EMPLOYEE SALARY PREDICTION SYSTEM USING MACHINE LEARNING

V.SARALA [1], K.GANESH[2]

[1] **Assistant Professor MCA, DEPT,** Dantuluri Narayana Raju College, **Bhimavaram, Andhra Pradesh**
Email id:- vedalasarala21@gmail.com
[2]**PG Student of MSc Computer Science,** Dantuluri Narayana Raju College, **Bhimavaram, Andhra Pradesh**
Email id :- abganesh7@gmail.com

ABSTRACT

Machine learning is a technology which allows a software program to became more accurate at pretending more accurate results without being explicitly programmed and also ML algorithms uses historic data to predicts the new outputs. Because of this ML gets a distinguish attention. Now a day's prediction engine has become so popular that they are generating accurate and affordable predictions just like a human, and being using industry to solve many of the problems. Predicting justified salary for employee is always being a challenging job for an employer. In this project, a salary prediction model is made with suitable algorithm using key features required to predict the salary of employee. The main aim of the project is to predict the salary of graduates and make a suitable user-friendly graph. From this prediction the salary of an employee can be observed according to a particular field according to their qualifications. It helps to see the growth of any field. In the project, we have used Linear Regression as an algorithm for prediction. Linear regression carries out a task that may predict the value of a dependent variable (y) on basis of an independent variable (x) that is given. Therefore, this kind of regression technique looks for a linear type of relationship between input x and output y. Apart from Linear Regression, other types of regression techniques are also used like the Decision Tree Regressor and Random Forest Regressor. Since nothing in this universe can be termed as "perfect", thus a lot of features can be added to make the system more widely acceptable and more user friendly. This will not only help to predict salaries of other fields but also will be more user beneficial. In the upcoming phase of our project we will be able to connect an even larger dataset to this model so that the training can be even better. This model should check for new data, once in a month, and incorporate them to expand the dataset and produce better results.

1 INTRODUCTION

Nowadays, one of the major reasons an employee switches a company is the salary of the employee. Employees keep switching the company to get the expected salary. And it results in loss for the company and to overcome this loss we came with an idea what if the employee gets the desired/expected salary from the Company or Organization. In this Competitive world everyone has a higher expectation and goals. But we cannot randomly provide everyone their expected salary there should be a system which should measure the ability of the Employee for the Expected salary. We cannot decide the exact salary but we can predict it by using certain data sets. A prediction is an assumption about a future event. A prediction is sometimes, though not always, is based upon knowledge or experience. Future events are not necessarily certain, thus confirmed exact data about the future is in many cases are impossible, a prediction may be useful to help in preparing plans about probable developments. In this project, the salary of an employee of an organization is to be predicted on basis of past experience and the educational qualifications of the individual. Here the history of salary has been observed and then on basis of that salary of a person after a certain period of time it can be calculated automatically. In order to gain useful insights into the job recruitment, we compare different strategies and machine learning models. The methodology different phases like: Data collection, Data cleaning, Manual feature engineering, Data set description, Automatic feature selection, Model selection, Model training and validation, Model comparison. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly. Machine learning algorithms are broadly classified into three divisions, namely;

Supervised learning, Unsupervised learning and Reinforcement learning.

• Supervised learning:- Supervised learning is a learning in which we teach or train the machine using data which is well labelled that means some data is already tagged with correct answer. After that, machine is provided with new set of examples so that supervised learning algorithm analyses the training data and produces a correct outcome from labelled data. Basically, they can apply what has been learned in the past to new data using labelled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

• Unsupervised Learning:- In contrast, unsupervised machine learning algorithms are used when the information used to train is neither classified nor labelled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabelled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabelled data. Unsupervised learning is the training of machine using information that is neither classified nor labelled and allowing the algorithm to act on that information without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data. Unlike, supervised learning, no teacher is provided that means no training will be given to the machine. Therefore, machine is restricted to find the hidden structure in unlabelled data by our-self.

• Reinforcement learning:- Reinforcement learning is an area of Machine Learning. Reinforcement. It is about taking suitable action to maximize reward in a particular situation. It is employed by various software and machines to find the best possible behaviour or path it should take in a specific situation. Reinforcement learning differs from the supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of training dataset, it is bound to learn from its experience. Reinforcement machine learning algorithms is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal. The project uses various regression techniques for predicting the salary of the employees. The techniques are listed as follows.

1. Linear Regression: In Linear regression we are given a number of predictor variables and a continuous response variable, and we try to find a relationship between those variables that allows us to predict a continuous outcome.

2. For example, given X and Y, we fit a straight line that minimize the distance using methods to estimate the coefficients like Ordinary Least Squares and Gradient Descent between the sample points and the fitted line.

3. Decision Tree Regressor: Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches, each representing values for the attribute tested. Leaf node represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

4. Random Forest Regressor: Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time For regression tasks, the

mean or average prediction of the individual trees is returned.

