2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)

Design of a novel Prediction Engine for predicting suitable salary for a job

Sananda Dutta
Department of Computer Science and Engineering
Kalyani Government Engineering College
Kalyani, India
sananda.dutta95@gmail.com

Airiddha Halder
Department of Computer Science and Engineering
Kalyani Government Engineering College
Kalyani, India
airiddha@gmail.com

Kousik Dasgupta
Department of Computer Science and Engineering
Kalyani Government Engineering College
Kalyani, India
kousik.dasgupta@gmail.com

***Abstract*—Prediction engine is a tool which can forecast a future outcome using a set of past observation. In present days Prediction engines have become increasingly popular as they are producing accurate and affordable predictions almost similar to human. The additional advantage of using prediction engine is, it does not take decisions by itself it predicts and leaves the decision making to human/users. These prediction engines are now being using by researchers and industry for prediction of plethora of problems. Providing suitable and justified salary for any job has always been a challenging problem for not only the employer but it is also an important factor for the employee. For an individual searching for a job, salary may not be the prime factor but it is an important factor along with the others factors to meet one's basic human needs. In this paper, machine learning techniques are used to automate and formulate a proposed model for salary prediction. The proposed Prediction engine can predict the salary on the basis of some key features. The proposed approach a raw dataset is fit into decision models like decision tree and ensembles model. The results obtained are encouraging and with high precisions.**

Keywords— Data analysis, Data Processing, Prediction engine, Decision tree classifier, Random Forest classifier

I. INTRODUCTION

In present days industries and business are growing rapidly. A business or institute which has a website, a profile in social media, and accepts electronic payments in any form, etc., must have data about customers, user experience, web traffic, and more. According to figures from GEM Global Report [1] every second over three businesses are launched and with this amount of total data is also increasing.

Modern businesses houses are increasingly stressing on capability to analyze data and act accordingly to make their business successful. Ever increasing competition in the business industry makes it inadvertent for business houses to react quickly to the change in demands from customers and environmental conditions. It becomes more difficult, when the decisions are increasingly complex and the companies compete in a global marketplace, where so many factors are affecting the business. Managers need to analyze large amount of data before they can take the necessary decisions. According to present day statistics, any business that has been running for a year or more than that likely has "a ton of data" [2] in their record which can help them to make better decisions. But to make a reliable decision based on such big amount of data is impossible for a human being, which is why we need a model which can do this for us in most relevant way. These models also need to make an estimate of the financial and economic consequences associated with the decision or phenomena that are always having an iota of uncertainty. By using machine learning techniques one can design such model or Prediction engine that can do the task intelligently. Such a prediction engine [3] takes certain key features as input and gives the most favorable outcome by analyzing all the factors based on the previous data.

The problem of salary projections or prediction has always been a challenging proposing. To decide a fair salary for a job in general we need to consider one's qualification, work experience, no of projects, place where he or she will be working, category, company etc. Each of these aspects needs particular assessment and rules to decide salary, i.e. to decide one's salary we need to consider each aspect separately to decide possible amount of salary based of each attributes. Finally, the employers need to merge them by making all possible combinations and then only one can calculate fair and precise salary for the job.

The work of such projection has been studied by Shapiro [4] and Carriere, J. F., & Shand, K. J. [5] where the problem has been seen through historical context for projection of pension of employees. However, in the present day economic the context and scenario has changed a lot where the present salary of an employee is the major you too crux. Bone & Mitchell [6] have put forward for obtaining more data and constructing better models so that actuaries can make better estimates. Similarly, Shiller [7] also has emphasized on the importance of good data sources. Braskamp et.al. [8] has made a detailed study on determine salary equity. CB Johnson et. al. [9] have studied on prediction of salary in education institutions. Shaun Jackman and Graham Reid [10] have worked Predicting Job Salaries from Text Descriptions, they have tested a variety of regression models including maximum-likelihood regression, lasso regression, artificial neural networks and random forests. Singh. R [11] has done a detailed study to find the salary determinants of fresh undergraduate engineers in Indian job markets. Compounded is the fact that there are several job search sites that proposes jobs for individuals. Also individual profiles of employees are rarely modeled thus a moderate or average salary is not attractive for modern age individuals. A balance should be found between the not only the "social side" of the individual and the financial viability of the employee. Such a complex problem with certain amount of uncertainty is always a challenging proposition. In this paper we concentrate on individual salary models and predictions. To do this efficiently and ethically according to latest market policy and trends, we need to take the reference of large

978-1-5386-7638-7/18/$31.00 ©2018 IEEE

amount of previous data. Taking all the features and huge previously related data into consideration is not possible for a human being which is why we have tried to make it easy by introducing this prediction engine.

