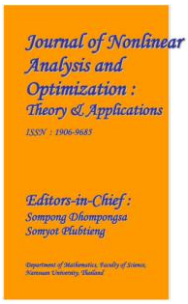
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 2, No.1 : 2024
ISSN : **1906-9685**



SALARY FORECASTING USING ML ALGORITHM WITH HIGH ACCURACY AND PERFORMING DATA ANALYSIS WITH SEABORN LIBRARY

CHRISTO JOSHY Student, III Year (Digital Cyber Forensic Science) Rathinam College of Arts and Science, Coimbatore-21
Dr. T. VELUMANI Assistant Professor Department of Information Technology Rathinam College of Arts and Science, Coimbatore–21

Introduction:

The project aims to predict salaries with high accuracy using machine learning algorithms and analyze the data using the Seaborn library. Accurate salary forecasting is crucial for both employers and employees in making informed decisions. By leveraging machine learning models, we can analyze historical salary data and relevant factors such as job title, experience, and education level to predict future salaries. Additionally, the Seaborn library will be used for visualizing and understanding the data, providing valuable insights for stakeholders. This project combines the power of machine learning and data visualization to improve salary forecasting accuracy and decision-making processes. This machine learning-based algorithm is based on real-time data set with different Employment statuses, Job roles and locations. In this, I have used the Linear Regression - Multivariate algorithm(regression-based algorithm) for training the machine and the one-hot encoding technique to handle multiple location-based data. In this project, I am trying to increase the accuracy of the machine learning model by removing the outliers present in the data. This project will be helpful to find out the salary based on the employment status, Job location and the job profile they are searching for.

Existing System:

In the existing system (Reference 4 - SALARY PREDICTION USING MACHINE LEARNING) the data set they have used is experience vs salary which us a univariate linear regression algorithm. Based on the experience the algorithm will predict the salary. Also, there are some online resources to check the salary based on employment status, job profiles and locations. They use the data mining concept to fetch the data to which usual google search we can refer. The data shown will be based on availability.

Drawbacks of existing system:

While using machine learning (ML) algorithms for salary forecasting can provide high accuracy and valuable insights, there are several drawbacks and considerations to keep in mind:
Data Quality and Bias: The accuracy of the ML model heavily relies on the quality of the input data. Biases in the data, such as gender or racial biases, can lead to biased predictions and unethical outcomes.
Overfitting: ML models may overfit the training data, meaning they perform well on the training data but poorly on unseen data. Regularization techniques and cross-validation can help mitigate overfitting.
Regarding performing data analysis with the Seaborn library, some drawbacks include:
Limited Customization: While Seaborn provides a high-level interface for creating attractive and informative visualizations, it may have limited customization options compared to lower-level libraries like Matplotlib.
Complexity for Advanced Plots: Seaborn may not be as suitable for creating highly customized or complex plots, requiring users to switch to Matplotlib or other libraries for advanced plotting needs.

Proposed System:

In the proposed system I have used multi-variate linear regression Machine learning algorithms to analyse the data and build a system to predict salary based on a few parameters. In the existing system, there is no customised search is available. In the proposed method by analysing the realtime dataset we can predict the salary based on custom search requirements such as employment status, Job location and job profile. Also, I am performing data analysis using seaborn to visualise and analyse the real-time data sets. Using Scikit learn library I am implementing a regressionbased algorithm which is linear regression for this salary prediction. Machine learning is applied to different data sets such as data set with multiple outliers, data set with few outliers removed and data set with no outlier and creating the machine learning models and comparing the accuracy of all of them. Finally, the high accuracy-based model we are going to use for the application backend.

Advantages of proposed system:

Using machine learning (ML) for salary forecasting with high accuracy and performing data analysis with the Seaborn library offer several advantages:

Accurate Predictions: ML models can analyze historical salary data and other relevant factors to make accurate predictions about future salaries. This can help individuals and organizations make informed decisions.

Insights and Patterns: ML models can uncover hidden patterns and insights in the data that may not be apparent through traditional analysis methods. Seaborn's visualizations can further enhance understanding of the data by highlighting trends and relationships.

