**CERTIFICATE FROM THE INDUSTRY**



**ABSTRACT**

The aim of this project was to evaluate and investigate the CTC of new hires in order to reduce the workload on the Human Resource team in the companies. This will benefit the company in automating the task of deciding the salaries to a great extent and will leave more time to focus on more important issues. To this end, training as well as test data were provided. The data from these datasets was preprocessed and cleaned to remove the outliers and imputer the rows with non-existent values. From these datasets, models were built using Jupyter Notebook to examine the error and observe how accurate the prediction is. Cross-validation was used to check the performance of the data on a validation set and the mean and standard deviation of the scores was checked accordingly. After a preliminary study of machine learning algorithms and data review, it became apparent that the problem fell under the linear regression category. Thus, the model was then finetuned using different hyperparameters and an appropriate parameter was decided based on the cross-validation score given by them on the test data. The model is then tested on test cases and it is concluded that the model can predict accurately on the test data with less bias.

**TABLE OF CONTENTS**

Certificate i

Acknowledgement ii

Certificate from industry iii

Abstract iv

Table of Contents v-vi

List of Figure vii

List of Tables viii

**CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW 1**

* 1. Introduction
     1. Data, Models and Machine Learning Task 2
     2. Types of Machine Learning techniques 3
  2. Basic terms of project 5
  3. Literature Overview 6
  4. Motivation 7
  5. Organization of Project Report 9

**CHAPTER 2: METHODOLOGY ADOPTED**  **10**

2.1 Frame the objective 10

2.2 Data Collection 13

2.3 Data Visualization 14

2.4 Data Preparation 15

2.5 Model Selection and Training 16

2.6 Fine Tuning the Model 17

2.7 Presenting the Solution 17

2.8 Launch Monitor and Maintain 18

**CHAPTER 3: DESIGNING AND RESULT ANALYSIS 20**

3.1 Loading the Libraries 20

3.2 Loading the dataset 20

3.3 Exploratory Data Analysis 22

3.4 Splitting the dataset into train and test data 22

3.5 Training a linear regression model 28

3.6 Evaluating results of the linear regression model 29

3.7 Training a decision tree regression model and evaluating results 31

3.8 Comparing models- cross validation 32

3.9 Confidence Interval 33

**CHAPTER 4: MERITS, DEMERITS AND APPLICATIONS 35**

4.1 Merits 35

4.2 Demerits 36

4.3 Applications 36

**CHAPTER 5: CONCLUSIONS AND FUTURE SCOPE 38**

5.1 Conclusion 38

5.2 Future Scope 38

**REFERENCES**

**40**

**List of Figures**

**Figure No. Title of Figure Page No.**

1 Data analysis introduction 2

2 Implementation of Machine Learning tasks 10

3 Data Visualization 15

4 Underfitting and Overfitting 17

5 Flow Diagram ML 19

5 Jupyter Notebook code snippets 20-34

6 Future Scope of Machine learning 39

**List of Tables**

**Table No. Title of Table Page No.**

1 Flow table of ML methodology 10

2 Feature-significance table 13

**CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW**

* 1. **INTRODUCTION**

A major reason, nowadays, as to why companies face employee resignment is because the salary they get is a value that they think is not enough for them. This switching leads to a loss to the company and the company starts doubting the HR team for not being able to correctly assess the worth of a potential employee. In order to prevent such loss in this highly competitive world, an automated calculation of the salary of the newly hired employees based on their experience and many other factors, seems like a great machine learning algorithm to implement that will not only lead to ease of work done by the HR team but also lead to satisfactory CTC being offered to the employees. The exact CTC however cannot be determined and only a prediction can be made and the aim is to make this prediction which enough accuracy to grant legibility to the model. A prediction is an assumption about a future event.

In this project, the main aim is to predict the CTC of the newly hired employees based on their qualifications and experience, and other factors. For developing this system, a Linear Regression model of supervised machine learning technique is used since the CTC that is to be predicted is a continuous number.

Machine Learning is a form of predictive analysis associated with the motive of getting computers to learn without being explicitly programmed. This field of study is crucial in solving those problems, the answers to which are either not known by the humans or even if they know how to get the answers, the solutions to them cannot be easily programmed to be able to give the knowledge of the solutions to the computer. Data is a crucial element in machine learning and the entire algorithm is data driven with an automated program design.



Figure 1 Data analysis introduction

**1.1.1 DATA, MODELS AND ML TASK**

Data is an important element in machine learning and thus it needs to be handled carefully as only about 10-15% time is spent in building the machine learning model, the rest of the time is spent in capturing and pre-processing the data and then making wise decisions by consulting domain experts and product managers in the organization after the model has been implemented.

It is represented as a collection of vectors. Metadata is also an important class of data which is used to get information about the data. Metadata does not matter for the computer algorithms but it is useful for human interpretation of data and understanding the essential features regarding the data to model.

A model is a mathematical simplification of reality and is a core player in data science. It is almost always representing reality but is simpler and compact than reality. It is a learnt function that helps in prediction without requiring the training data.

Predictive models are deterministic and use a procedural approach to make the prediction by analyzing patterns in the input data.

Probabilistic models on the other hand score different configurations of reality and depending on which configuration has a higher score, choose the appropriate prediction.

A learning algorithm is responsible for converting data into models by choosing from a collection of models with same structure but different parameters.

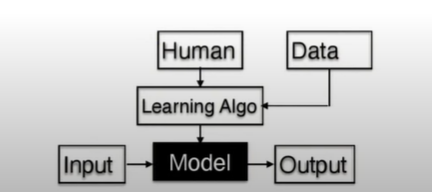


Figure 2 Implementation of Machine Learning tasks

**1.1.2. TYPES OF MACHINE LEARNING**

* **Supervised machine learning**

Supervised Learning is a technique in Machine Learning that involves the prediction of labels using the features given to the model. It is called supervised as there is a supervision given already in the form of the desired output the algorithms should predict based on the learning function that maps an input to a desired output. The mapping is based on input-output pairs.

Examples: Regression, Classification, Structural Learning, Ranking, etc.

* **Unsupervised machine learning**

Unsupervised machine learning is an approach that involves little or no idea about the desired output expected from the algorithm and there is no specific label to predict. It is usually considered for pre-processing of the data before applying the supervised machine learning algorithm on it and is an essential component. The structure of the data can be derived via clustering based on the relationship among the variables. It is self-organized and is associated with observing the previously unknown patterns in dataset without pre-existing labels.