Whereas the Section I gives an introduction to the problem statement and does literature survey. The rest of the paper is organized as follows; Section II explains the preprocessing steps involved with the raw data set. Section III details the data processing and the proposed Prediction engine is explained in Section IV. Results of the model is given in section V and section VI concludes the paper with future scope.

II. PREPROCESSING

A. Overview of the Dataset

The proposes Prediction engine has been tested on to decide the salary for jobs in UK to improve the experience of peoples searching for jobs, and help employers and jobseekers to figure out the latest market worth for different positions. The dataset contains id, job description, title, location, contract type, contract time, company, category, source name. Based on the above mentioned features our Prediction engine will predict salary.

B. Description of the Dataset

The dataset used in this work is collected and uploaded to www.kaggle.com [12] by ADZUNA [13], from where we've downloaded this dataset. It contains 244,768 data points; a snapshot of the data set is given in Fig. 1. We have divided our dataset in two parts- train dataset and test dataset and the divide ratio is 1/3rd in test set and 2/3rd in train set.

	Id	Title	FullDescription	LocationRaw	LocationNormalized	ContractType	ContractTime	Company	Category	SalaryRaw	SalaryNormalized	SourceName
0	12612628	Engineering Systems Analyst	Engineering Systems Analyst Dorking Surrey Sal...	Dorking, Surrey, Surrey	Dorking	NaN	permanent	Gregory Martin International	Engineering Jobs	20000 - 30000/annum 20-30K	25000	Cv-library.co.uk
1	12612830	Stress Engineer Glasgow	Stress Engineer Glasgow Salary **** to **** We...	Glasgow, Scotland, Scotland	Glasgow	NaN	permanent	Gregory Martin International	Engineering Jobs	25000 - 35000/annum 25-35K	30000	Cv-library.co.uk
2	12612844	Modelling and simulation analyst	Mathematical Modeller / Simulation Analyst / O...	Hampshire, South East, South East	Hampshire	NaN	permanent	Gregory Martin International	Engineering Jobs	20000 - 40000/annum 20-40K	30000	Cv-library.co.uk
3	12613049	Engineering Systems Analyst / Mathematical Mod...	Engineering Systems Analyst / Mathematical Mod...	Surrey, South East, South East	Surrey	NaN	permanent	Gregory Martin International	Engineering Jobs	25000 - 30000/annum 25K-30K negotiable	27500	Cv-library.co.uk
4	12613647	Pioneer, Miser Engineering Systems Analyst	Pioneer, Miser Engineering Systems Analyst Do...	Surrey, South East, South East	Surrey	NaN	permanent	Gregory Martin International	Engineering Jobs	20000 - 30000/annum 20-30K	25000	Cv-library.co.uk

Fig. 1. A snapshot of the dataset

Machine learning approaches may require the raw data sets to be divided into two parts, target attributes and non-target attributes.

The target attribute(s) is a special kind of attribute. The column of target attribute in the training dataset contains the previously recorded values used to train the model. The column of target attribute in the test dataset contains the previously recorded values to which the predictions are compared [13].

Non target attributes are the rest of the attributes present in dataset other than target one. Among these non-target attributes there are some attributes (it may be all non-target attributes also) which are used to predict target value, i.e. the attributes based on which we need to perform our predictive task are called features attributes. The target and non-target attributes defined for the proposed data set is given as below:

- Non target attributes

Id - A unique identifier for each job advertisement. It is a unique value to make each row distinguishable from other (like primary key of a dataset).

Title – It contains title of the job, it is actually the job designation.

Full Description - The full description of the job i.e. sub designation and responsibilities of the employee.

