Determination of the most effective machine learning method for predicting performance of employees

Dr.N.Satheesh[1],Dr.G.Jawaharlel Nehru[2],Dr.B.Rajalingam[3], Dr.M.Narayanan[4],Avinash seekoli[5],

[1,2,3,4]Associate Professor
[5]Assistant Professor
Department of computer science and engineering,St.Martins engineering college,

Abstract

In today's society, being able to predict one's pay is critical. Nowadays, a person's worth is determined by the amount of money he or she earns. The greater the pay, the greater the individual's ability to be a wonderful person while still enjoying a wonderful lifestyle. This document discusses how a corporation predicts its employees' salaries, including all of the important factors such as the jobs held by employees, their years of experience, their income and level, their age, their expected salary, and how much time they have taken off. Here's where we put forward a candidate. Machine learning algorithms are being used to predict salaries. Not only is it necessary to anticipate the output, but it is also necessary to determine which method is the best one among all of the supervised and unsuprivised algorithms and to demonstrate the comparison between them by plotting the numbers.

Keywords:Prediction,supervised,machine learning algorithms,ploting

1. INTRODUCTION

Machine Learning is the method for analyzing the data/datasets with the help of models. It is a part/subset of Artificial Intelligence which works mainly on the algorithms who predict the outputs and accuracy on the basis of past data or from past experiences. This paper is fulfilled by ML programming with the help of python language.addition features like geometric shapes, ellipsis are added. addition features like attendance "An heuristic codebook was proposed of good generalization and discriminative properties, enabling multipath interferences mechanisms on propagation of1 conditional livelihood"[6] .Here pareto distributions are used for parametrics" Pareto distribution model is implemented to improve the reliability prediction rate"[3].This paper is about the salary prediction of the employees of a certain company with the key elements such as: the positions of the employees, year of experience, salary, level, age of the employee, estimated salary, and leave taken by each employee. This paper is executed with ten different algorithms of Machine Learning. Among them some are based on Regression algorithms and rest are based on Classification algorithms. In this paper the maximum scored accuracy is from the regression part with 97% and the least accuracy is scored from the classification part with 85%. This paper also contains an algorithm with an overfitted accuracy which shows that this paper is fulfilled with different forms of accuracy structures. The paper also has certain visualized materials which makes the predicted values more accurate and clearer to the picture. At the end, this paper is completed with a bar graph which shows the differences among the accuracies predicted by different algorithms.

Banking, bioinformatics, business, computer vision, and education are just a few of the industries that are interested in data mining and learning from it in general. It is the purpose of

data mining to extract usable information from acquired data, which is accomplished via the application of a broad variety of machine learning algorithms. The topic of which learning algorithm is best suited for a particular dataset is one that is frequently asked. The data may be used to run tests with different algorithms to determine which algorithms are the best candidates. Getting input from machine learning specialists can also be beneficial in determining which algorithms are the best prospects.

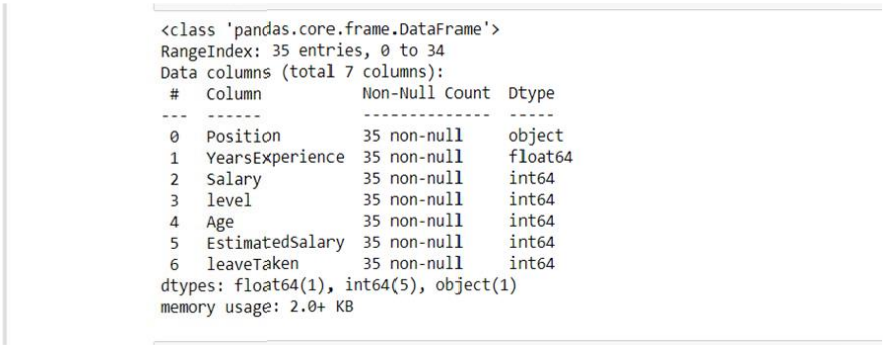
Using the concept of meta-learning, we are able to address the problem of picking the best machine learning method for a certain dataset by using supervised learning to solve it. Specifically, we demonstrate an effective method of dealing with a non-standard format in a real-world dataset in order to collect training data. Our study goes farther in addressing the problem of well-known algorithm selection, which has lately been brought to our attention.