Examples: Clustering, Representation Learning, etc.

* **Sequential machine learning**

Sequential machine learning is an approach that does not involve the supervision in the form of the complete data given already. As the data gradually comes, the model learns sequentially by observing the feedback given by the present set of data. The feedback obtained after prediction is the form of supervision and the model is updated based on the feedback obtained.

Linear Regression is used to build an efficient linear model that fits all the data points and gives best results on the test set. It is used when the labels to be predicted take real values and thus approximates the mapping function to get the best results.

Examples: Online Learning, Reinforcement learning, Multi-armed Bandits

* 1. **BASIC TERMS OF PROJECT**

1. **Algorithm**: A sequence of steps or instructions to solve a problem.
2. **Machine Learning**: It is a data analysis method that automates analytical model building and is a branch in artificial intelligence used to make decisions with minimal human intervention.
3. **Data**: A collection of vectors.
4. **Python**: A high level programming language. Here, it has been used to implement machine learning algorithms.
5. **Data frame**: Two-dimensional, size-mutable, potentially heterogeneous tabular data.
6. **Array**: A collection of items of same data type stored at contiguous memory locations.
7. **Hyperparameters**: Parameters which are passed to the algorithm beforehand.
8. **Library**: A collection of related modules. It contains bundles of code that can be used repeatedly in different programs.
9. **Data Visualization:** graphical representation of information and data.
10. **Data preprocessing:** Data mining technique which is used to transform the raw data in a useful and efficient format
    1. **LITERATURE REVIEW**

The core objective of a learner is to generalize from experience. Computational learning theory is a branch in computer science that deals with the computational analysis of machine learning algorithms and measuring their performance on a set theory. The guarantees regarding the performance of the algorithms do not hold due to the finite training sets and uncertain future. In order to ensure optimal performance of the algorithms, the hypothesis should match the learning function of the underlying data. If the hypothesis is less complex than the function, then the model has underfit the data and if it is more complex, then the model has overfit the data. Both the cases will lead to more training error. Learning theories, in addition to performance bounds, also involve a systematic study of time complexity and algorithm feasibility that builds the efficiency in machine learning tasks and is thus an important aspect.

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.

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.

* 1. **MOTIVATION**

As an emerging and advanced technology, machine learning has been evidently contributing to the development of enterprises in various industries all over the world. Data is the new oil and it powers every little process taking place in the industry. The markets have seen an evident increase in the quality resourced and skills in the competitive environment hosting for an incentive for the organizations to invest more in human resources. This increases the essential requirement to meet a proper recruitment process so that the right candidates get an expected salary withstanding business competitiveness and efficiency. Thus, this project gives an emphasis on automating this entire process using the power of machine learning. The prediction is based on various factors like college tier, role, city type, previous CTC, previous job changes, graduation marks, experience, etc.

There are several motivations for developing and using salary prediction machine learning models, including:

1. Determining fair and competitive salaries: Employers can use salary prediction models to determine fair and competitive salaries for new hires and existing employees, which can help attract and retain top talent.
2. Improving efficiency: With the automation of the salary prediction process, organizations can save time and resources that would have been spent on manual calculations and evaluations.
3. Supporting data-driven decision making: By providing accurate predictions of salaries, salary prediction models can help organizations make data-driven decisions about their workforce, such as determining the budget for the next fiscal year.
4. Improving the job market: By providing accurate information about expected salaries, salary prediction models can help job seekers make informed decisions about their career paths and negotiate fair compensation packages.
5. Identifying patterns and trends in the labour market: Researchers and analysts can use salary prediction models to study the labour market and identify patterns and trends in compensation and employment.
6. Benchmarking: Employers can use salary prediction models to compare their compensation packages to industry standards and make adjustments as needed.
7. Supporting education and career development: By providing predictions of expected salaries for specific majors and careers, salary prediction models can help students make informed decisions about their college and career choices.
8. Supporting government and policy: Government agencies and policy makers can use salary prediction models to study the labour market and develop policies that promote fair compensation and employment.

Overall, salary prediction machine learning models can be used to determine fair and competitive salaries, improve efficiency, support data-driven decision making, improve the job market, identify patterns and trends in the labor market, and support education and career development. They can be a valuable tool for organizations, researchers and analysts, job seekers, and policy makers.

* 1. **ORGANISATION OF PROJECT REPORT**

**CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW**

This gives a brief introduction about the theory of machine learning and throws light on the different types of machine learning and its importance in today’s world. A literature review has been provided based on a scholarly research-based approach on the various articles scanned as per the current knowledge pertaining to the subject.

**CHAPTER 2: METHODOLOGY ADOPTED**

This gives a full-fledged flow of how the CTC prediction machine learning model has been built by following a sequential algorithmic flow theoretically in a detailed manner along with discussing the extreme possibilities of the model.

**CHAPTER 3: DESIGNING AND RESULT ANALYSIS**

This gives the practical implementation of the machine learning algorithm associated with snippets from the jupyter notebook codes and their results in well-oriented manner. The results of each and every step have also been portrayed along with proof of correctness. In the end it has been proved that the linear regression model is an appropriate model to vouch for.

**CHAPTER 4: MERITS, DEMERITS AND APPLICATIONS**

The advantages and disadvantages of building a CTC prediction machine learning model have been discussed along with the applications which illustrate the use of the model in real life scenarios.

**CHAPTER 5: CONCLUSION AND FUTURE SCOPE**

This section concluded the CTC prediction machine learning project and emphasizes on its importance in the future as machine learning continues to acquire various industrial domains.

**CHAPTER 2: METHODOLOGY ADOPTED**

Data Collection

Data Visualisation

Framing the problem

Data Preparation for ML algorithms

Fine Tuning the model

Model Selection and Training

Launch, Monitor, Maintain

Presenting Solution

Figure : Flow of machine learning algorithm

**2.1 Framing the problem**

This involves looking at the big picture that is associated with the entire objective of the problem.

**OBJECTIVE: To predict the CTC of the new hires using the dataset provided.**

It involves listing the input and the output associated with the model. The input is a full-fledged dataset in order to implement supervised machine learning while the output is the predicted value of CTC of the new hires.

