Predictive Analysis of HR Salary using Machine Learning Techniques

[1]Ritvik Voleti, [2] Bappaditya Jana
[1,2]Department of CSE, KCC Institute of Technology and Management, Gr Noida

Abstract:- **Information irregularity amongst employers and employees has become a problem that needs immediate solving. The probable applicants are most often kept blind with regards to the interview procedure and only are aware of it at the end. In the meantime, the employers must be committed to rightly meeting up with the candidate's prospects for making new HR strategies that satisfy the demands of the applicant. Therefore, one must be vigilant enough to not offer too low a salary, which would result in the decline in not just the salary but also will build more irresponsible, lack-luster individuals with longer untaken positions. Whilst the vice-versa would also be a cause of concern leading to wastage of companies' vital resources. Therefore, it is imperative to provide an unbiased salary for an employee which he/she truly deserves, and also has to be appropriate to the market demands. This paper is based on predicting the salary by training a Machine Learning model and performing comparative analysis on Logistic Regression and Support Vector Machine using their classification reports.**

Keywords— Logistic Regression, Support Vector Machine, Data Mining, Statistics, Machine Learning

I. INTRODUCTION

Machine Learning is emerging as an advanced technology that is supporting organizations within the domains of organizational aspects, people management, and business strategies [1]. There has been an evident escalation in Human Resources in recent times, as the market is experiencing a boom in quality resources and skills which therefore is a competitive incentive for enterprises to invest in [2]. Proper recruiting has become more crucial than ever and is an essential element in the enterprise's corporate plans due to its consequences on the businesses' competitiveness and efficiency. This paper put an emphasis on predicting the employee's new salary based on several factors such as age, work class, education, experience, previous occupation, past income, hours-per-week. Based on the above dataset parameters, salary can be predicted accurately. By training an ML model using Logistic Regression and Support Vector Algorithms one can receive highly efficient results.

II. PROCESS OF PREPARING MODEL

The process of predictive analysis begins with importing the necessary Python libraries like Numpy, Pandas, Sklearn, etc, and then performing data mining over the dataset which is to be trained thereafter by using Logistic Regression and SVM. Once it is trained, comparison using Classification Report is executed and if a user gives a new input the model has the capability to predict the salary either >50K or <=50K. These complex calculations are all due to the powers of computers, and libraries which enable not only collecting but also assisting in manipulating the imported dataset [3].

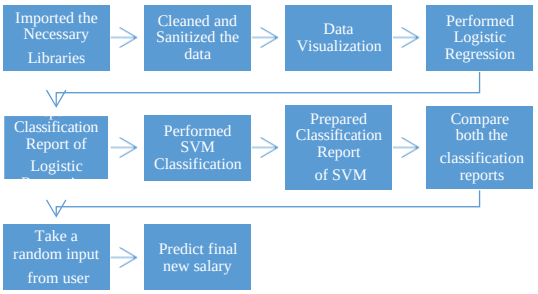


Figure 1: Process involved in performing predictive analytics for salary parameter.

A. Dataset Selection

Dataset selection/ importing always depends on the type of algorithm used, whether supervised or unsupervised or semi-supervised and also the number of valid records and attributes must be considered important selection criteria. In this paper, the supervised algorithm is chosen and a dataset consisting of about 50000 records and 15 attributes from Kaggle. This dataset was chosen, as the number of records, attributes are large enough to build an efficient model and the prediction has a great scope. It is in CSV format. The attributes referred to in the dataset are logical and competent to have a good idea in predicting the new salary efficiently and accurately.

B. Data Mining

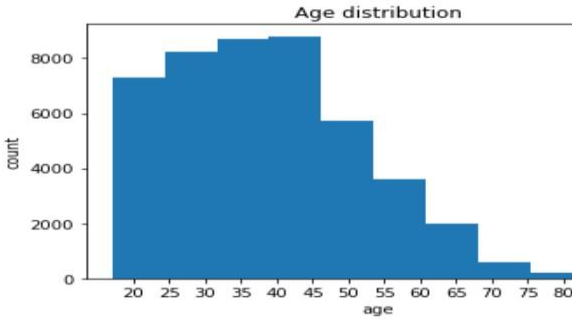
The data cleaning procedure involves creating visualist and analytic ways useful in multi-fields such as business intelligence, statistics, scientific visualization, and ML which are capable of handling multivariate, multidimensional data sets [4]. This technique executes the steps of identifying the data, extracting it, cleaning, and also integrating the dataset to be examined as needed [5]. Since there is never a possibility to obtain a perfect dataset that is noiseless, it is the programmer's responsibility to

remove as much noise, incompleteness, and other limitations for minimizing errors. There has to be a good enough model.

