WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Salary Estimator Using ML Algorithms

Aruna A G[1], Narmadadevi S D[2], Reena R[3]

Assistant Professor[1], Master of Science[2,3]
Decision and Computing Sciences Coimbatore Institute of Technology
agaruna@cit.edu.in[1], narmadadevi16@gmail.com[2], reenaramachandran2001@gmail.com [3]

ABSTRACT:

The increasing demand for data scientists in today's digital age has driven the need for efficient and accurate tools to estimate their salaries. In this paper, a Data Science Salary Estimator application is presented, which leverages web scraping techniques to gather salary data from Glassdoor. The paper outlines the data cleaning and preprocessing steps to prepare the collected data for modeling. Machine learning algorithms for supervised learning were employed to build predictive models. The application's user-friendly interface, developed using HTML and CSS, is powered by the Flask framework. Additionally, user details are securely stored in a MySQL database, ensuring data privacy and security. This work demonstrates a comprehensive solution for salary estimation in the field of data science, combining data acquisition, processing, modelling, and user interaction in a cohesive application.

Keywords: Flask, machine learning, data science, modelling, salary estimator

I. INTRODUCTION

The field of data science is expanding rapidly, and the demand for data scientists has increased correspondingly. Determining appropriate compensation for data science roles can be challenging, as it depends on various factors. The Data Science Salary Estimator application is a versatile decision support tool that gathers data from Glassdoor, cleans and preprocesses it, and uses machine learning models to predict data science salaries. This application also offers a user-friendly interface and securely stores user information in a MySQL database.

The development of this tool involves several key components: data acquisition, data preprocessing, modeling, and user interaction. Web scraping techniques were employed to collect salary data from Glassdoor, a popular platform for job seekers and employers. The acquired data was then carefully cleaned and preprocessed to ensure the quality and reliability of the models. These models take into account various features such as job location, experience, education, and company ratings. The combination of these models enables users to receive salary predictions with high accuracy.

The frontend of the application is developed using HTML and CSS, providing an intuitive and interactive user experience. Flask, a micro web framework, is used to handle user requests and serve the predictions. To maintain user privacy and data security, user details are stored in a MySQL database, ensuring that sensitive information remains protected. This paper presents the comprehensive development of the Data Science Salary Estimator application, highlighting the process from data acquisition to modeling and user interaction. By providing accurate salary estimations for data science roles, this tool aims to assist both job seekers and employers in the data science industry.

II. LITERATURE SURVEY

Pokpong Songmuang along with Pornthep Khongchai proposes system of salary prediction to enhance the motivation of students in college. A seven-feature prediction model was generated by using the technique of decision tree. Jobs which are related to the field of study of the students was also included in the paper to enhance the performance of the system.

A research study was done by Bhatia to estimate the UG student's salary using data mining techniques. Data mining is an essential concept as it provides the effective and efficient learning to the students and researchers.

Karla et al used hierarchal linear regression to build a prediction model with students and program characteristics as control variables and salary as the predictor variable.

A deep learning technique was used by Thongchai which estimates the monthly salary of job for a labour workforce in Thailand where it helped job seekers in Thailand solving a regression problem which is a numerical outcome is effective.

III. METHODOLOGY

A. Data Collection

The dataset for the Salary Estimation App for Data Science Jobs was obtained from Glassdoor, a leading job listing and company review platform, through a web scraping process facilitated by Selenium. This approach automated the search for data science job listings, navigation through multiple result pages, and extraction of essential information including job titles, company names, job descriptions, location, and salaries. Strict adherence to ethical considerations, such as respecting rate limits and avoiding the collection of sensitive data, was maintained. The collected data underwent thorough cleaning and preprocessing to ensure data quality and reliability, serving as the foundational dataset for subsequent machine learning model development. It has the below attributes.

Unnamed: 0	int64
Job Title	object
Salary Estimate	object
Job Description	object
Rating	float64
Company Name	object
Location	object
Headquarters	object
Size	object
Founded	int64
Type of ownership	object
Industry	object
Sector	object
Revenue	object
Competitors	object

B. Data Pre-processing

Salary parsing is done, minimum and maximum salary is found and average salary is kept as a separate attribute. Similarly, the name and state attribute are processed and data cleansing is done. The company age is found with the help of the founded year attribute. After which, job description parsing is done where the key skills for the data science job is found and made as a separate attribute. Job title simplification is also done and the jobs are categorized into seniority order and made as a separate attribute. Finally, the unnecessary columns are dropped out. Figure 1 shows the dataset with the created attributes job title and seniority division. Also, it shows the added attributes of the skillset that is extracted from the job description attribute.

age	python_yn	R_yn	spark	aws	excel	job_simp	seniority
16	1	0	0	1	1	mle	na
15	1	0	0	0	0	analyst	senior
172	0	0	0	0	1	manager	na
12	1	0	0	1	1	data engineer	na
6	0	0	0	0	0	na	senior

Fig.1 Data Cleaning

C. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed to extract valuable insights from the dataset. EDA is a fundamental data science practice that facilitates a comprehensive understanding of dataset characteristics. This analysis involved various visualization techniques, including histograms, barplots, box plots, and word cloud analysis.