2. LITERATURE SURVEY AND RELATED WORK

1) Susmita Ray," A Quick Review of Machine Learning Algorithms," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-ITCon), India, 14th -16th Feb 2019 a brief review of various machine learning algorithms which are most frequently used to solve classification, regression and clustering problems. The advantages, disadvantages of these algorithms have been discussed along with comparison of different algorithms (wherever possible) in terms of performance, learning rate etc. Along with that, examples of practical applications of these algorithms have been discussed.

2) Sananda Dutta, Airiddha Halder, Kousik Dasgupta," Design of a novel Prediction Engine for predicting suitable salary for a job" 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN) - 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. model is well capable to predict precise value.

3) Pornthep Khongchai, Pokpong Songmuang, "Improving Students' Motivation to Study using Salary Prediction System" - proposed prediction model using Decision tree technique with seven features. Moreover, the result of the system is not only a predicted salary, but also the 3-highest salary of the graduated students which share common attributes to the users. To test the system's efficiency, they set up an experiment by using 13,541 records of actual graduated student data. The total result in accuracy is 41.39%.

4) Phuwadol Viroonluecha, Thongchai Kaewkiriya," Salary Predictor System for Thailand Labour Workforce using Deep Learning" - used Deep learning techniques to construct a model which predicts the monthly salary of job seekers in Thailand solving a regression problem which is a numerical outcome is effective. We used five-month personal profile data from wellknown job search website for the analysis. As a result, Deep learning model has strong performance whether accuracy or process time by RMSE 0.774 x 104 and only 17 seconds for runtime.

3 EXISTING SYSTEM

Easily identifies trends and patterns: Machine Learning Models can review large volumes of data and discover specific trends and patterns that would not be apparent to humans. For instance, for an e-commerce website like Amazon, it serves to understand the browsing behaviours and purchase histories of its users to help cater to the right products, deals, and reminders relevant to them. It uses the results to reveal relevant advertisements to them.

No human intervention needed (automation): With implementation of ML model, there is no need to have any eye on the project at every step of the way. Since, giving machines the ability to learn, lets them make predictions and also improve the algorithms on their own. A common example of this is anti-virus softwares; they learn to filter new threats as they are recognized. ML is also good at recognizing spam.

Continuous Improvement: As ML algorithms gain experience, they keep improving in accuracy and efficiency. This lets them make better decisions.

.

4 PROPOSED WORK AND ALGORITHM

An Architectural Diagram or a pipeline is used to help automate machine learning workflows. They operate by enabling a sequence of data to be transformed and correlated together in a model that can be tested and evaluated to achieve an outcome, whether positive or negative. The pipeline/ Diagram consists of several steps to train a model. Machine learning pipelines are iterative as every step is repeated to continuously improve the accuracy of the model and achieve a successful

algorithm. To build better machine learning models, and get the most value from them, accessible, scalable and durable storage solutions are imperative, paving the way for onpremises object storage. The steps include:

• Data Collection: Collecting raw data from billions of datasets available.

• Data Exploration: Exploring the data & the features related and being familiar with the data-types

• Data Manipulation: Includes Cleaning of data, treating missing, repetitive values that are present.

• Data Analysis: Analysing the data to increase efficiency while applying the best Algorithm & feature selection according to our preferences.

• Application of Algorithm: Applying the algorithm to the model. • Evaluation: Using evaluation metrices to calculate the least error and following the above to make further changes.

5 METHODOLOGIES

MODULES

List of algorithm used in implementation of our experiment are:

Support Vector Machine

K-nearest algorithm

Artificial Neural Network

Random forest algorithm 5.Logistic Regression algorithm

6 RESULTS AND DISCUSSION

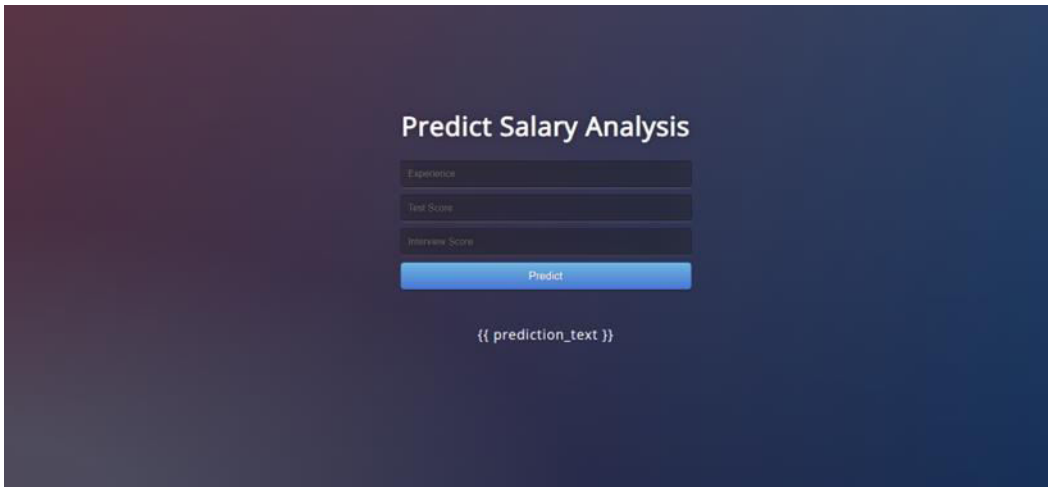

Fig 1:- Homescreen

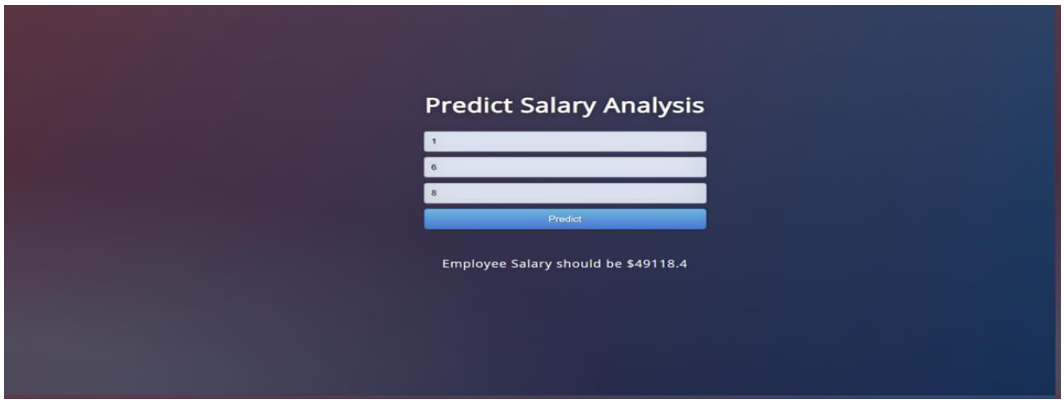

Fig 2:- Predict Analysis for salary

6.CONCLUSION AND FUTURE SCOPE

In today's real world, it has become tough to store such huge data and extract them for one's own requirement. Also, the extracted data should be useful. The system makes optimal use of the Linear Regression Algorithm. The system makes use of such data in the most efficient way. The linear regression algorithm helps to fulfil customers by increasing the accuracy of estate choice and reducing the risk of investing in an estate.

Our Model Predicted An Accuracy score of 95.68% on the training dataset while it predicted an Accuracy score of 95.33% on the testing dataset. Since there is a very minute difference between the training and testing scores, we can say that our model has performed extremely well on the given dataset, that with such a high % score. It is illustrated that the approach contributes positively according to the evaluation.

FUTURE SCOPE :

Since nothing in this universe can be termed as "perfect", thus a lot of features can be added to make the system more widely acceptable and more user friendly. This will not only help to predict rates of other aras in the city but also will be more user beneficial.

In the upcoming phase of our project we will be able to connect an even larger dataset to this model so that the training can be even better. This model should check for new data, once in a month, and incorporate them to expand the dataset and produce better results

We can try out other dimensionality reduction techniques like Uni-variate Feature Selection and Recursive feature elimination in the initial stages.

Another major future scope that can be added is providing the model with estate database of more cities which will provide the user to explore more graduates and reach an accurate decision. More factors like training period that affect the job salary of a graduate shall be added. In-depth details of every individual will be added to provide ample details of a desired estate. This will help the system to run on a larger level.

7 REFERENCES

1.https://expertsystem.com/machine-learning-definition/#:~:
text=Machine%20learning%20is%20an%20application,use%20it%20learn%20for%20th emselves.

2.Fisher, R.A. (1922). "The goodness of fit of regression formulae, and the distribution of regression coefficients". Journal of the Royal Statistical Society. 85 (4): 597–612.

3.Schmidhuber, J. (2015). "Deep Learning in Neural Networks: An Overview". Neural Networks. 61: 85–117.

4.Yan, Xin (2009), Linear Regression Analysis: Theory and Computing, World Scientific, pp. 1–2.

5.Vapnik, V. N. The Nature of Statistical Learning Theory (2nd Ed.), Springer Verlag, 2000.

6.https://www.python.org/doc/essays/blurb/#:~:text=Python%20is%20an%20interpreted%

2C%20object,programming%20language%20with%20dynamic%20semantics.&text=Pyt
hon's%20simple%2C%20easy%20to%20learn,program%20modularity%20and%20code
%20reuse.
7.https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day- life-112a47a429d0
8.https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data
9.http://ijcsma.com/publications/march2019/V7I302.pdf
10. https://towardsdatascience.com/predicting-house-prices-with-linear-regression-machine- learning-from-scratch-part-ii-47a0238aeac1
11. https://flask.palletsprojects.com/en/1.1.x/