It is a challenge to enhance the overall accuracy of the ML model by aiming to remove all the unnecessary and meaningless features which are non-contributing towards the target variable. This paper is based on dropping all the null, missing, duplicate, insignificant values from the dataset through data cleaning. It not only cleans the dataset but also reduces the records and attributes. Selecting the correct feature is key in ML to affect the model's performance in a positive manner [6]. The accuracy can be bettered using the dataset's attributes for the training of the model. The variables must be removed which are no longer essential, or are just insignificant [7]. No data cleaning has a threat to harm the model's prediction in a negative manner. In preparing this model, there were 5 attributes removed and the remaining records were about 45000 which were efficient and sufficient enough to get an accurate machine learning model capable of predicting the salary efficiently.

C. Data Visualization

Scientific data visualization has seen rapid transformation in the last 45 years [8]. Visualization of data is the mechanism to represent the data in terms of graphs, and charts. It acts as the link between data and images. This is imperative as the patterns and trends become more observable due to data visualization. ML supports in conducting regular analyses like predictive analytics, that are valuable aid towards presenting visualization. Data visualization is a field independent tool which is assisting all the jobs in one way or another. It is an effective application to both showcasing importance of big data and also supporting in its depth analyses [9]. As the worth of data is all due to its meaning and visualization enhances the meaning of data so this tool is a necessity. The best representation of age parameter is via histogram which represents the flow of the frequency and also parameters like educational level are best suited to be shown via bar- chart with high accuracy. Interactive graphs using Plotly library we can examine three factors at a time in a fairly simplistic manner as shown below.