Location Raw - This is the actual job location given by the job advertiser.

Location normalized- It is the normalized location derived from location raw. It categorized the location into various group of location.

Contract Type - Contract type denotes whether the job contract is full-time or part-time.

Contract Time - Contract time denotes whether the job is given in contract basis or permanently.

Company - The name of the company or employer who is offering the job.

Category - Standard job categories like it, finance, sales etc.

Source Name - The name of the website or advertiser from whom the job advertisement is received.

- Target attributes

Salary Raw – it is the amount which will be paid for the job.

The next step for preprocessing of raw data set is data analysis.

C. Data analysis

Data analysis also called data visualization) refers to data collection, data cleaning, data aggregation and data reshaping with features exploration [14].

The pie chart given in Fig. 2 has been obtained from attribute ContractType using matplotlib [15]. From the chart we can see that 23.51% data are specified as full time, 3.23% data are specified as part-time but 73.26% data are not given, i.e. most of the job advertiser doesn't specify whether the contract is part-time or full-time which makes our dataset noisy though it is real world situation that most of the advertiser doesn't specify which type of job contract they are offering.

The pie chart given in Fig. 3 has been obtained from by plotting the attribute ContractTime using matplotlib. From the chart, we can see that 61.9% data are specified as permanent, 12% data are specified as contractual. However 26.1% data are not specified, which means that there's a lot of noise in our data.

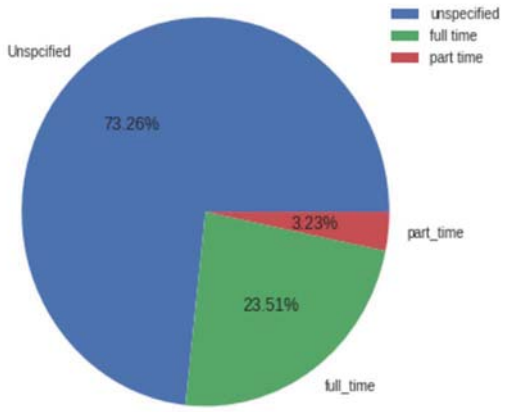

Fig. 2. A snapshot of the dataset

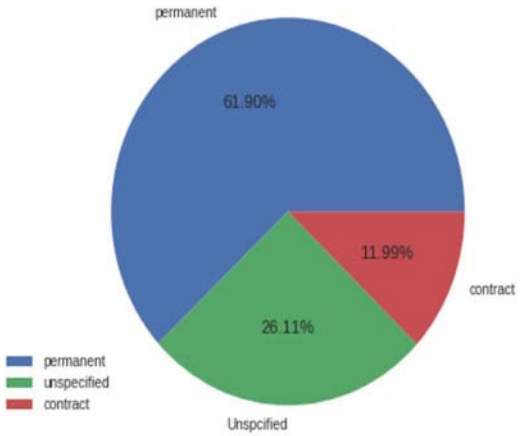

Fig. 3. Pie chart of ContractTime

We can see from the histogram of raw salary as given in the Fig. 4 that the salary distribution is much skewed to the left, meaning that there are a lot more low salaries on the low side than there are on the high side. This distribution of salaries leads to a standard deviation of 28193.095.

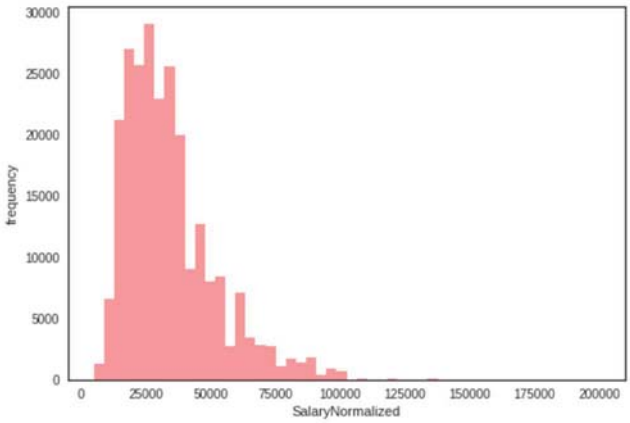

Fig. 4. Frequency distribution of salary normalized

Chart in Fig. 5 shows the correlation between each and every pairs of attributes using Pearson method of correlation. From this chart we can say that our predictive task is not dependent on contract type, contract time as the correlation value of these attributes are negligible. From the Fig 5 we can also conclude that the raw salary is highly dependent on id, title, full description, and moderately dependent on location raw and company.