Efficiency: ML models can automate the process of salary forecasting and data analysis, saving time and effort compared to manual methods. This allows organizations to quickly derive insights and make decisions.

Customization: ML models and Seaborn visualizations can be customized to suit specific needs and requirements. This flexibility allows for the creation of tailored solutions that address unique challenges.

System Analysis:

Machine learning algorithms used for natural language processing (NLP) currently take too longto complete their learning function. This slow learning performance tends to make the modelineffective for an increasing requirement for real time applications such as voice transcription,language translation, text summarization topic extraction and sentiment analysis. Moreover,current implementations are run in an offline batch-mode operation and are unfit for real timeneeds. Newer machine learning algorithms are being designed that make better use of samplingand distributed methods to speed up the learning performance. In my thesis, I identify unmetmarket opportunities where machine learning is not employed in an optimum fashion. I willprovide system level suggestions and analyses that could improve the performance, accuracy andrelevance.

FEASIBILITY STUDY:

The feasibility of the project is analyzed in this phase and the business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis, the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

The feasibility study investigates the problem and the information needs of the stakeholders. It seeks to determine the resources required to provide an information systems solution, the cost and benefits of such a solution, and the feasibility of such a solution. The analyst conducting the study gathers information using a variety of methods, the most popular of which are:

Interviewing users, employees, managers, and customers.

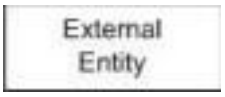
Developing and administering questionnaires to interested stakeholders, such as potential users of the information system.

Observing or monitoring users of the current system to determine their needs as well as their satisfaction and dissatisfaction with the current system.

Collecting, examining, and analyzing documents, reports, layouts, procedures, manuals, and any other documentation relating to the operations of the current system.

1FILE DESIGN

- **External Entities**



- External entities determine the system boundary. They are external to the system being studied. They are often beyond the area of influence of the developer.
- These can represent another system or subsystem. These go on margins/edges of data flow diagram. External entities are named with appropriate name.
- **Processes**

- Processes are work or actions performed on incoming data flows to produce outgoing data flows. These show data transformation or change. Data coming into a process must be "worked on" or transformed in some way. Thus, all processes must have inputs and outputs. In some (rare) cases, data inputs or outputs will only be shown at more detailed levels of the diagrams. Each process in always "running" and ready to accept data.
- The major functions of processes are computations and making decisions. Each process may have dramatically different timing: yearly, weekly, and daily.
-
- **Naming Processes**
-
- Processes are named with one carefully chosen verb and an object of the verb. There is no subject. The name is not to include the word "process". Each process should represent one function or action. If there is an "and" in the name, you likely have more than one function (and process). For example, get invoice update customer and create Order Processes are numbered within the diagram as convenient. Levels of detail are shown by decimal notation. For example, top level process would be Process 14, next level of detail Processes 14.1-14.4, and next level with Processes 14.3.1-14.3.6. Processes should generally move from top to bottom and left to right.
- **Data Flow**
- 

- Data flow represents the input (or output) of data to (or from) a process ("data in motion"). Data flows only data, not control. Represent the minimum essential data the process needs. Using only the minimum essential data reduces the dependence between processes. Data flows must begin and/or end at a process.
- Data flows are always named. The name is not to include the word "data". Should be given unique names. Names should be some identifying noun. For example, order, payment, complaint. • **Data Stores**
- 

- or
-

- Data Stores are repositories for data that are temporarily or permanently recorded within the system. It is an "inventory" of data. These are the common link between data and process models. Only processes may connect with data stores.
- There can be two or more systems that share a data store. This can occur in the case of one system updating the data store, while the other system only accesses the data.
- Data stores are named with an appropriate name, not to include the word "file", Names should consist of plural nouns describing the collection of data. Like customers, orders, and products. These may be duplicated. These are detailed in the data dictionary or with data description diagram.

PROCESS DESIGN

INPUT DESIGN

The necessity for the success of a project is the input design. Input design is the process of converting user-originated inputs into machine-readable form. Input specification describes the way in which the data is arranged to enter into the system for processing.