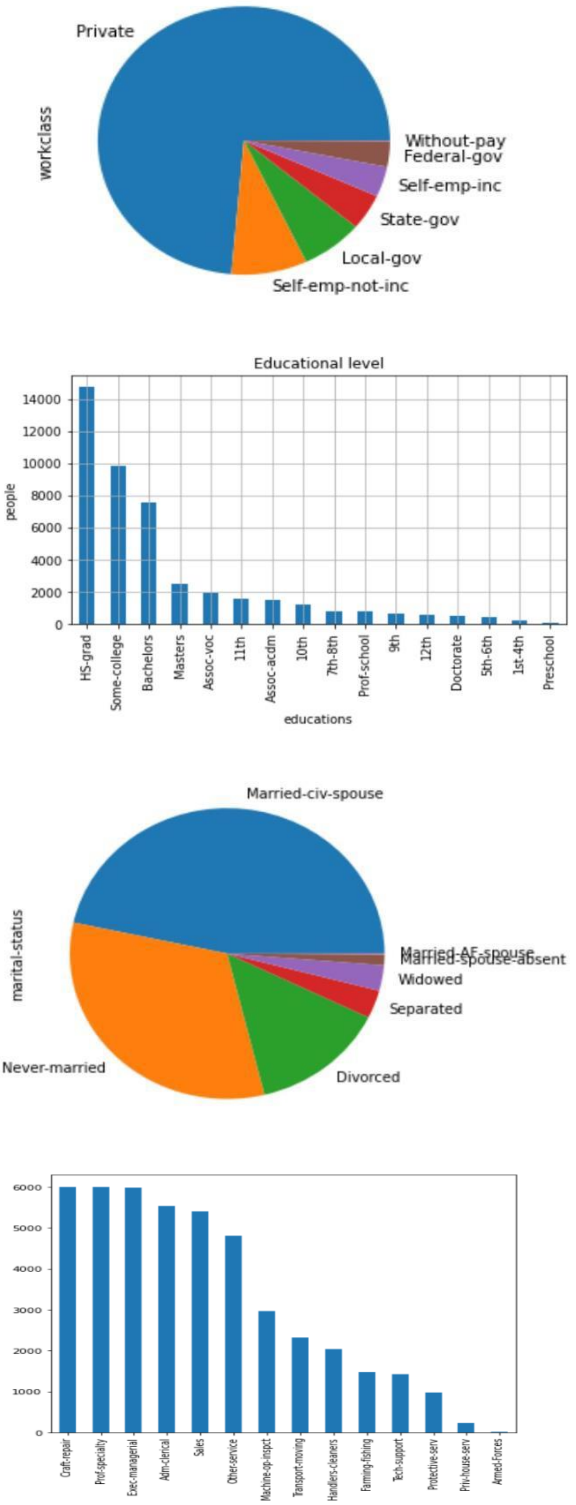


Figure 2: Graphical representations of different attributes in the dataset

D. Algorithm

The data has no significance without the supporting useful information extracted from it. Henceforth, predictive analytics is concerned with data analyzation to receive meaningful information [10]. All this is only possible via algorithms that support executing different tasks in the ML model. This

paper is driven using 2 Machine Learning algorithms i.e Logistic Regression and Support Vector Machine.

E. Importance of Performance Parameters

The classification model has to be evaluated to analyze real-world programs. Machine Learning classifying models are examined by the performance measures which assess how well the algorithms are performing within a given situation [11]. These performance measures comprise precision, recall score, accuracy, and F1 score. The model's performance criteria are now becoming a necessity in understanding the strengths and weaknesses of the model while performing new predictions in many cases.

The formula for the F1 Score is based on the weighted mean of Recall and Precision. Hence, this measure considers the false negatives and false positives into consideration. In situations of uneven distribution of class, it is a hurdle to properly interpret accuracy, but on the contrary F1 score acts as a more resourceful measure. Accuracy is better in cases of false negatives and false positives having a similar cost [12]. Otherwise in other scenarios the better measure is to look beyond F1- Score and into either Recall or Precision criteria.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

Figure 3. Formula of Performance Measures

F. Logistic Regression

There is a need for building classification models to solve various classification problems thereby finding the main target class within the data sample. Logistic regression is an important algorithm to facilitate organized data analysis [13]. It is part of a supervised machine learning algorithm [14]. Even though logistic regression, appears as a regression algorithm but in reality, it is a classification algorithm. This algorithm is based on the sigmoid function and has a conditional value of 0.5. If the input is above 0.5 it belongs to one class and the input value which lies below 0.5 belongs to another class. It is a type of data modeling procedure where the result is the final probability in the discrete outcome based on input values [15]. Binary logistic regression is the most common algorithm consisting of true or false, yes or no values, etc.

```
print(classification_report(ytest,pred))
```

	precision	recall	f1-score	support
<=50K	0.82	0.93	0.87	8539
>50K	0.62	0.38	0.47	2755
accuracy			0.79	11294
macro avg	0.72	0.65	0.67	11294
weighted avg	0.77	0.79	0.77	11294

Figure 4: Classification Report of Logistic Regression

G. Support Vector Machine

SVM (Support Vector Machine) is a type of supervised machine learning algorithm, that has applications in both regression and classification problems. But, the usage lies in the problems related to classification in ML. It has great potential in the ML research field [16]. The SVM algorithm's aim is to build the decision boundary which has the potential to find n-dimensional (space) into different classes so that we can assign the latest data point within the right futuristic category. The decision boundary is known as a hyperplane. Support Vector Machine selects the extreme vectors/points which assist in constructing the hyperplane. Extreme cases are known as SV (support vectors), so the algorithm is called Support Vector Machine. The Gaussian populations and data, lack in having sharp linear margins in real-life problems. So, the SV classifiers would not be able to compete, versus other techniques, in situations where supporting vectors are above the fixed number to support in finding the output classification weights and predicted rules [17].

```
print(classification_report(ytest,svm_pred))
```

	precision	recall	f1-score	support
<=50K	0.85	0.94	0.89	8539
>50K	0.72	0.47	0.57	2755
accuracy			0.83	11294
macro avg	0.79	0.71	0.73	11294
weighted avg	0.82	0.83	0.81	11294

Figure 5: Classification Report of Support Vector Machine

III. LIMITATIONS

- No. of records is never enough to perfectly classify and predict the salary with complete accuracy as the larger the data, the more accurate is the ML Model. The number of columns in the dataset is assumed to be accurate but is not sufficient like the records.
- The binary classification used does not have the same number of examples from each class like the class distribution is either skewed or imbalanced.
- The dataset predicted personal income levels being above or below 50,000 per year based on personal details such as marital status, education level, etc. There could have been many more cases of incomes less than 50K or above $50K, although the skew is not severe.
- Henceforth there are techniques for imbalanced classification that can be used whilst model performance can still be

reported using classification report, as is used with balanced classification problem.