Histograms were employed to depict the distribution of numerical variables. Figure 2, for instance, illustrates the 'age' histogram, revealing a bell-shaped curve characteristic of a normal distribution and providing insights into the age distribution among job records. Barplots were used to visualize the distribution of categorical data, offering a view of the diversity of job titles and other relevant categorical variables.

Box plots, as depicted in Figure 4, aided in comparing the distribution and spread of numerical data across different job titles, facilitating the identification of potential outliers and variations within the dataset. Additionally, word cloud analysis was utilized to unveil frequently occurring terms within job descriptions, as presented in Figure 5. This process shed light on recurring themes and keywords, contributing nuanced insights into the qualifications and skills sought by employers in data science job postings.

In summary, the EDA process proved instrumental in extracting crucial insights and patterns from the dataset, setting the stage for subsequent analyses and model development. These visualizations significantly enhance the effectiveness of the Salary Estimation App for Data Science Jobs.

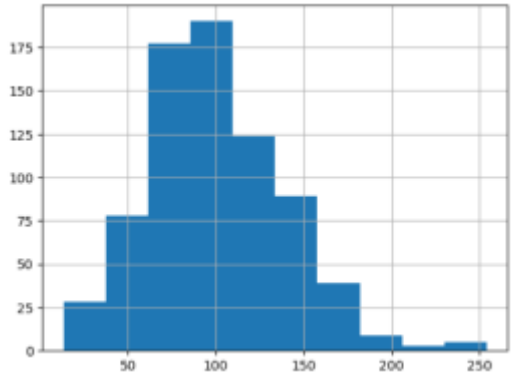

Fig.2 Histogram (Age)

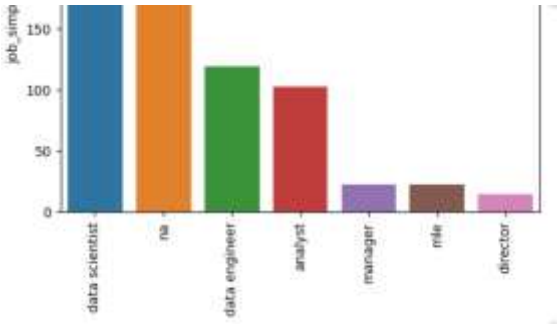

Fig.3 Barplot (Job Title)

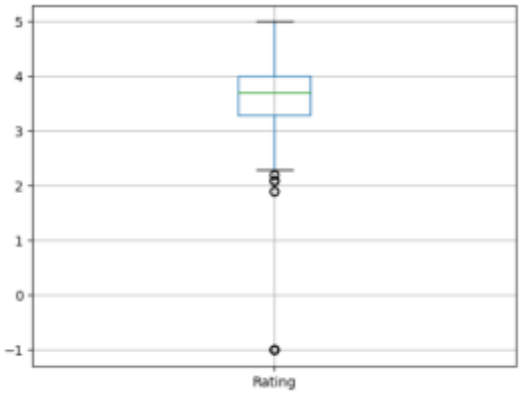

Fig.4 Boxplot (Rating)

Fig.5 Word Cloud (Job Description)

D. Feature Selection

Feature selection aims to remove features that don't contribute to our predictive modelling. It includes features that don't contribute to target class differences as well as highly correlated features, which can cause multicollinearity issues. Heat map correlation is applied for feature selection. Multi collinearity does not exists between the independent features. Hence, all the features are taken into consideration. The correlation plot is shown in Figure 7.

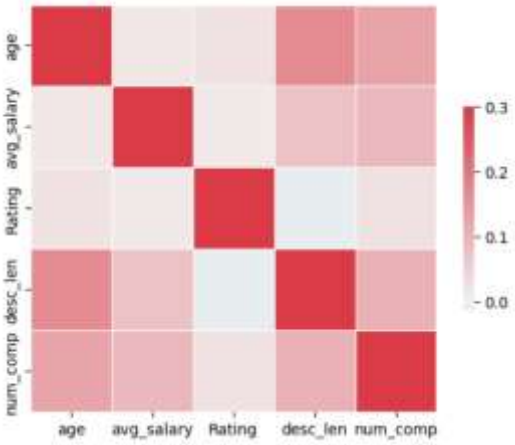

Fig.7 Correlation plot

E. Model Building and Evaluation metrics

Categorical features are converted into numerical features using pandas get dummies function and then train test split is done with the testing dataset size 20 percent in order to give it to the model. Figure 8 shows the metrics of the negative mean squared error.

Linear regression is a widely used statistical method in predictive modeling. Linear regression is particularly valuable for prediction, forecasting, and understanding the relationships between variables. However, it's essential to acknowledge the assumptions and limitations of linear regression, such as linearity, independence of errors, and homoscedasticity.

