EasyChair Preprint

№ 5960

Projection of Employee Wages

Shirisha Gangarapu, M. Raja Sekar and T. Sunil Kumar

June 30, 2021

PROJECTION OF EMPLOYEE WAGES

G. Shirisha[1], M. Raja Sekar[2], T. Sunil Kumar[3]

[1]M-Tech, Dept. of Computer Science and Engineering VNR Vignana Jyothi Institute of Engineering Technology, Hyderabad 500090, India.

[2]Professor, Dept. of Computer Science and Engineering VNR Vignana Jyothi Institute of Engineering Technology, Hyderabad 500090, India.

[3]Asst. professor, Dept. of Computer Science and Engineering VNR Vignana Jyothi Institute of Engineering Technology, Hyderabad 500090, India.

shirishagangarapu@gmail.com[1], rajasekar_m@vnrvjiet.in[2], sunilkumar_t@vnrvjiet.in[3]

ABSTRACT

The graphical representation of anticipating pay is a process that attempts to build a computerised system that can follow all daily salary growth graphs in any location and predict revenue after a certain period of time. This application may access the company's salary database and generate a graph based on the information contained therein. Before importing a graph to aid in the visual representation, it will validate the salary fields. It can then forecast a wage for a specified time period using the prediction method. Accurate staff recruiting is a vital component of a company's business strategy because it has such a large impact on productivity and competitiveness. This research looks at an effective and efficient job portal analysis and pay prediction method that allows any student (freshman) or professional to choose the best platform for a career by learning which skills are in demand in the IT sector.

Keywords: Predict Salary, Logistic Regression, Classification, and Machine Learning

1. INTRODUCTION

Knowing where you are now and where you want to go next is one of the most significant aspects of the job search. A job seeker's perspective is similar to that of a product manager, who is looking for the best product and market fit. Product managers that are good at what they do know what their customers want. As a result, they thoroughly research consumers' wants, which are presented in job descriptions as "what you must have." Smart product managers would personalise

their products, such as resumes, cover letters, online profiles, and portfolios, to display these features and attributes as much as possible after understanding what is usually needed in the market. Because of its impact on productivity and competitiveness, accurate personnel recruiting is a critical component of every company's business plan.

- Helping to see the growth at any field.
- With the help of machine learning it can easily produce a graph.
- Marketing easy to estimate the salary between x-y axes.
- User can give any point to get the salary through the program.
- Salary of the employees can be observed to give them a particular field according to their qualifications.

Several algorithms have been presented in the field of job recommendation that recommend jobs based on candidate profiles gathered using clustering algorithms.

1.1 How are the salary estimated?

Use your salary forecast to get a sense of where you are now and to help you evaluate your current income and job offers. CTC is never equal to his or her take-home pay. There are several components of the CTC that are not included in the take-home pay. A person's basic pay is the amount of money he or she earns on a regular basis. It is a portion of one's compensation package that cannot be changed. A basic pay is determined by the employee's position as well as the industry in which they work.

CTC = Gross Salary + PF + Gratuity

We can estimate prices for even unusual combinations of location, title, experience, and abilities by projecting how each component affects prices using a predictive machine-learning model rather than a data sampling methodology. This avoids the problem that many other salary calculators have in providing you with a useful wage estimate due to sample size limitations.

Net Salary = Basic Salary + HRA + Allowances - Income Tax - Employer's Provident Fund - Professional Tax

1.2 Salary Deductions:

The purpose is to make sure that this is a constant component in your pay check that serves as the foundation for other parts of your income. HRA, for example, is defined as a percentage of this basic wage (as determined by the firm). Your PF is withdrawn from your base pay at a rate of 12%. It normally accounts for a significant amount of your entire pay.

Salaried employees account for the majority of all taxpayers in the country, and their contribution to tax collection is substantial. With the help of these deductions and exemptions, a person's tax liability can be significantly reduced. The goal of implementing a standard deduction is-

- Provident Fund - This is a statutory deduction that every employer is required to make.
- Income Tax - IT is deducted according to the income tax slabs in India.
- Insurance deductibles are determined by an employee's position, amount of experience, and years of service.

2. LITERATURE SURVEY

We show in this study that there is a tremendous quantity of data available in the medical business to acquire knowledge from big volumes of data using an analytical method. Incomplete values, higher dimension values, and interference, among other issues, are common in real-time data sets and are ineffective for the full classification.

[1] Ashok Kumar S et.al tells that candidate profile functionality and features, posting new recruiting programmes by company, publishing new carrier guidance placement programmes by Admin, and so on. Viewing resumes by company is a great way to start.

[2] R.A. Huebner et.al gives in-depth understanding of the teaching and learning process, allowing for more effective educational planning. This survey work focuses on the components, research trends, and educational outcomes of EDM, highlighting its related Tools, Techniques, and outcomes. It also emphasises EDM's challenges.