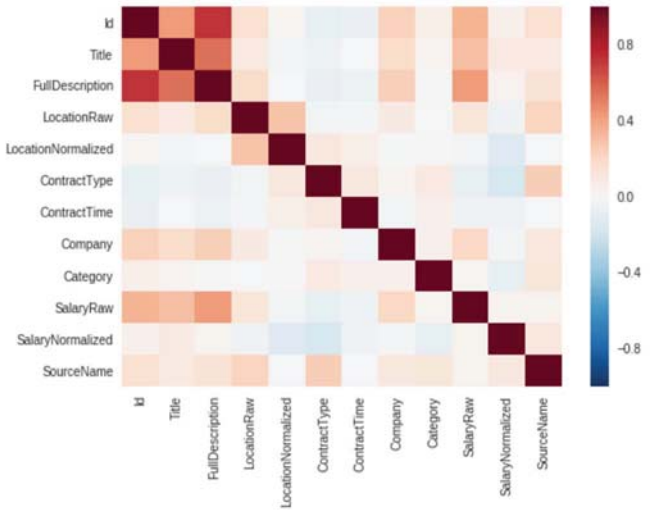

Fig. 5. Correlation chart between every pair of attributes using Pearson method of correlation

III. Data Processing

From the fig. 5 we can see that the correlation value of contract type and contract time is negligible and also from the given pie chart in previous section we can see that these two attribute mostly contains noisy data. So we can eliminate these two attribute from our dataset for simplification and noise reduction as this will not affect our predictive task.

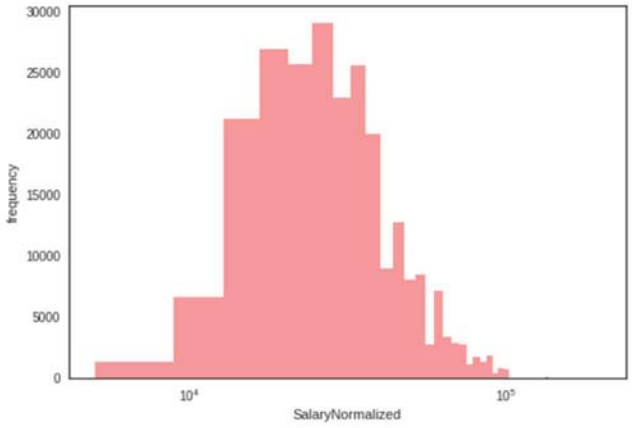

Fig. 6. Frequency distribution of log of salary normalized

In the histogram of raw salary of Fig. 4 we have seen that distribution is much skewed to the left, the possible solutions to this problem is taking the log of the salaries. This results into a more evenly distributed set of salaries. From the graph of Fig. 6 we can see that the distribution closely resembles a bell curve.

IV. THE PROPOSED PREDICTIVE ENGINE

After we have considered the problem as decision problem as we have to decide salary of a person based on certain condition and we have to consider all possible solution based on each attribute. The prediction engine starts the treatment with a decision tree classifier.