Lasso regression, an extension of linear regression, introduces regularization to the model. It includes a penalty term that encourages the selection of the most important independent variables and pushes the coefficients of less important variables to zero. λ is the regularization parameter. Lasso regression is valuable for feature selection and preventing overfitting, particularly in high-dimensional datasets. Researchers can choose an appropriate value for the regularization parameter (λ) to control the level of regularization.

Random Forest consists of multiple decision trees, each trained on a random subset of the data and a random subset of the features. The predictions from each tree are combined to produce a final prediction, making Random Forest robust, resistant to overfitting, and capable of handling large datasets.

Comparing the models linear regression, lasso regression and random forest, random forest provides the less negative mean squared error so we are finally considering the random forest model and pickling is done for the random forest model.

The mean absolute error for the testing set for the random forest model is around 11 and the r square value is around 80% which is comparatively better than ordinal least squares which had given an r square value of around 70% only.

	Model	Accuracy
0	Linear Regression	-20.770234
1	Lasso Regression	-19.262516
2	Random Forest	-15.199075

Fig.8 Performance metrics

IV. API DEVELOPMENT

The salary estimator web application is built using flask framework. The application allows users to log in, input various job-related data, and get a predicted average salary based on the provided information and also it gives feedback to learn technologies that helps in getting better salary in related field. A Flask application is created and MySQL database is configured to store user account information, including usernames, passwords, and emails. Figure 9 shows the user information stored in the MYSQL database when the user register in the application. A machine learning model is loaded from a file using the pickle module. This model is used to predict average salaries based on the input data. User authentication is handled by checking username and password against the database. If the credentials are correct, the user is logged in. When users submit job-related data on the input page, the data is collected and processed. The machine learning model is used to predict the average salary based on the provided data. The predicted salary is displayed on the output page along with the feedback. Figure 10 shows the salary prediction page where after registering and logged in, user can input the required details, average salary along with the feedback to improvise the salary will be displayed to the user in the output page.

id	username	password	email
2	Reena	1234	reenaramachandran2001@gmail.com
3	abishek	1234	abishek@gmail.com
4	jothi	qwerty	jothi@gmail.com
5	rishi	123456	rishi@gmail.com
6	raveena	123456	raveeena@gmail.com
7	Vijay	7890	vijay@gmail.com
8	Narmada	12345	narmadadevi16@gmail.com

Fig.9 User Information

Fig.10 Salary Prediction

V. CONCLUSION

The developed Salary Estimation App for Data Science Jobs represents a comprehensive solution for job seekers in the data science field. This research effort successfully combines web development, machine learning, and database management, offering users a feature-rich platform for estimating their potential salaries and acquiring valuable insights with feedbacks as well.

The application commences with user registration and login procedures, ensuring secure storage of user information and consistent access to the platform. Through an intuitive HTML and CSS front-end, users can input their details and receive estimates of average salaries relevant to their profiles along with the feedbacks. These estimates are accompanied by informative feedback, contributing to an enhanced understanding of the factors influencing salary levels in the data science industry.

Central to the salary estimation functionality is a Random Forest-based machine learning model. This model is trained on a dataset that is extracted by web scraping from Glassdoor, containing pertinent features, ensuring the provision of accurate salary predictions. The model is serialized using the pickle library, facilitating its seamless deployment within the Flask web framework.

In addition to its user-facing capabilities, the application effectively manages user data, ensuring secure storage in a MySQL database. This meticulous approach guarantees the confidentiality and privacy of user information.

The successful integration of front-end web development, machine learning, and database management underscores the potential of such integrated tools in improving job search experiences. This project demonstrates the efficacy of data science in addressing real-world challenges and providing valuable insights to job seekers, empowering them to make informed career decisions.

In summary, the Salary Estimation App for Data Science Jobs serves as an exemplar of data-driven technology's potential in career advancement. By offering data science job seekers accurate salary estimates and informative feedback, this application plays a crucial role in streamlining the job search process and fostering individual success within the field.

References

[1]http://www.ijmtst.com/volume6/issue12/59.IJMTST0612210.pdf

[2]https://ijirt.org/master/publishedpaper/IJIRT151548_PAPER.pdf

[3]http://www.ir.juit.ac.in:8080/jspui/bitstream/123456789/3763/1/Salary%20Prediction.pdf

[4]https://www.ijcseonline.org/spl_pub_paper/IJCSE-ICACICT-2019-16.pdf

[5]https://www.researchgate.net/publication/362280362_Salary_Prediction_in_Data_Science_Field_Using_Specialized_Skills_and_Job_Benefits_-A_Literature_Review

[6]https://www.academia.edu/79455029/SALARY_ESTIMATOR_A_LITERATURE_REVIEW

[7] https://www.ijariit.com/manuscripts/v8i3/V8I3-1357.pdf

[8]https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3990877

[9]http://103.47.12.35/bitstream/handle/1/9318/BT4234_RPT%20-%20Mr.%20Sreenarayanan%20N%20M.pdf?sequence=1&isAllowed=y

[10]http://ijasret.com/VolumeArticles/FullTextPDF/842_47._SALARY_PREDICTION_USING_MACHINE_LEARNING.pdf