[3] Sayan Das et.al created a graph based on the information in the database. It will check the salary fields before importing a graph to aid in the visual depiction. Then, using the prediction method, it may forecast a wage for a specific time period. It can also be used to make other accurate predictions.

[4] Jackman, S., & Reid, G.* et.al this research offers a lot of information, mostly in the form of unstructured text descriptions of the available positions. It's useful to estimate a salary based on the job description for those advertisements that don't specify one.

[5] Singh, R. et.al this data also demonstrated that in Indian labour markets, cognitive skills such as English and quantitative competence, as well as a desire to complete a work properly, are

major contributors to the starting income of engineering graduates.

3. METHODOLGY

Employers are fully aware that certain qualities, such as a strong work ethic, critical thinking, excellent communication, and teamwork, are desired by undergraduate students. This project will assist you in becoming that smart product manager by advising you on what attributes and top abilities the market seeks, as well as assisting you in navigating pay negotiations and making the best decision possible.

Proposed method for Projection of Salary

Step 1: Salary information was extracted from the dataset.

Step 2: In the graph, the points corresponding to an individual's salary data have been plotted. Using the dataset, we plotted the points on the graph according to number or input, resulting in the true dataset.

Step 3: We then create lines between the sites using linear regression.

Step 4: We use polynomial regression to curve the points if they are not in a linear order. We may create a smooth and curved path using the clustering points.

Step 5: We can then anticipate salary using the linear/polynomial graph on the x-y axis.

Step 6: We also forecast a person's future salary position based on the graph. Take a specific person's position, and then use the graph to calculate the prediction response.

We examine multiple models to discover the classifier that best describes the data in a job recruitment situation characterised by high levels of noise, high dimensionality, and a restricted quantity of samples after phrasing the prediction issue as a classification challenge. Our main focus is on pay range prediction, as this might lead to improved categorization of job postings and hence easier navigation for end-users.

3.1 Dataset:

This dataset contains various features to define how the employee performance is measured for future benefits. It consists of fields like time spent on work which is quite important, work accident which is a bit risk to any company and if higher risk or work accident then employee have less salary and risk of eliminating too. This helps organizations to use data to assess the effectiveness of a strategy and gathering data allows you to see how effectively

your solution is working and whether it needs to be altered in the long run.

3.2 Data Pre-processing:

Missing values are supplemented, outliers are identified and removed, and self-contradiction is resolved via data pre-processing. Because the sample code number has no bearing on illnesses, it has been removed from the dataset. In the dataset, there are 16 trait values that are missing. The class's missing traits are replaced by the mean. In addition, the dataset uses random selection to ensure that the data is circulated properly.

3.3 Data Visualization:

Almost every profession requires data visualisation. Teachers may use it to display student exam results, computer scientists may use it to research artificial intelligence (AI), and CEOs may use it to share information with stakeholders. It is also crucial in large-scale data efforts. It's also part of the larger data presentation architecture discipline, which strives to find, package, and convey data in the most efficient way possible.

Figure 1: Graph for estimating salary vs experience

The following are some of the other advantages of data visualisation:

- Ability to absorb information rapidly, develop insights, and make quicker judgments;
- A better grasp of the next steps that must be made to develop the company; a better ability to keep the audience's attention with information that they can grasp;
- A simple means of disseminating information that increases the possibility to discuss ideas with all parties involved.

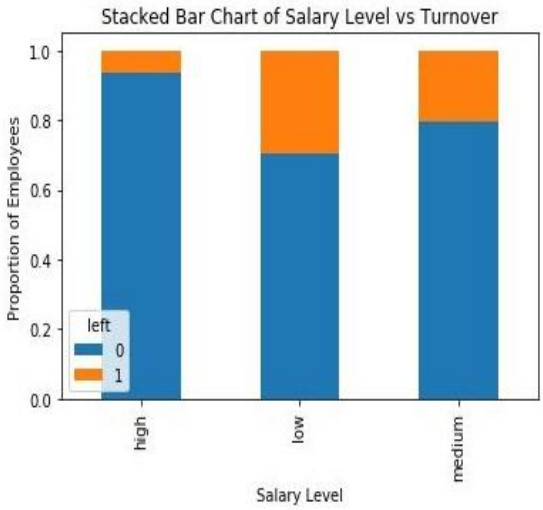

Figure 2: Bar graph for salary vs proportion of Employees

3.4 Training and Testing Phase:

The training step extracts the dataset's features, whereas the testing phase determines how the appropriate model performs in terms of prediction. This is the stage of training and testing. K fold cross-validation shows that a single fold is used for testing and k 1 folds are used in a circular fashion for training. Cross-validation is a technique for avoiding overfitting. In our work, data is partitioned using a ten-fold cross-validation methodology, with nine folds used for training and one-fold utilised for testing in each iteration.