As it is known that Machine learning learns from the past experiences and it is the method for analyzing the datasets with the help of models. So, in this paper also a dataset is being read.. This dataset contains 35 rows and 7 columns in total. The main elements of the dataset are the positions of the employees, year of experience, salary, level, age of the employee, estimated salary, and leave taken by each employee. The position of the employees contains different names and positions of the employees from director to the least graded position which is the clerical staff. The salary also differs in the same way as of the highest salary is paid to the director (Rs.139465) then CEO (Rs.135675), then Chief Operating Officer then it goes on to the least paid position of clerical staff (Rs. 37731). The age and year difference also matters here a lot with the senior most is the director and the youngest is the health care official of 18years who works as an intern. The programming-based information of the dataset is given below:

s.no	Position	Years Exp	Salary	level	Age	Estimated Salary	leaveTaken
0	staff	1.1	39343	1	19	19000	0
1	Clerical staff -2	1.3	46205	2	35	20000	0
2	Clerical staff-1	1.5	37731	3	26	43000	0
3	Clerk -4	2	43525	4	27	57000	0
4	Clerk -3	2.2	39891	5	19	76000	0
5	clerk - 2	2.9	56642	6	27	58000	0
6	Clerk -1	3	60150	7	27	84000	0
7	IT personal -8	3.2	54445	8	32	150000	1
8	IT personal-7	3.2	64445	9	25	33000	0
9	IT personal-5	3.7	57189	10	35	65000	0
10	IT personal-3	3.9	63218	11	26	80000	0
11	IT personal-2	4	55794	12	26	52000	0
12	IT personal-1	4	56957	13	20	86000	0
13	Web designer	4.1	57081	14	32	18000	0
14	Health Care official	4.5	61111	15	18	82000	0
15	Advisor	4.9	67938	16	29	80000	0
16	Management Consultant	5.1	66029	17	47	25000	1

17	Public Relation Officer	5.3	83088	18	45	26000	1
18	Media Officer	5.9	81363	19	46	28000	1
19	C-Service Officer	6	93940	20	48	29000	1
20	C- information Ofiicer	6.8	91738	21	45	22000	1
21	C-gaming officer	7.1	98273	22	47	49000	1
22	C- design officer	7.9	101302	23	48	41000	1
23	C- data officer	8.2	113812	24	45	22000	1
24	C- compliance Officer	8.7	109431	25	46	23000	1
25	C- brand officer	9	105582	26	47	20000	1
26	C- Business Development officer	9.5	116969	27	49	28000	1
27	C- commercial officer	9.6	112635	28	47	30000	1
28	C-legal Ofiicer	10.3	122391	29	29	43000	0
29	C-Marketing Officer	10.5	121872	30	31	18000	0
30	C-Technology Officer	11.2	127345	31	31	74000	0
31	C-Financial Officer	11.5	126756	32	27	137000	1
32	C-Operating Officer	12.3	128765	33	21	16000	0
33	CEO	12.9	135675	34	41	44000	0
34	Director	13.5	139465	35	55	90000	0