**BUSINESS OBJECTIVE**: To increase the efficiency of the HR teams in calculating the CTC to be offered to the new hires with less trouble and reduction of loss for the company in terms of resigning of jobs by the employees due to an unexpected salary.

This model will benefit the company by automating the task of prediction of CTC to a great extent and thereby reduce the workload on HR teams.

This information will be useful in algorithm and performance measure selection and will aid in overall effort estimation.

The current solution to the problem involves mathematical calculations done by the HR teams in the companies by weighing the features based on their importance which they know with respect to their domain knowledge and thereby assessing the CTC of the new hires through the parameters manually or in excel.

Given the baseline, machine learning will surely provide a reliable and better solution to the problem using the required data.

It is useful to have knowledge about the design consideration in problem framing as to whether a problem is - supervised or unsupervised, single output or multiple output, continuous learning or periodic updates, batch learning style or online learning style?

Performance measures need to be selected in order to be able to judge how accurate a model performs on a given dataset.

A regression model has the performance measures –Mean Squared Error (MSE) or Mean Absolute Error (MAE).

A classification model has the performance measures – F1- Score, Precision, Recall, Accuracy.

Assumptions regarding the machine learning task need to be listed and reviewed by consulting with the domain experts before progressing as it ensures correctness in the algorithm developed using the data.

**2.2 Data Collection**

Data spread across multiple tables, files, documents with appropriate access controls. Thus, appropriate authorization is needed to access the data. SQL is an important tool that can be used here in order to extract rows from the data and get familiarized on how the data looks.

The first step is to load the basic libraries that are needed to perform the required machine learning task. Libraries used in this project include:

* Pandas – This library focuses purely on data and its statistics and is useful in calculating mean, average and other important statistics along with building a data frame.
* Numpy: This library provides very fast computation like quantum computation, image processing, astronomy and works well with arrays which form the core data structure in machine learning.
* Matplotlib: This library is used for statistical visualisation- line charts, scatter plot, bar graph, histogram.
* Seaborn: This library is used for data visualisation and is based on matplotlib - box plot, migration plot, etc.
* Scikit learn: This library is the core machine learning library and is well loaded with datasets to analyse and a pool of models to select from.

Once the dataset has been loaded, the data is checked.

The CTC prediction dataset involves a total of 1338 data samples (rows) and 9 features(columns).

The 9 columns are: - College, Role, City type, Previous CTC, job changes, graduation marks, Exp (months), CTC.

Each feature needs to be carefully examined by consulting the experts and the meaning of each column needs to be understood.

|  |  |
| --- | --- |
| Feature | Significance |
| College | It is associated with 3 tiers – Tier 1 being the best. |
| Role | The role for which the new employee has been hired for whether it is for a manager or an executive. |
| City type | Whether the employee belongs to a metro type (urban) or a non-metro type (rural) city. |
| Previous CTC | Value of CTC of previous job of the employee. |
| job changes | The number of times the employee changes his/her job. |
| graduation marks | Marks attained by the employee when he/she graduated. |
| Exp (months) | Number of months of experience that employee has in his domain. |

The data statistics are studied various plots are made in order to get an intuitive insight on how the data looks like.

It is observed that the values of features are at different scales and have different distributions.

When we look at the test set, it is likely to notice patterns in that and based on that we may select certain models. This leads to biased estimation on test set, which may not generalize well in practice. This is called data snooping bias.

Thus, we apply random sampling on the database in order to split the data into train and test since the model needs to learn from training data and needs to predict on test data.

Scikit learn provided a class called train\_test\_split in the model\_selection package and it performs random sampling in order to randomly divide data into train and test. A good practice is to have 80% training data and 20% test data.

**2.3 Data Visualization**

Data visualisation is performed on training set and is useful to understand features and their relationship with output label. In case of large number of samples, we take samples from our actual dataset and this set is used for further analysis. Such a set is called exploration set. But in the CTC prediction dataset, it is not required. The libraries Seaborn and Matplotlib are used for visualising data in the form of scatter plots, box plots, histograms, bar graphs, etc.

The correlation between the features and the label is studied and plotted using a heatmap.

Correlation gives the linear relationship among the features and the labels and rank correlation is used to give the non-linear relationship.

A similar analysis can be carried out with combined features – features derived from other features.

Exploration is an iterative process and thus, need not be thorough. More insights can be obtained once the model is built.