IV. CONCLUSION

The main aim of this review paper focuses on finding the right future salary of an applicant based on some parameter of a particular domain. The algorithms used are Logistic Regression and Support Vector Machines to train and perform predictions using the ML model. Once both are imported the classification report is used as a comparison criteria to examine the overall efficiency of both algorithms. Out of the two, it is Support Vector Machine which is more accurate with an accuracy (F1-Score) of 89
% accuracy for salary<=50K and 57% for salary>50K. All this is only possible if proper data cleaning is done by removing all the missing, incorrect and noisy data from the dataset to get an efficient result. Hence, this model has the capability to act as an aid for HR to predict salary precisely quite conveniently.

REFERENCES

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

[2] Fallucchi, F.; Coladangelo, M.; Giuliano, R.; William De Luca, E. Predicting Employee Attrition Using Machine Learning Techniques. *Computers* 2020, *9*, 86. https://doi.org/10.3390/computers9040086

[3] Ritvik Voleti, Unfolding the Evolution of Machine Learning and its Expediency, International Journal of Computer Science and Mobile Computing, Vol. 10, Issue. 1, January – 2021, pg. 1-7, Doi: 10.47760/ijcsmc.2021.v10i01.001.

[4] Wong, Pak Chung, "Visual data mining." IEEE Computer Graphics and Applications 19.5 (1999): 20-21.

[5] Ritvik Voleti, 2020, Data Wrangling – A Goliath of Data Industry, International Journal of Engineering Research and Technology(IJERT) Volune09, Issue 08 (August 2020).

[6] Abdulhamit Subasi, Practical Machine Learning for Data Analysis Using Python, Elsevier, ISBN 978-0-12-821379-7, DOI https://doi.org/10.10.1016/C2019-0- 03019-1

[7] Sandro Sperandei, Understanding logistic regression analysis, Biochemia Medica Volune- 24, issue- 1, DOI: 10.11613/BM.2014.003.

[8] Christa Kelleher, Thorsten Wagener, Ten guidelines for effective data visualization in scientific publications, Enviornmental Modelling & Software, Volume 26, Issue6, 2011, Page 822-827, ISSN 1364-8152, https://doi.org/10/1016/j.envsoft.2010.12.0 06.

[9] D. Keim, H. Qu and K. -L. Ma, "Big-Data Visualization," in *IEEE Computer Graphics and Applications*, vol. 33, no. 4, pp. 20-21, July-Aug. 2013, doi: 10.1109/MCG.2013.54.

[10] Acharjya, Debi, and A. Anitha. "A comparative study of statistical and rough computing models in predictive data analysis." *International Journal of Ambient Computing and Intelligence (IJACI)* 8.2 (2017): 32-51.

[11] Thomas W. Edgar, David O. Manz, in Research Methods for Cyber Security, 2017, Elsevier, ISBN 978-0-12-805349-2

[12] Hoss Belyadi, Alireza Haghighat, in Machine Learning Guide for Oil and Gas Using Python, 2021, Elsevier, ISBN 978-0-12-821929-4, DOI https://doi.org/10.10.1016/C2019-0- 03617-5

[13] Scott Menard, Applied Logistic Regression Analysis Quantitative Applications in the Social Sciences, Sage Publication, ISBN 1544332580978154455258l.

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

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

[16] M. Pal & P. M. Mather Support vector machines for classification in remote sensing, International Journal of Remote Sensing, Volume – 27 issue- 5, 2005 https://doi.org/10.1080/014311605123313 14083.

[17] S Raudys, How good are support vector machines?, Neural Networks, Volume 13, Issue 1, 2000, Pges 17-19, ISSN 0893-6080, https://doi.org/10.1016/S0893- 6080(99)00097-0.