Table-1:Dataset of employes



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35 entries, 0 to 34
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Position         35 non-null     object
 1   YearsExperience  35 non-null     float64
 2   Salary           35 non-null     int64
 3   level            35 non-null     int64
 4   Age              35 non-null     int64
 5   EstimatedSalary  35 non-null     int64
 6   leaveTaken       35 non-null     int64
dtypes: float64(1), int64(5), object(1)
memory usage: 2.0+ KB
```

Fig-1:Data set information from python programming

2. Evaluation of Algorithms

This paper based on different algorithms, namely:
Simple Linear Regression
Polynomial Regression
Multi-Linear Regression
Logistic Regression
KNN (KNearer Neighbour)
SVM (Support Vector Machine)
Decision Tree Classification
Decision Tree Regression
Random Forest
KMeans Cluster

Simple Linear Regression: Simple Linear Regression is a type of regression that has only one input and one output, and it has a straight-line graph as its representation. This dataset has an accuracy of 97 percent, and its input data is the employee's year of experience, while its output data is his orher pay paid.

Polynomial Regression: It is a sort of regression that provides the accuracy of a single observation when several observations are used. Non-linear regression is the linked variant of simple linear regression that has a curved graph, which is why it is also referred to as non-linear regression.

(Eg: Suppose a person came to give an interview in a company and he said that his previous company paid him the salary of Rs.60,000 per month. So, to judge that he is telling a truth or bluff the new joining company will recheck the salary statement.) In this type of works polynomial regression is used. It also gives better and much accurate accuracy than simple linear regression.

Multi-linear :Multivariate regression is a type of regression in which there are numerous inputs and a single output. This regression does not produce a graph since it contains several types of independent variables as input and a single dependent output, resulting in a graph that could be 3D, 4D, or more complex than that. In this dataset, the input data is the employee's year of

experience and wage, whereas the output data is the employee's level, with an overfitted accuracy of 154 percent in this dataset.

Logistic Regression Logistic regression is a kind of binary classification that uses an S-shaped graph as its visual representation. It is based on the sigmoid function and provides a more accurate value than a standard linear regression since the S-shaped graph encompasses all of the points in comparison to a linear regression. The input data in this case is the year of experience, the salary, the level, the age, and the predicted salary, and the output data is the amount of time taken off. It had an accuracy rating of 85 percent.

KNN classification is a type of classification in which the data is separated into classes and the total number of clusters is determined with the help of the'k' value. The inputs and outputs are identical to those of logistic regression, and the accuracy of training (78 percent) and testing (85 percent) are both high.

SVM In this research, SVM is utilised for classification, and it follows a multi-class classification approach. The optimal line is the one that has the greatest distance between two points, and the two points from which the greatest distance is measured are referred to as the support vectors. The accuracy of the training is 96 percent, while the accuracy of the testing is 85 percent in this case.