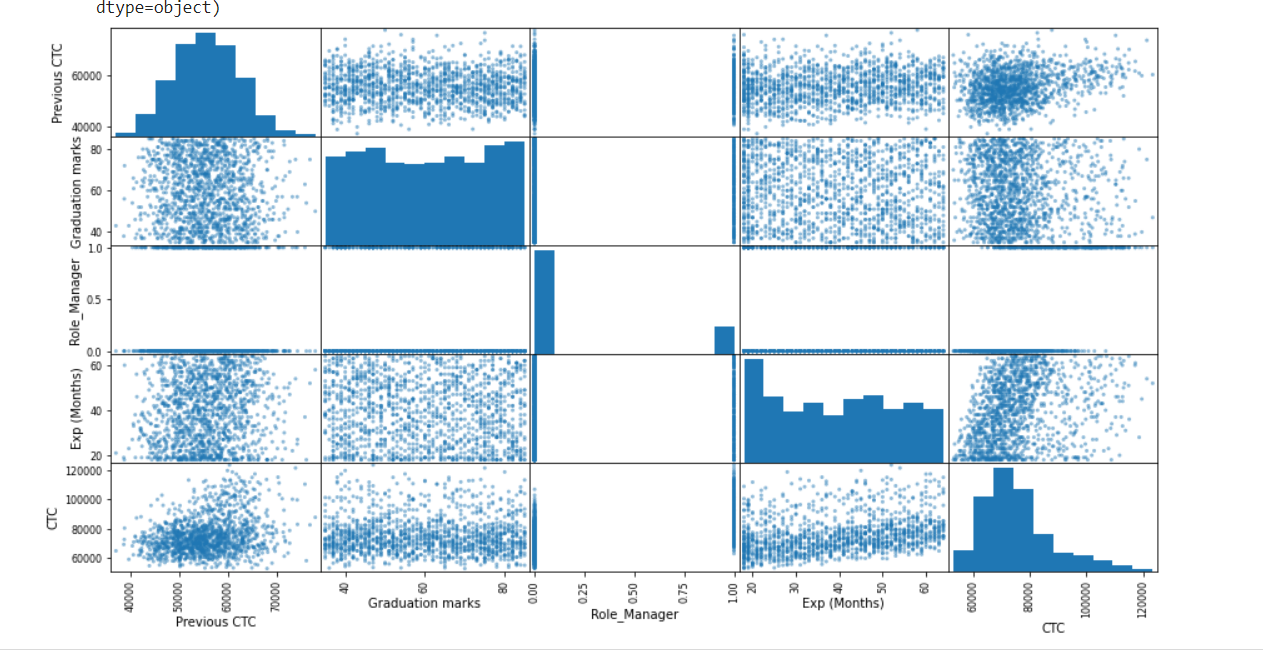
****

Figure 3 Data visualisation

**2.4 Data Preparation**

* Data may contain outliers or missing values due to errors in data capture.
* Different features may be at different scales.
* The current data distribution may not be amenable to learning.

Thus, data needs to be preprocessed by following the steps:

* Separating features and labels.
* Handling missing values and outliers.
* Feature scaling to bring all features on the same scale.
* Applying certain transformation like log, square root, etc.

If the values are not recorded or recording errors are there, then they should be imputed using SimpleImputer class in sklearn or the rows containing NaN should be dropped drop(), dropna().

If they do not exist, it is better to keep them as NaN only.

The categorical variables need to be converted to numerical variables via the get\_dummies() method, in order to use them in machine learning tasks.

Ordinal\_Encoder class and OneHotEncoder class are 2 other classes provided by sklearn in order to convert the categorical variables to numerical variables.

They can also be converted into low dimensional learnable vectors called embeddings.

Min-max scaling or Normalisation is a method employed to bring the features to a common scale. MinMaxScalar in sklearn is used in which we subtract minimum value from current value and divide it by difference between maximum and minimum value of the feature overall. Values are shifted and scaled so they range between 0 and 1. Hyperparameter- feature\_range. Standardization is another method used for the same, using StandardScalar in sklearn - subtract minimum value from current value and divide it by standard deviation so the resulting feature has unit variance. It does not bound values in specific range and less affected by outliers compared to normalisation.

**2.5 Model Selection and Training**

It is a good practice to build a quick baseline model on the preprocessed data and get an idea about the model performance.

There are several types of models to choose from, such as linear regression, decision trees, and neural networks. Linear regression is a linear approach, where the relationship between the input and output variables is represented by a straight line. Decision trees are non-linear models that use a tree-like structure to represent the relationships between the input and output variables. Neural networks are a type of deep learning model that are good for handling high-dimensional data.

In the CTC prediction dataset, first the linear regression model is chosen and the mean square error of both the training set as well as the test set is analyzed. If the mean square errors of both are close, it means the model works well. After this a Decision Tree model is also trained and analyzed using mean square error. Likewise, the best model is selected by cross-validation that provides a separate mean squared error for each validation set which can be used to ger a mean estimation of MSE as well as the standard deviation that helps to decide the accuracy of the model.

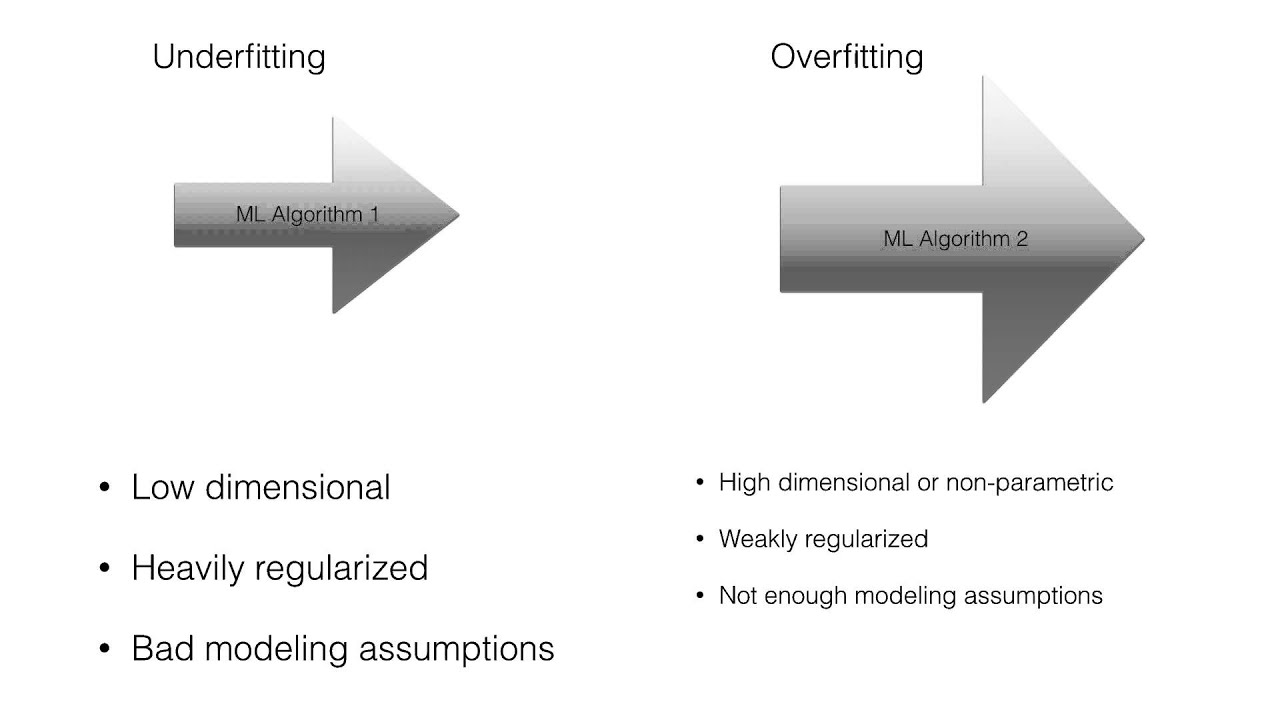


Figure 4 Underfitting and overfitting

**2.4 Fine Tuning of model**

Hyperparameter are the parameters in the model that are set manually and thus tuning them leads to a better accuracy. Finding the best possible combination of these hyperparameters is a search problem in the space of hyperparameters, which is huge.