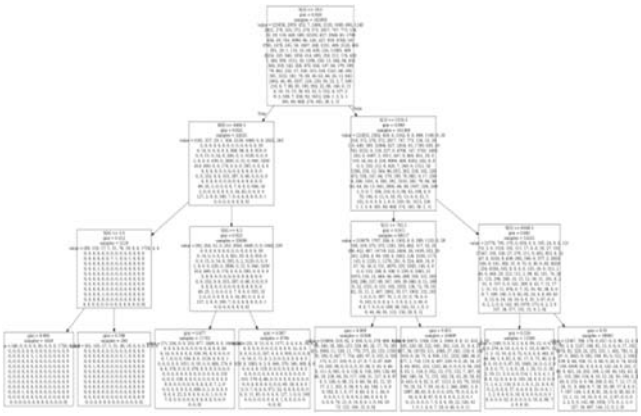

Fig. 7. Snapshot of Decision Treeof salary normalized

A. Decision tree classifier

Decision tree is a machine learning technology used for classification and prediction in a supervised way. Decision tree algorithm is a typical causative algorithm based on instance. It focuses on generating classification rules displayed as decision trees that is deduced or concluded from a group of disorder and irregular instances. The algorithm compares different features between internal nodes of decision tree in a top-down approach recursively. Generates the downward branches according to different attribute of the node, and calculates the prediction value. Further if the attribute is a new combination then store it for future use else draw a conclusion from leaf nodes in the decision tree [16]. In the task of the proposed job salary prediction, it is needed to consider all possible solution based on each attribute like at first we have to take decision about what should be the prime attribute to decide one's salary, then next prime attribute to decide further and so on. Thus in the proposed work decision tree classifier is used for the predictive task. In decision tree one can perform these jobs using CART (Classification and Regression Trees) algorithm by calculating GINI index and INFORMATION GAIN. Decision tree used to calculate information gain for all attribute and then whichever has the greater value will be the next decision node, i.e. the attribute which is more important to decide salary will be the next decision node and dataset will split based on that attribute.

But there were some problems with decision tree. The major drawback of Decision trees algorithm is they are sensitive to the specific dataset on which they are trained. If the training dataset has been changed the resulting decision tree may be quite different and also the accuracy of the prediction may be quite different. These problems may be addressed using Random forest classifier.

B. Random forest classifier

Random Forest Classifier is a type of ensemble machine learning algorithm. It generates multiple decision trees and then merges them together for more stable and accurate prediction. An ensemble method is a technique that creates multiple models using machine learning algorithms and then combines the predictions of each model together to make more accurate predictions than any individual model. The general procedure that can be used to reduce the variance for those algorithms that have high variance is Bootstrap Aggregation [17, 18].

To overcome the shortcomings of decision tree (Dataset sensitivity and high variance) we have introduced Random Forest Classifier algorithm. Here we can use Bootstrap Aggregation (Bagging) method for our purpose.

Steps of bagging method:

Step 1: Make multiple sub-samples (e.g.100) from the dataset randomly with replacement.

Step 2: Train each sample using CART method.

Step 3: Calculate the average prediction from each model.

Step 4: prediction is given based on the aggregation of predictions from n number of trees (here n is the no of sub-samples like 100).

V. RESULT

To improve performance further we've tried different models in our data set. We tried with random forest classifier, k nearest neighbor (27%), support vector machine(0%), and decision tree among them decision tree and random forest classifier are giving valuable accuracy score.[23, 25] The results are tabulated in Table 1.

Among decision tree classifier and random forest classifier algorithm, random forest classifier is giving more accuracy.

TABLE I. SCORES OF THE PREDICTION ENGINE FOR DIFFERENT MODELS

Method	Decision tree classifier	Random forest classifier
Accuracy score	0.844	0.873
Mean absolute error	6.04	5.04
Mean squared error	389.64	329.12
f1_score (macro)	0.615	0.656
f1_score (weighted)	0.844	0.869

VI. CONCLUSION

In this paper we have focused on the problem of predicting salary for job advertisements in which salary are not mentioned and also tried to help fresher to predict possible salary for different companies in different locations. The corner stone of this study is a dataset provided by ADZUNA [13]. All of the data is real, live data used in job. As our

dataset is a real world dataset we can say that our model is well capable to predict precise value but it is not sufficient. In future the work can be extended to design more powerful engine for more precise value. For this purpose we can collect more dataset and train our model with those. By train with different dataset we can make research about best fitting classifier and do more data analysis and data processing and can also go beyond the correlation value by features extraction. [19] We can also use this model for pickling and it's easy to deploy the model through REST api. [21, 22]

.

REFERENCES

[1] GEM Global Entrepreneurship Monitor. [Online]. Available: https://www.gemconsortium.org/report. [Accessed: 14-Oct-2018].

[2] "Changing Dynamics in the Retail Industry," Adobe Blog, 06-Dec-2017. [Online]. Available: https://theblog.adobe.com/changing-dynamics-in-the-retail-industry/. [Accessed: 14-Oct-2018].