3.5 Model Selection:

Applying Machine Learning to Any Dataset is at its most interesting stage. For forecasting the best results, it's also known as Algorithm selection. When dealing with huge data sets, Data Scientists typically employ a variety of Machine Learning algorithms.

However, all of those distinct algorithms may be divided into two categories at a high level: supervised learning and unsupervised learning. The term "supervised learning" refers to a system in which both the input and the desired output data are provided. To offer a learning framework for future data processing, input and output data are labelled for classification. Regression and Classification are two types of supervised learning issues.

Learning without Supervision: Unsupervised learning occurs when an algorithm is given information that is neither classed nor labelled and is allowed to act on it without supervision. Given the statistical noise in the data, the incompleteness of the data sample, and the limits of each model type, all models have some predictive inaccuracy. In order to optimally reveal the structure of the problem to the learning algorithm, certain techniques require specialized data preparation. The most prevalent approaches for supervised learning are regression and classification.

Algorithm	Cross-Validation	Accuracy
Logistic Regression	0.96	77%
SVM	0.96	90%
Random Forest	0.96	97%

Table 1: Accuracy for different algorithms

Random forest is a decision tree-based methodology for modelling predictions and behaviour analysis. It contains a number of decision trees, each of which represents a unique classification of data fed into the random forest. The random forest methodology considers each case separately, selecting the one with the most votes as the chosen prediction.

	precision	recall	f1-score	support
0	0.98	0.98	0.98	3462
1	0.95	0.95	0.95	1038
accuracy			0.98	4500
macro avg	0.97	0.97	0.97	4500
weighted avg	0.98	0.98	0.98	4500

Figure 3: Performance of Model using Random Forest

Logistic regression is a statistical model that uses a logistic function to represent a binary dependent variable in its most basic form, though there are many more advanced variants. Logistic regression is a type of regression analysis that involves estimating the parameters of a logistic model. A binary logistic model mathematically has a dependent variable with two possible values, such as pass/fail, which is represented by an indicator variable, with the two values labelled "0" and "1." An example of a logistic regression equation which is shown below and describes each term as:

$$y = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)})$$

	precision	recall	f1-score	support
0	0.81	0.92	0.86	3462
1	0.51	0.26	0.35	1038
accuracy			0.77	4500
macro avg	0.66	0.59	0.60	4500
weighted avg	0.74	0.77	0.74	4500

Figure 4: Performance of Model using Logistic Regression

For two-group classification issues, a support vector machine (SVM) is a supervised machine learning model that uses classification techniques. Support vector machine (SVM) models can categorise new text after being given sets of labelled training data for each category. A simple example is the easiest way to understand the fundamentals of Support Vector Machines and how they work.

	precision	recall	f1-score	support
0	0.96	0.92	0.94	3462
1	0.77	0.86	0.81	1038
accuracy			0.91	4500
macro avg	0.86	0.89	0.88	4500
weighted avg	0.91	0.91	0.91	4500

Figure 5: Performance of Model using SVM

They aim to maximise whatever assessment criteria you're using, and they might take shortcuts to do so. For example, if you had two classes, one with 99 examples and the other with only one, a model could always correctly predict the first class 99 percent of the time! The model would have a high accuracy rating, but it wouldn't be able to help you find examples of the smaller class.

Figure 6: Final analysis

Random forests, also known as neural nets, provide estimates for variable relevance. They also provide a preferable way for dealing with data that is missing. Random forests outperform all other classification methods in terms of accuracy. The random forest methodology can also handle large datasets with thousands of variables.

4. CONCLUSION

It is self-evident that industries will not be able to survive in the long run unless they have predictive analytics skills in human resource management. Employee attrition is one of the most pressing issues confronting businesses today. The cost of replacing an employee and the time it takes to do so might divert attention away from the company's bottom line. In comparison to Logistic Regression and Support Vector Machine, the results showed that Random Forest Regression produces the best results. Instead, by identifying at-risk employees, analytics may be utilised to prevent this from happening in the first place. Predictive analytics can identify who is most likely to leave the company by analysing criteria such as compensation, promotions, corporate culture, and relationships with managers.

5. REFERENCES

[1] Ashok Kumar S, Santhosh V, "Prediction of Job Applicant Salary and Designation using Machine Learning", Department of Computer Science and Engineering Kalyani Government Engineering College, India.

[2]R.A. Huebner, "A survey of educational data mining research," Research in Higher Education Journal.

[3] Rupashri Barik, Sayan Das,‟ Salary Prediction Using Regression Techniques", JIS College of Engineering, Kalyani, Nadia.

[4] Jackman, S., & Reid, G. (2013), "Predicting Job Salaries from Text Descriptions", Doctoral dissertation, University of British Columbia.

[5] Singh, R. (2016), "A Regression Study of Salary Determinants in Indian Job Markets for Entry Level Engineering Graduates", Dublin Institute of Technology.