Sklearn provides GridSearchCV and RandomisedSearchCV class that provides the evaluations of all possible combinations of hyperparameters in cross-validation. Thus, the best estimator is selected form the combination of results.

Feature importance is studied and those features which do not affect the CTC are dropped. And then the final model is built after having assured that the mean squared error belongs to a confidence interval of 95%.

The predictions are compared with the actual values and error is recorded using evaluation metrics.

**2.5 Presenting the solution**

After the model has been successfully implemented based on the performance on the test set, the prelaunch phase starts that involves presenting the solution that highlights learnings, assumptions, and system limitations. Everything is documented and clear visualizations are created. In case the model does not work better than the experts, it may still be a good idea to launch it and free up bandwidths of human experts.

Deployment of the model involves wrapping it in a web service or an API, so that it can be easily integrated into existing systems. This step requires knowledge of web development and API design.

**2.6 Launch, monitor and maintain**

Monitoring the model's performance over time is important to ensure that it remains accurate. This step can involve implementing methods such as online learning, concept drift detection, and re-training the model as necessary.

Launch involves plugging in the input and writing test cases in order to ensure and end-to-end correct working and to give more confidence about working of different pieces together.

Once the model is launched, it needs close monitoring for system outages and find ways to minimize the outages. If the model performance degrades, there should be strategy to maintain it either via retraining it within 2 weeks or a threshold metric can be kept which when crossed, would trigger the retraining of the model.

Sample predictions for human evaluation and it can be presented to the experts to get feedback on how the model is performing.

Regular assessment of data quality is critical for model performance and if the features are not captured properly then it will adversely affect the working of the model as well as the predicted output.

Model should be retrained after fixed intervals with fresh data as the data keeps changing with time.

Pushing the training model to production without disrupting the live path is also something to keep track of.

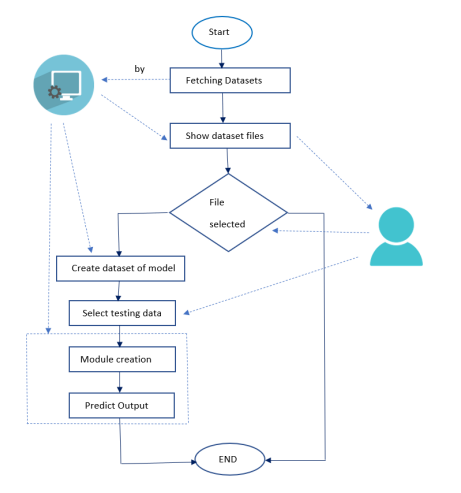
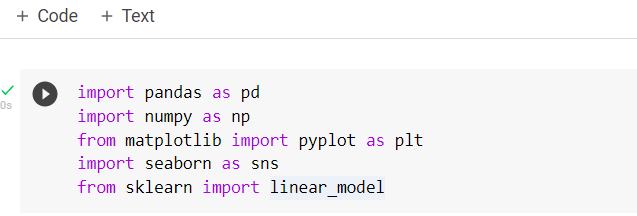
****

Figure 5 Flow diagram ML

**CHAPTER 3: DESIGNING AND RESULT ANALYSIS**

**3.1 LOAD THE LIBRARIES**

The libraries used in CTC prediction ML model are loaded in the jupyter notebook.

****

**3.2 LOAD THE DATASET**

The data file is first uploaded into the jupyter notebook kernel and then it is loaded using pandas library command:

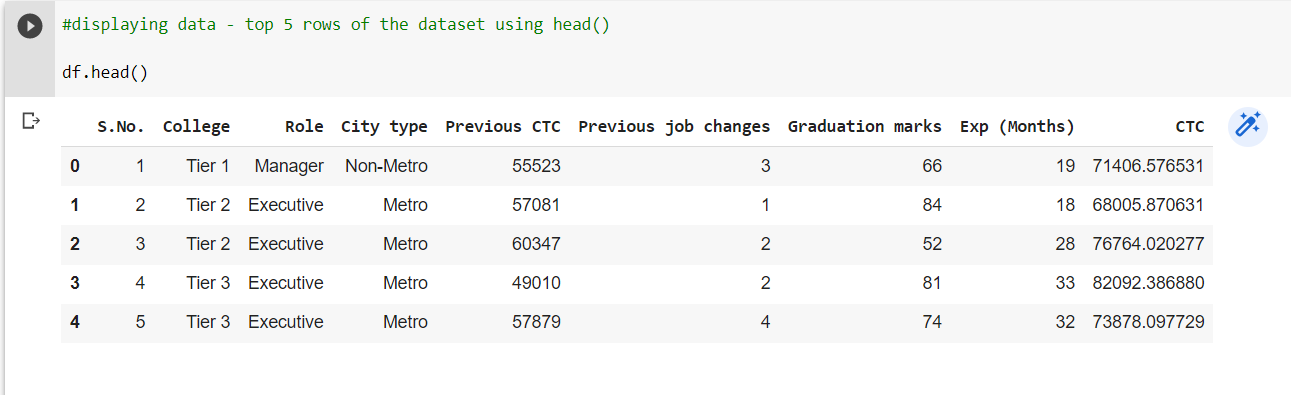
pd.read\_excel(‘<file\_path>’,header = <number>)

****

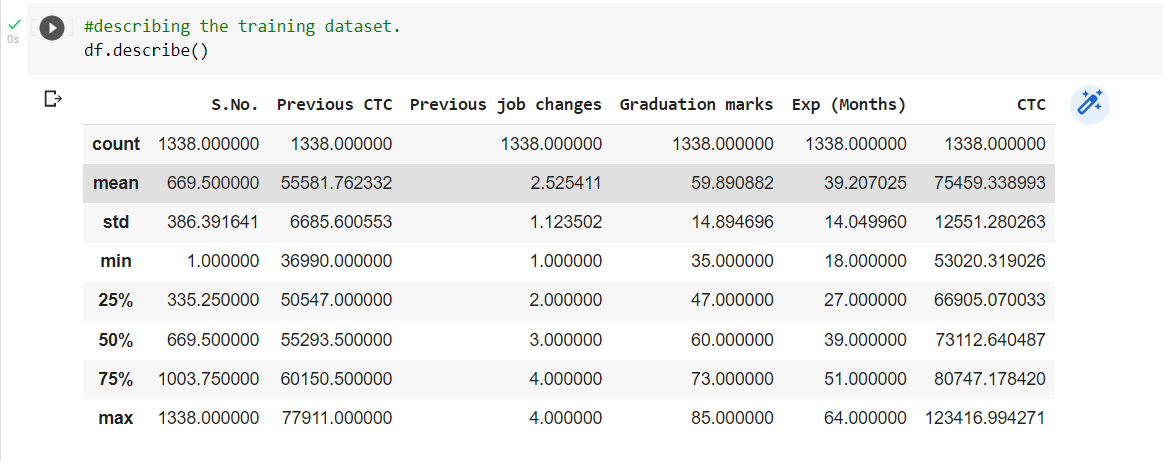
**3.3 LOAD THE DATASET**

Analysing the entire dataset using:

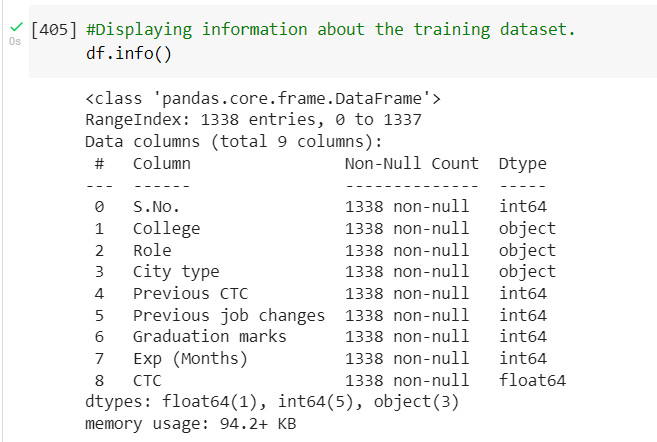
* Df.head() – gives the first 5 rows to analyse the features and their corresponding values.

****

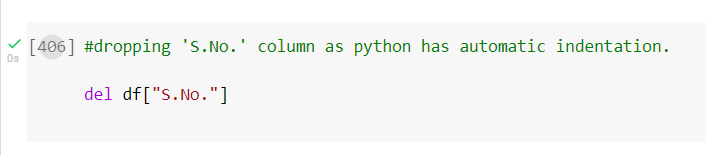
* Df.describe() – to give the summary statistics of the data including – mean, standard deviation, minimum value, maximum value, and 25th, 50th and 75th percentiles.

****

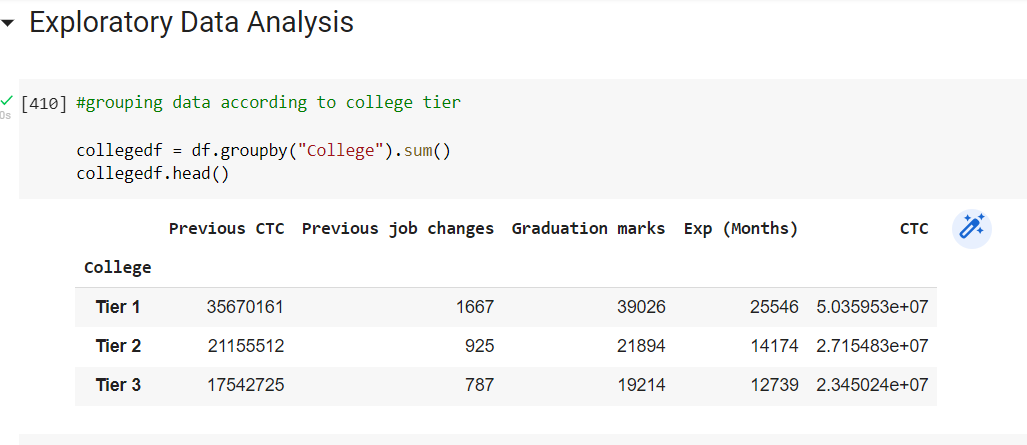
* Df.info() that gives information about the type of columns in the dataset.

****

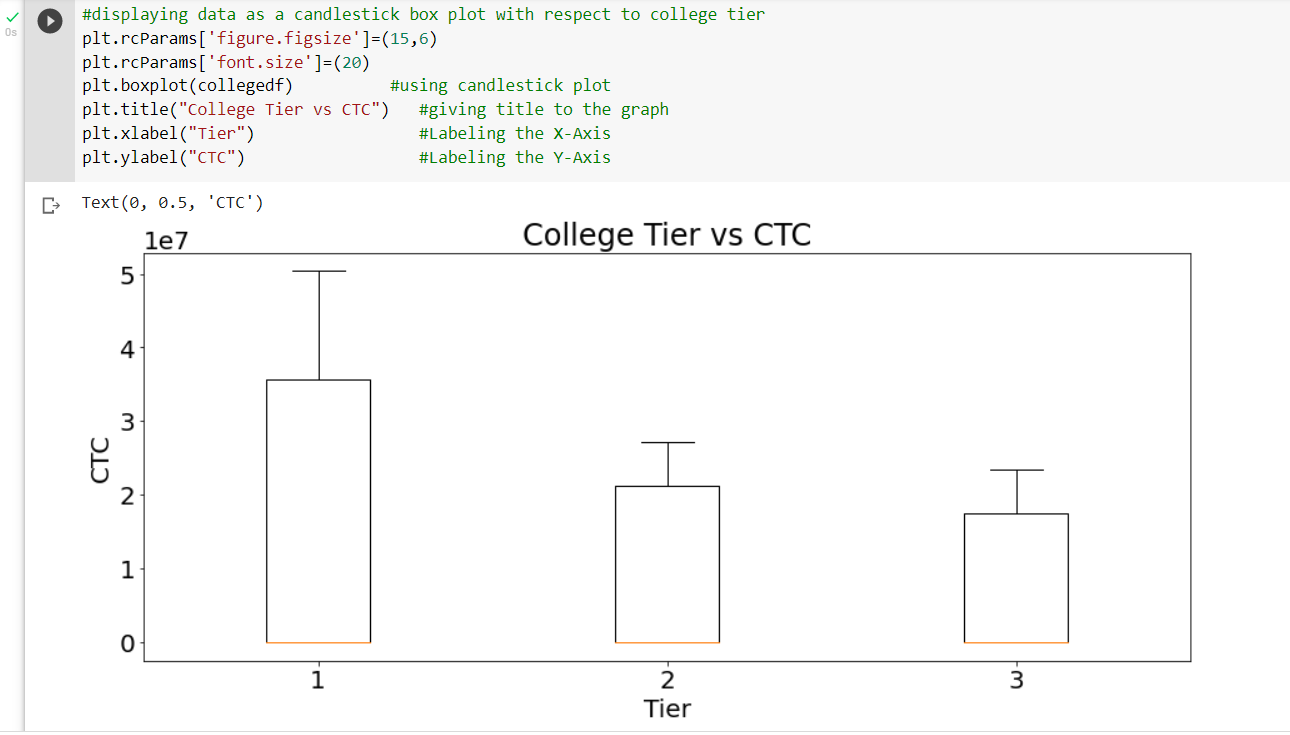
The serial number column is dropped due to automatic indentation of python.