Decision Tree The CART method is used to construct the Decision Tree, which implies it follows both regression and classification rules. An ordinary decision tree asks a question and then categorises the respondent according to the answer. The root node, leaf node, and internal node are the three most important aspects of the tree in this case. Gini Impurity, which is also known as entropy, is used to locate the root node in a graph. The age and estimated income are provided as input data, and the leave taken is provided as output data. When it comes to accuracy, the classification section here is 92 percent accurate. The level of the employee is the output data of the regression component, and the input data for the regression part is the anticipated salary.

Random Forest Similarly to the CART Algorithm, Random Forest is a combination of multiple numbers of Decision Trees that is based on the CART Algorithm. It examines the vast majority of cases before announcing the ultimate outcome. The Gini impurity approach is used to identify the root node in this case as well. The accuracy of the categorised component in this instance is 92 percent.

KMeans Cluster In this paper, KMeans Cluster is an unsupervised learning technique that is used to identify the output/dependent variable, which is called the output variable. The number 'k' represents the total number of random points/centroids in this example. This paper makes use of five different clusters, each with its own set of random points, which are represented graphically by a scatterplot.

3. Proposed work:

The algorithm suggested in this paper is the most accurate algorithm currently available. Consequently, the Simple Linear Regression is proven to have the highest accuracy accuracy, with a precision of 97 percent, according to this article. The simple linear regression model has only one input and one output, where the input is the number of years of experience and the output is the amount of money that was paid. Moreover, a straight-line graph is used to depict the correctness of this technique, which is displayed in the most picturesque manner possible. This method is also compared to other algorithms because it exhibits the highest accuracy among all other algorithms, as seen by the visualisations presented below.

Fig-2:Simple linear regression graph

4. Simple linear regression

This **Simple Linear Regression** algorithm is done and followed by different steps, such as:

a).Firstly, all the important libraries are imported and the dataset is read with the help of **Pandas** library.

b). Secondly, the slicing is done to identify the input and output data.

c).Thirdly, the model is being created and the **test_train_split()** function is introduced to divide the model into training and testing parts. Here the training is of 99.66% and that of the testing is 0.33%

d).Fourthly, the **StandardScaler** is introduced to standardized the dataset by round off the values.

e).Fifthly, the dataset is transformed to do training by **.fit** function.

f).Lastly, the intercept(c), coefficient(x) and the accuracy are found and a graph is being executed for visualization.

Modelling:

The modelling is done with the sliced dataset having the input data as years experienced and the output data as salary paid to each employee. The **test_train_split()** function is called for splitting the model into training and testing parts. The training part done is of **99.66%** and rest of the **0.33%** of the model is send for testing.

Evaluation:

Generally, the models are being evaluated for getting the accuracy by **cost function** and **gradient descent**. Here the **cost function/lost function** is generated in ML programming to verify the actual data and the predicted data is same or not by checking the difference between actual output and predicted output. Naturally, the cost function must be always low then only the accuracy will be high. The **gradient descent** algorithm is used to change the values of the slope of the line and the coefficient(x). Here, in this paper also the same thing is being done by **.predict()** function and its been evaluated to find the accuracy.

Accuracy:

Here in this paper the best accuracy calculated and seen among all the algorithms is of the simple linear regression with an accuracy of **97%**. The accuracy is calculated by using the **r2_score()** function from the **sklearn.metrics** package.

