IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 7, July 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907149 |

Salary Estimation Using Model Building Regression Techniques

Chaitanya Lahari Nayudu, Chetana Laasya Nayudu, G.Mounika, S.Sridevi

Student, Dept. of Computer Science and Engineering, Velammal Engineering College, Chennai, India

Student, Dept. of Computer Science and Engineering, Velammal Engineering College, Chennai, India

Student, Dept. of Computer Science and Engineering, Velammal Engineering College, Chennai, India

Assistant Professor, Dept. of Computer Science and Engineering, Velammal Engineering College, Chennai, India

ABSTRACT*:* The goal of this paper is to predict salary of a person after a certain year. The graphical representation of predicting salary is a process that aims for developing computerized system to maintain all the daily work of salary growth graph in any field and can predict salary after a certain time period.These days, the problem faced by employees is the lack of knowledge base to negotiate their salaries during their employment. Often, HR asks the interviewee – "How much salary are you expecting?". Since, the interviewee is not aware of various parameters to come to a conclusion, they settle for less. So, this proposed project lets you to estimate various salaries in a particular field from a beginner level job role to highest managerial job role. The Estimator gives you estimation by using the data collected from a particular website of around 1000 companies. This project optimizes linear, lasso and random forest regressions and will reach the best model and will built client facing API using flask.

KEYWORDS*:* API,regression.

I. INTRODUCTION

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains. Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data.The project undergoes predictive analysis to estimate the salaries and exploratory data analysis to establish the relations between the parameters that are used to estimate the salaries for the particular job role.

II. LITERATURE SURVEY

A."Salary Prediction Using Regression Techniques", 2020.

The goal of this paper is to predict salary of a person after a certain year. The graphical representation of predicting salary is a process that aims for developing computerized system to maintain all the daily work of salary growth graph in any field and can predict salary after a certain time period. This application can take the database for the salary system from the organization and makes a graph through this information from the database. It will check the salary fields then import a graph which helps to observe the graphical representation. And then it can predict a certain time period salary through the prediction algorithm. It can also be applied in some other effective prediction also.

B."Random Forest for Salary Prediction System to Improve Students' Motivation", 2016.
A salary prediction model was generated for graduate students using a data mining technique to generate for individuals with similar training attributes. An experiment was also conducted to compare the two data mining techniques Decision Trees ID3, C4.5 and Random Forest to determine the most suitable technique for salary prediction, tuned with key important parameters to improve the accuracy of the results. Random Forest gave the best accuracy at 90.50%, while Decision Trees ID3 and C4.5 returned lower accuracies at 61.37% and 73.96%, respectively for 13,541 records of graduate students using a 10-fold cross-validation method. Random Forest generated the best efficiency model for salary prediction. A questionnaire survey was conducted to determine usage evaluation with 50 samples. Results indicated that the system was effective in boosting students' motivation for studying, and also gave them a positive future viewpoint. The results also suggested that the students were satisfied with the implemented system since

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907149 |

it was easy to use, and the prediction results were simple to understand without any previous background statistical knowledge.

III. SYSTEM STUDY

A.SALARY PREDICTION TECHNIQUE

The graphical representation of predicting salary is a process that aims for developing computerized system to maintain all the daily work of salary growth graph in any field and can predict salary after a certain time period. This application can take the database for the salary system from the organization and makes a graph through this information from the database. It will check the salary fields then import a graph which helps to observe the graphical representation. And then it can predict a certain time period salary through the prediction algorithm. It can also be applied in some other effective prediction also.

B.NECESSITY FOR SALARY PREDICTION

Salary negotiation can run smoothly if you mention a salary range you would be satisfied with instead of stating a single amount.The main aim is predicting salary and making a suitable user-friendly graph. From this prediction the salary of an employee can be observed according to a particular field according to their qualifications. It helps to see the growth of any field. It can produce a person's salary by clustering and predict the salary through the graph. Using linear regression, lasso regression and random forest regression it makes a graph. This graph helps to predict the salary for any positions.

IV. DESIGN AND IMPLEMENTATION

This project attempts to predict the salary with respect to the company's previous value. It requires historic data of the company as the project also emphasizes on data mining techniques. So it is necessary to have a trusted source having relevant and necessary data required for the prediction. We will be using galssdoor website as the primary source of data. This website contains all details such as: Job title, Job description, rating, company, location, company headquarters, salary range etc. For each companies. It also provides the overall performance of companies of different categories.