****

**3.4 EXPLORATORY DATA ANALYSIS**

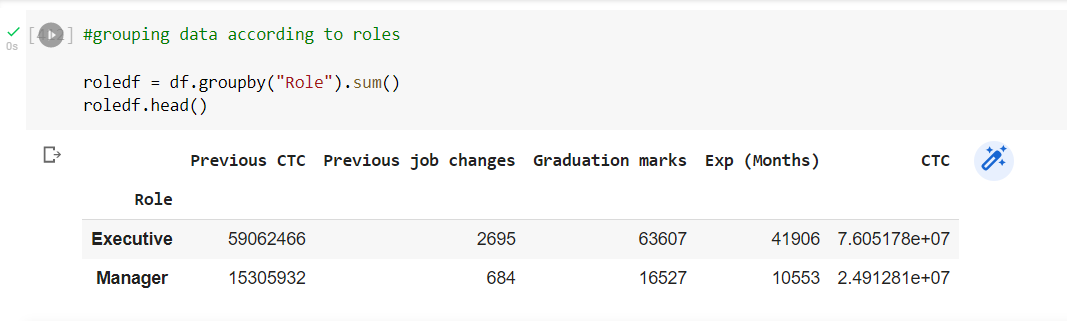
****

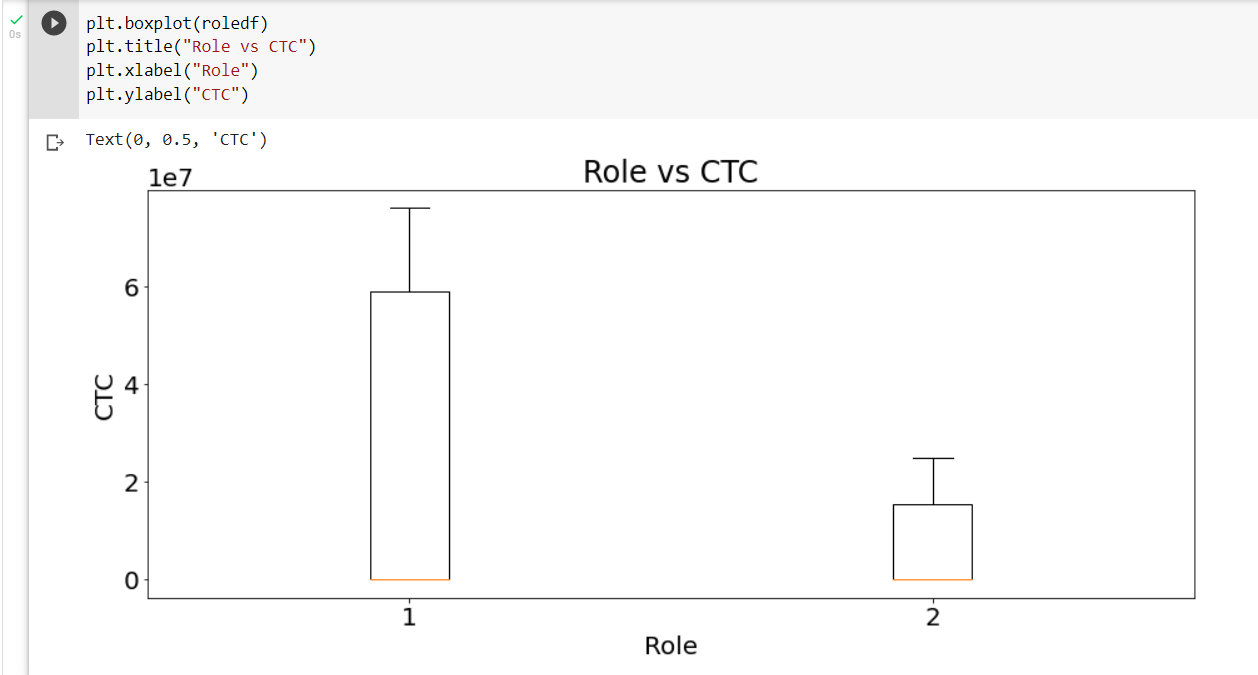
The data is grouped by the college tear and all the values corresponding to the features are summed up and stores in the respective cells. This dataframe is stored in a variable called collegedf.

****

A box plot is made using this dataset. It is also called a candlestick plot and is made between college tier and CTC in order to analyze the relationship between both the features.