5. Result and Analysis:

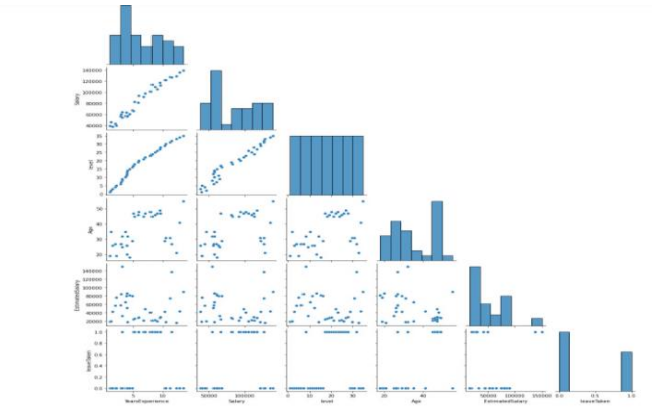


Fig-3:Pairplot for checking the similarties

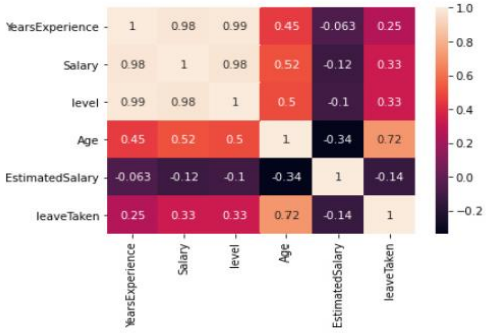


Fig 4:Heat map to identify correlation

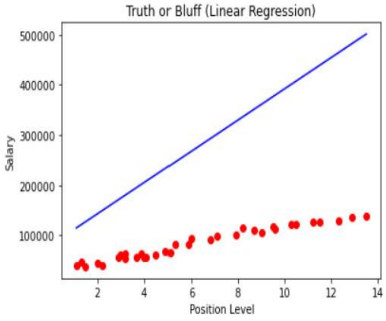


Fig 5:Simple linear regression

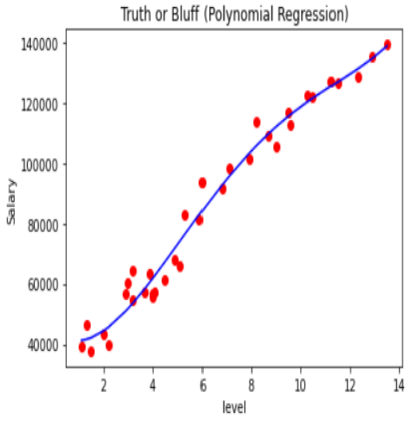


Fig 6:Non linear regression

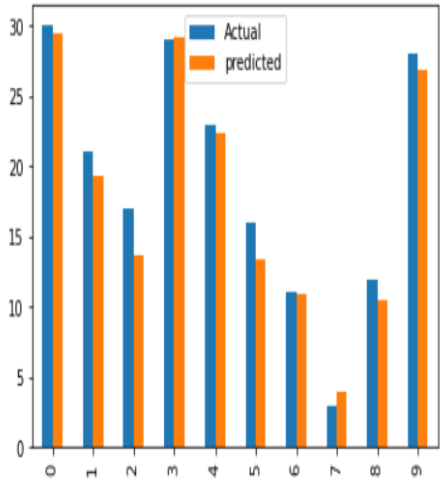


Fig 7:Actual and predicted value graph of multiple regression

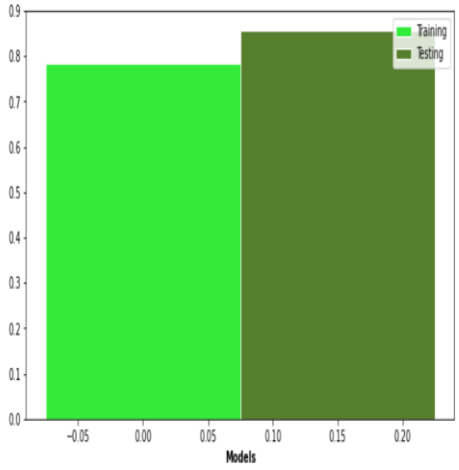


Fig 8:Logistic regression

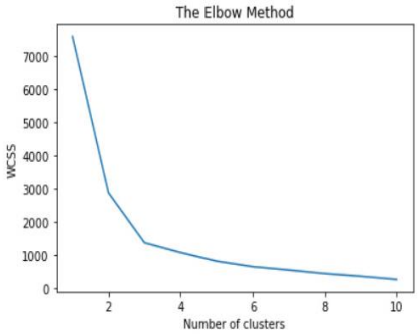


Fig 9:Elbow method

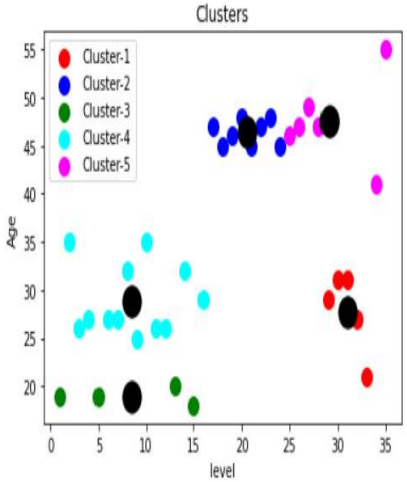


Fig 10:K means clustering

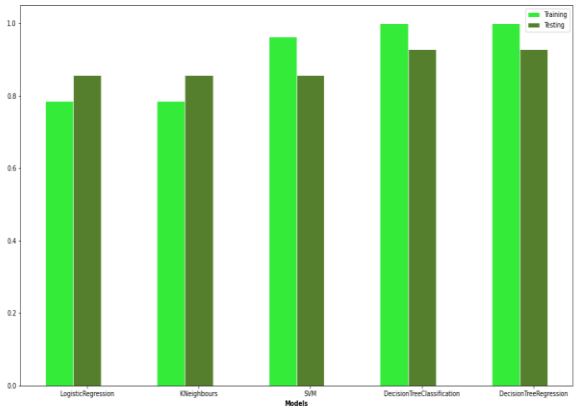


Fig 11:Bar graph showing the b difference between the accuracies of different classification algorithm

6. Conclusion

In the paper's conclusion, the authors state that we learned how to use various methods. Also demonstrated in this paper is the practical application of machine learning in today's world, and as the definition of Machine Learning states, it learns from past experiences, this is also demonstrated in this paper. When all accuracies are calculated, the basic linear regression is the most accurate for the data set when compared to all other methods. The best is represented in terms of overfitting and underfitting, as well as the cost function. After obtaining a more detailed understanding of the cost function, I modified the weights for the gradient descent process. But, at the end of the day, the outcome was beyond our expectations, as evidenced by this study.

7. Future enhancement

As Machine learning is playing now-a-days a vital role in this computer-generated world. So, from the errors/hurdles we faced and rectified a lot of problems. But in this paper the best accuracy we got in simple linear regression and we also got some overfitted accuracy in multi-linear regression. So, to get most of the accuracies up to the mark we should change some dataset values or standardized the values to make it perfect. We also learnt and suggest to do the slicing of the dataset to get the input and output data by checking and modeling the dataset in a perfect way to get the best accuracy in most of the algorithms.

8. Acknowlodgement

This dataset of our paper ***"Salary prediction"***is being collected from ***www.kaggle.com*****,** which is one the subparts of Google consisting of different works and opinions of many data scientists round the world. Kaggle is a website which provides a customized environment of **Jupyter** notebook. Here we can get different types of datasets and can also publish our works/documents/thesis. It is a free platform for all the programmers from beginners to the advanced ones

.

9. References

[1].MohammadAli,SwakkharShatabda(2020),"A Data Selection Methodology to Train Linear Regression Model to Predict Bitcoin Price",*International Conference on Advanced Information and Communication Technology (ICAICT),*IEEE.

[2].Lun Li et.al,"Bitcoin Options Pricing Using LSTM-Based Prediction Model and Blockchain Statistics",*IEEE International Conference on Blockchain (Blockchain)*.

***[3]Avinash seekoli,Dr.Y.Srinivas(2020),"A NEW METHODOLOGY FOR SOFTWARE RELIABILITY BASED ON PARETO DISTRIBUTIONS",*Solidstate technology,vol-63,issue-6,pp-10062-10067**

[4].Avinash seekoli,Y srinivas(2019),"A Methodology for Identification of Failures in Software Development Process",International Journal of Engineering and Advanced Technology (IJEAT),vol-9,issue-1,pp-1066-1069

[5].AkkalaAbhilasha,Avinashseekoli(2020),"*Multi-service ingress through AWS comprehend employ Identity and access management system",*Journal of Huazhong University of Science and Technology,Vol-49,pp:1-5.

[6].AkkalaAbhilasha,Avinashseekoli(2020),"Identify Occurrence of Substance Object of a Certain Classification in Fractional Videos and Pictures",International Journal of Engineering and Advanced Technology (IJEAT),vol-9,issue-3,pp:3728-3731.