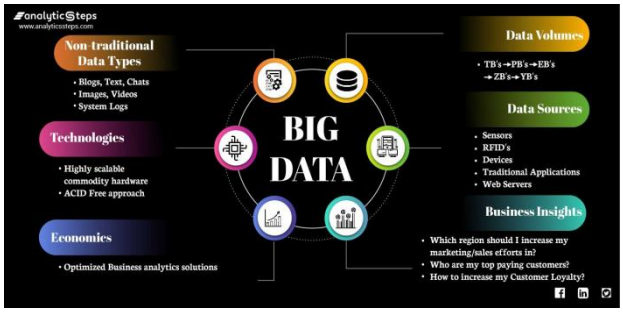
II. BIG DATA

A. What is Big Data

The definition of big data is data that contains greater variety, arriving in increasing volumes and with more velocity. This is also known as the three Vs. Put simply, big data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can't manage them. But these massive volumes of data can be used to address business problems you wouldn't have been able to tackle before.

B. How Big Data Works

Big data gives you new insights that open up new opportunities and business models.Big data brings together data from many disparate sources and applications. Traditional data integration mechanisms, such as extract, transform, and load (ETL) generally aren't up to the task. It requires new strategies and technologies to analyze big data sets at terabyte, or even petabyte, scale.During integration, you need to bring in the data, process it, and make sure it's formatted and available in a form that your business analysts can get started with. Big data requires storage. Your storage solution can be in the cloud, on premises, or both. You can store your data in any form you want and bring your desired processing requirements and necessary process engines to those data sets on an on-demand basis. Many people choose their storage solution according to where their data is currently residing. The cloud is gradually gaining popularity because it supports your current compute requirements and enables you to spin up resources as needed.Your investment in big data pays off when you analyze and act on your data. Get new clarity with a visual analysis of your varied data sets. Explore the data further to make new discoveries. Share your findings with others. Build data models with machine learning and artificial intelligence. Put your data to work.



III. METHODOLOGY

As this project is a Web Application there will be an option to provide input (Upload) to the user, this input will be passed to the backend and backend will perform the necessary computations on it and give the desired salary as output which will be passed back to the frontend. The data will be coming from the various job portals and we will create a dataset. We are assuming that the data acquired by the database is accurate. The whole of the output of the system depends on the accuracy of data.

A. Datasets

We had begun with comparing and taking datasets from multiple site base on various factors. For the following months we updated our dataset timely to be up-to- date.

B. Data Cleaning

- When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time. Data cleaning is no longer a new research field. It aims to improve the quality of data by identifying and removing errors and inconsistencies.
- Data cleansing process is complex and consists of several stages which include specifying the quality rules, detecting data error and repairing the error
- 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 salary is then parsed out from the collected data to be used in further computations for finding out an average salary.

C. Exploratory Data Analysis

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand,before getting them dirty with it.When some people claim that their methodology is exploratory, what they actually mean is that they are not sure what they are looking for. Unfortunately, poor research is often implemented in the name of EDA. During data collection, some researchers flood their subjects with hundred of survey items since their research questions are not clearly defined and their variables are not identified. While it is true that EDA does not require a pre-determined hypothesis to be tested, it does not justify the absence of research questions or ill-defined variables.

Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often with visual methods. EDA is used for seeing what the data can tell us before the modeling task. It is not easy to look at a column of numbers or a whole spreadsheet and determine important characteristics of the data. It may be tedious, boring, and/or overwhelming to derive insights by looking at plain numbers. Exploratory data analysis techniques have been devised as an aid in this situation.

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

D. Model Building

- Data Modeling is the process of analyzing the data objects and their relationship to the other objects. It is used to analyze the data requirements that are required for the business processes. The data models are created for the data to be stored in a database.
- Data Modeling helps create a robust design with a data model that can show an organization's entire data on the same platform.
- The data model makes sure that all the data objects required by the database are represented or not.
- The database at the logical, physical, and conceptual levels can be designed with the help data model.
- The relation tables, foreign keys, and primary keys can be defined with the data model's help.
- Data Modeling Tools help in the improvement of data quality.
- Data Model gives the clear picture of business requirements.
- Redundant data and missing data can be identified with the help of data models.
- In data models, all the important data is accurately represented. The chances of incorrect results and faulty reports decreased as the data model reduces data omission.
- The data models create a visual representation of the data. With the help of it, the data analysis gets improved. We get the data picture, which can then be used by developers to create a physical database.

- Better consistency can be qualified with the help of a data model across all the projects.
- The data model is quite a time consuming, but it makes the maintenance cheaper and faster.

ACKNOWLEDGMENT

It gives us great pleasure and satisfaction in presenting the final paper on 'Salary Estimator using Big Data'. We have furthermore to thank our Guides Dr.J.R.Pansare and Prof.S.Khade and Computer Department HOD Dr.(Mrs.) N. F. Shaikh to encourage us to go ahead and for continuous guidance..

REFERENCES

[1] Andreas Mullar, "Introduction to Machine Learning using Python: A guide for data Scientist," in O'Reilly Publisher, India.

[2] S. Marsland, Machine learning: an algorithmic perspective. CRC press, 2015

[3] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176,Oct., 2015

[4] Tzanis, George, et al. "Modern Applications of Machine Learning." Proceedings of the 1st Annual SEERC Doctoral Student Conference–DSC. 2006.

[5] Horvitz, Eric. "Machine learning, reasoning, and intelligence in daily life: Directions and challenges."Proceedings of. Vol. 360. 2006.

[6] Mitchell, Tom Michael. The discipline of machine learning. Carnegie Mellon University, School of Computer Science, Machine Learning Department,2006.

[7] Arum, R. (1998). The effects of resources on vocational student educational outcomes: Invested dollars or diverted dreams? Sociology of Education,71, 130-151.8. Lewis, C. D., 1982. Industrial and Business Forecasting Methods, London, Butterworths