The goal of input design is to make data entry easier, logical, and error free. The decisions made during the input design are:

- To provide cost-effective methods of input
- To achieve the highest possible level of accuracy
- To ensure that input is understood by the user

The most important aspect concerned with input design is that the data must be correct. For this input, validation must be performed. Validations are done in the testing phase of the project. The input screen is in such a way that it is user-friendly and easy to use. The input actions performed in this project are preprocessing for checking the delimiters and new file uploading is given as input in this application

OUTPUT DESIGN

Output design produces the hardcopy regarding the information requested or displays the output in a predefined format. In a web-based application hard copy reports are not mostly generated. Only the user can view the needed information. It is the direct source of information to the end user. Efficient and intelligible outputs improve the system’s relationships with the users and help in decision and design making.

The nature of processing and procedure related to the system were classified and gives the output results. Output from the computer storage is required primarily to communicate the result of processing to the users and provide a permanent copy of the result for later reference.

The outputs generated are similarity record already available and confirms whether the users to take duplicates or not. The existing record with the attributes is listed to avoid duplication and memory wastage.

Database Design

Store and sync data with our NoSQL cloud database. Data is synced across all clients in realtime and remains available when your app goes offline. The Firebase Realtime Database is a cloudhosted database. Data is stored as JSON and synchronized in realtime to every connected client. When you build cross-platform apps with our Apple platforms, Android, and JavaScript SDKs, all of your clients share one Realtime Database instance and automatically receive updates with the newest data.

Code Design:

Designing a machine learning system is an iterative process. There are generally four main components of the process: project setup, data pipeline, modeling (selecting, training, and debugging your model), and serving (testing, deploying, maintaining).

The output from one step might be used to update the previous steps. Some scenarios:

After examining the available data, you realize it's impossible to get the data needed to solve the problem you previously defined, so you have to frame the problem differently.

After training, you realize that you need more data or need to re-label your data.

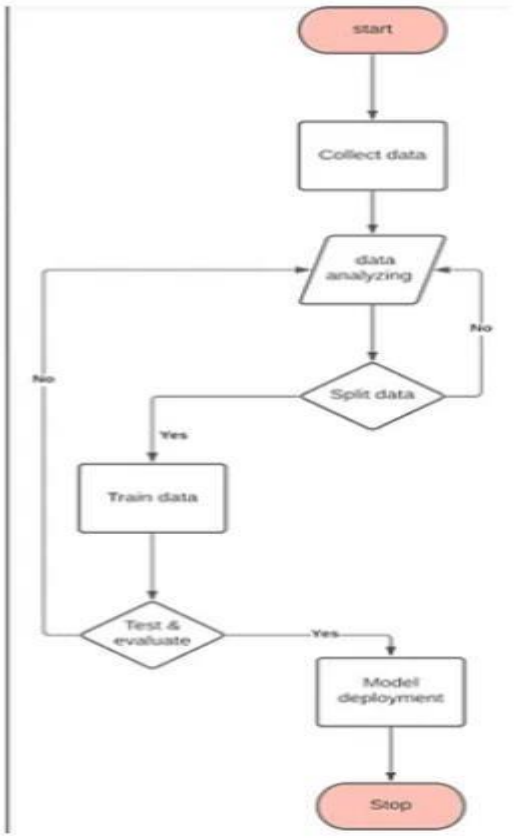
After serving your model to the initial users, you realize that the way they use your product is very different from the assumptions you made when training the model, so you have to update your model.

System Development:

System Module Specification

The objective of the module are as follows: to introduce the basics of artificial intelligence, the concept of intelligent agents, to analyze problem solving and search techniques, to understand how to represent knowledge, to do planning, to reason with uncertain knowledge, to analyze techniques for decision making, and finally to introduce neural networks.

Dataflow diagram:



Modules:

1. Data Set collection
2. Data analysis
3. Data Cleaning
4. Data Exploration
5. Data Visualisation
6. Data Preprocessing
7. Building a machine learning model
8. Developing a prediction function

Acknowledgment

This article / project is the outcome of research work carried out in the Department of **Computer Science** under the DBT Star College Scheme. The authors are grateful to the Department of Biotechnology (DBT), Ministry of Science and Technology, Govt. of India, New Delhi, and the Department of **Computer Science** for the support.