There is no API provided by the website for providing data. We have written scripts to scrape all the required data from glassdoor website.

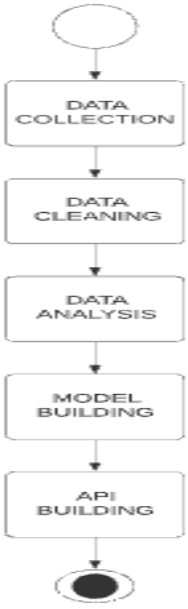

FIG 1: BLOCK DIAGRAM OF SALARY ESTIMATION

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907149 |

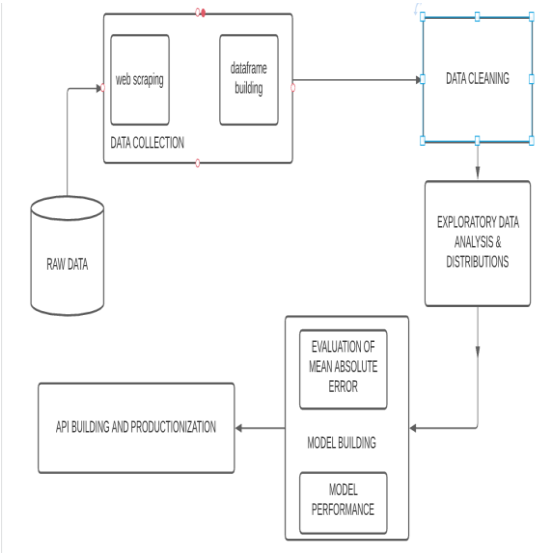

FIG 2: ARCHITECTURE DIAGRAM OF SALARY ESTIMATION

V. MODULES

A.Data Collection:
The data of around 1000 job postings from a website like glassdoor is collected which includes features like job title, job description rating, company, location, headquarters, salary range etc.
B. Data Cleaning:
The data collected is cleaned in this module which will be useful for analysis. So, from the existing features, we form new features and attributes that will be considered as various key parameters of the datasets.
C. Data Analyzing:
In this module, the exploratory data analysis (EDA) is used to look at the distributions of the data and value counts for the various categorical variables.
D. Model Building:
In this module, Different modules are evaluated using Mean Absolute Error. The models are multiple linear regression, lasso regression and random forest regression.
E. API building:
The flask API endpoint is builded which will takes in a request with a list of values from a job listing and returns an estimated salary.

VI. CONCLUSION

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. Random Forest models decide where to split based on a random selection of features. Rather than splitting at similar features at each node throughout, Random Forest models implement a level of differentiation because each tree will split based on different features. This level of differentiation provides a greater ensemble to aggregate over, producing a more accurate predictor. Rather than making one model and hoping this model is the best/most accurate predictor we can make, ensemble methods take a myriad of models into account, and average those models to produce one final model. Random Forest works well with: Data containing missing values and biased data. Since it randomizes the variable selection during each tree split it's not prone to over fit unlike other models. When fine-tuned, highly likely to perform better than LR. Linear Regression cannot handle non-linearity in the data. Other models don't perform very well, when the data set has more noise i.e. when target classes are overlapping. Naive Bayes makes a lot of assumptions like: two features are independent given the output class, normal distribution. Therefore, Random Forest Classifier is best for commentary classification.

VII. FUTURE ENHANCEMENTS

The project can be further deployed for other job roles and for other geographic locations.Accordingly, web scraping can be done to other websites and data can be collected for that particular roles.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907149 |

REFERENCES

[1] Ken Jee website
https://www.kennethjee.com/
[2] https://towardsdatascience.com/selenium-tutorial-scraping-glassdoor-com-in-10-minutes-3d0915c6d905
[3] Lumsden and L. S., "Student Motivation To Learn", *ERIC Digest*, no. 92, 1994.
[4] Y. Lee and M. Sabharwal, "Education—Job Match Salary and Job Satisfaction Across the Public Non-Profit and For-Profit Sectors: Survey of recent college graduates", *Public Management Review*, vol. 18, no. 1, pp. 40-64, 2014.
[5]J. Jerrim, "Do college students make better predictions of their future income than young adults in the labor force?", *Education Economics*, vol. 23, no. 2, pp. 162-179, 2013.
[6] P. khongchai and P. Songmuang, "Improving Students'Motivation to Study using Salary Prediction System", *13th International Joint Conference on Computer Science and Software Engineering (Preprint)*, 2016.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

Scan to save the contact details