[3] Bauman, "What is a prediction engine?," MindMeister, 20-Nov-2013. [Online]. Available: https://www.mindmeister.com/349147807/what-is-a-prediction-engine. [Accessed: 14-Oct-2018].

[4] Shapiro, A. F. (1998). "New Salary Functions for Pension Valuations", North American Actuarial Journal, 2(3), 26-27.

[5] Carriere, J. F., & Shand, K. J. (1998), "New salary functions for pension valuations", North American Actuarial Journal, 2(3), 18-26.

[6] Bone, C. M., & Mitchell, O. S. (1997), "Building better retirement income models", North American Actuarial Journal, 1(1), 1-10.

[7] Shiller, R. J. (2009). "The new financial order: Risk in the 21st century", Princeton University Press.

[8] Braskamp, L. A., Muffo, J. A., & Langston III, I. W. (1978), "Determining salary equity: policies, procedures, and problems", The Journal of Higher Education, 49(3), 231-246.

[9] Johnson, C. B., Riggs, M. L., & Downey, R. G. (1987),"Fun with numbers: Alternative models for predicting salary levels", Research in Higher Education, 27(4), 349-362.

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

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

[12] Job Salary Prediction | Kaggle. [Online]. Available: https://www.kaggle.com/c/job-salary-prediction/data. [Accessed: 14-Oct-2018]

[13] "Jobs in London, the UK & Beyond," Adzuna. [Online]. Available: https://www.adzuna.co.uk/. [Accessed: 14-Oct-2018]

[14] "Data Mining User's Guide," Moved, 15-May-2017. [Online]. Available: https://docs.oracle.com/database/121/DMPRG/GUID-2070E844-B79E-464A-954B-A7401D773C07.htm#DMPRG152. [Accessed: 14-Oct-2018]

[15] "Matplotlib: Python plotting - Matplotlib 3.0.0 documentation". [Online]. Available: https://matplotlib.org/. [Accessed: 14-Oct-2018]

[16] Dai, Q. Y., Zhang, C. P., & Wu, H. (2016), "Research of decision tree classification algorithm in data mining", International Journal of Database Theory and Application, 9(5), 1-8.

[17] Eesha Goel, Er. Abhilasha) Goel, E., & Abhilasha, E. (2017), "Random forest: a review", International Journal of Advanced Research in Computer Science and Software Engineering (ISSN: 2277 128X).

[18] Data Mining: Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016), "Data Mining: Practical machine learning tools and techniques", Morgan Kaufmann.

[19] "scikit-learn 0.20.0 documentation". [Online]. Available: http://scikit-learn.org/stable/modules/classes.html. [Accessed: 14-Oct-2018]

[20] Andreas C. Müller, S. (2018), "Introduction to Machine Learning with Python". [online] Shop.oreilly.com. Available at: http://shop.oreilly.com/product/0636920030515.do [Accessed 15 Oct. 2018]

[21] Learning, M. and Flask), T. (2018), "Tutorial to deploy Machine Learning model in Production as API with Flask". [online] Analytics Vidhya. Available at:

https://www.analyticsvidhya.com/blog/2017/09/machine-learning-models-as-apis-using-flask/ [Accessed 15 Oct. 2018].

[22] Medium. (2018), "A guide to deploying Machine/Deep Learning model(s) in Production". [online] Available at:

https://medium.com/@maheshkkumar/a-guide-to-deploying-machine-deep-learning-model-s-in-production-e497fd4b734a [Accessed 15 Oct. 2018].

[23] Colab.research.google.com. (2018). Google Colaboratory. [online] Available at:

https://colab.research.google.com/drive/1LlEsYsCtpoLQYr1zImVnweTbVeBPwWxg?authuser=1 [Accessed 15 Oct. 2018].

[24] Anon, (2018). [online] Available at: https://www.udemy.com/python-for-data-science-and-machine-learning-bootcamp/ [Accessed 15 Oct. 2018].

[25] Chrisalbon.com. (2018). Chris Albon. [online] Available at: https://chrisalbon.com/ [Accessed 15 Oct. 2018].