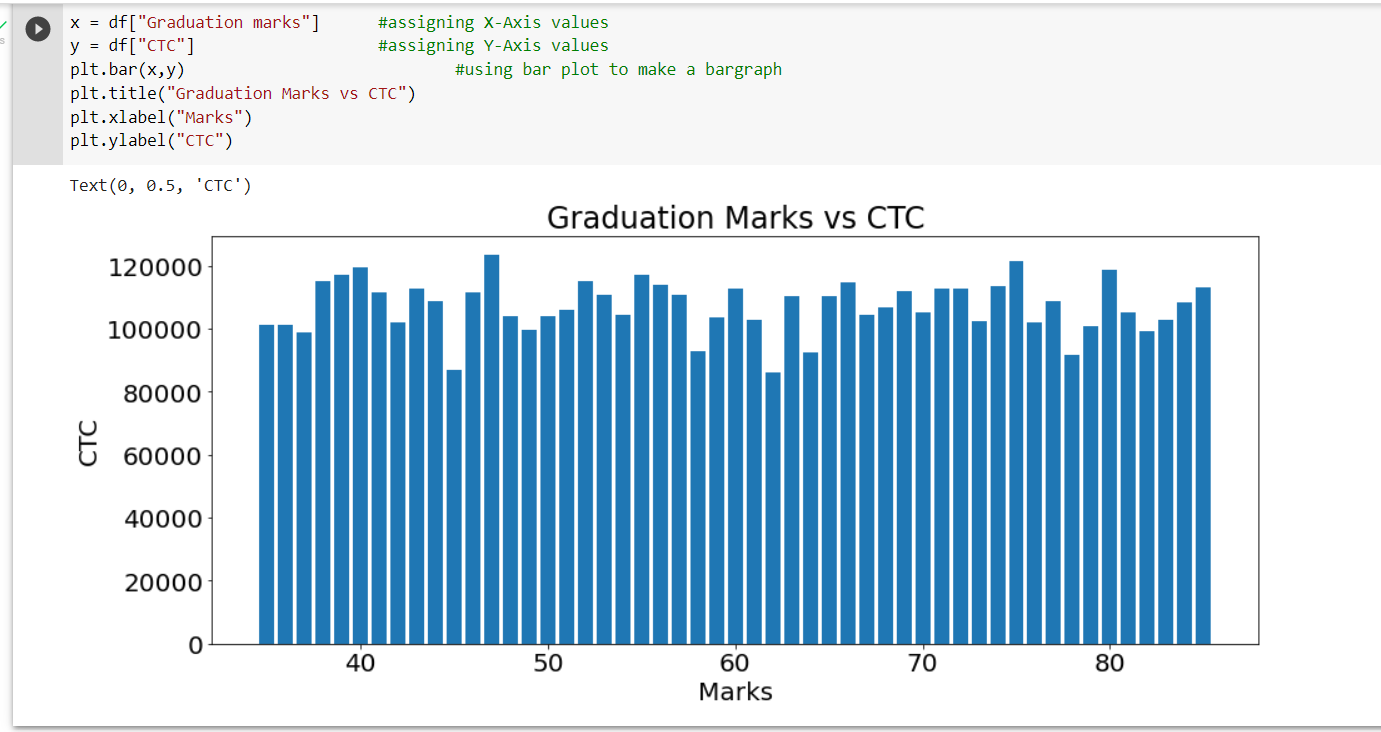
Similarly, the relationship between Role and CTC is studied.

****

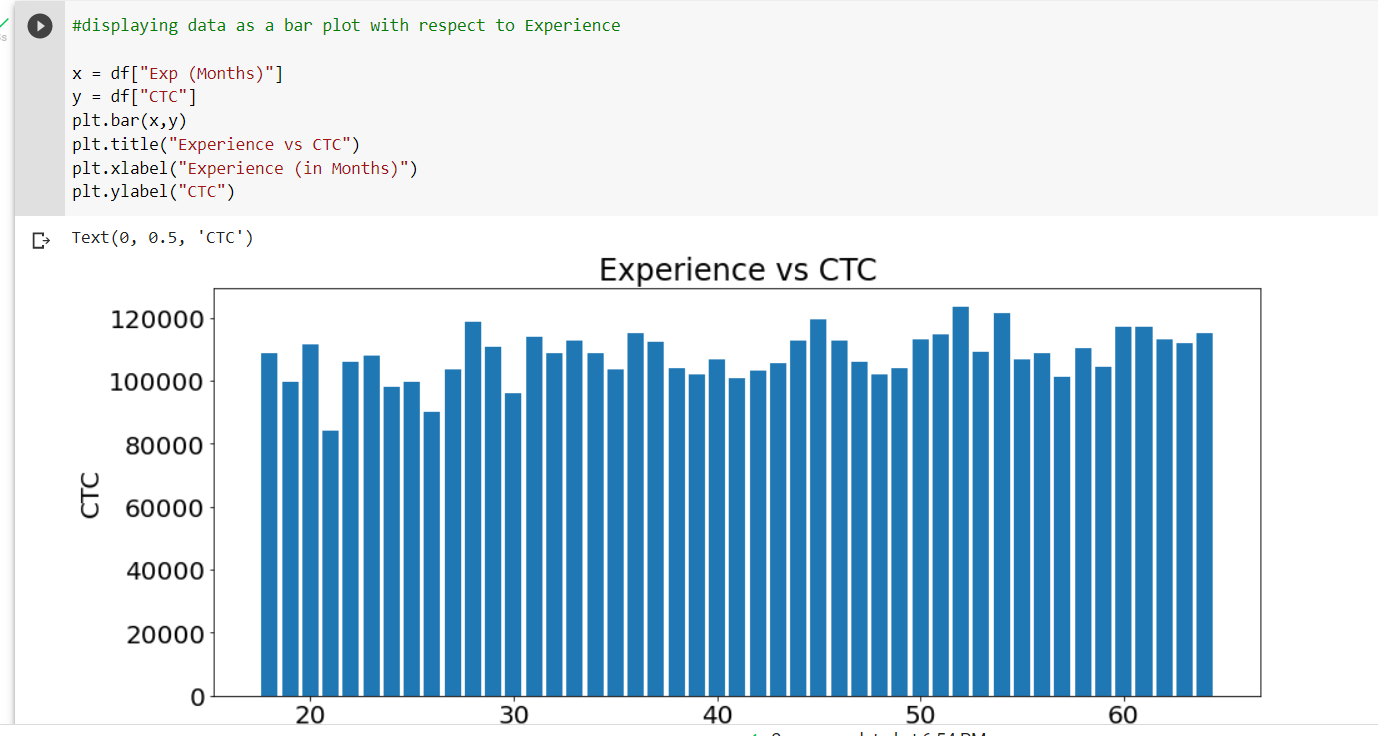
****

Similarly, bar graphs are designed using matplotlib and the relationships are studied:

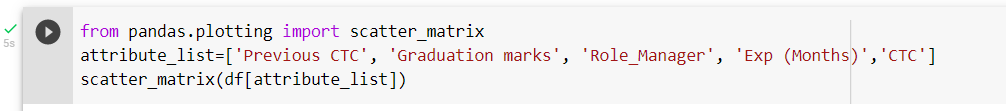
* **Graduation Marks and CTC**

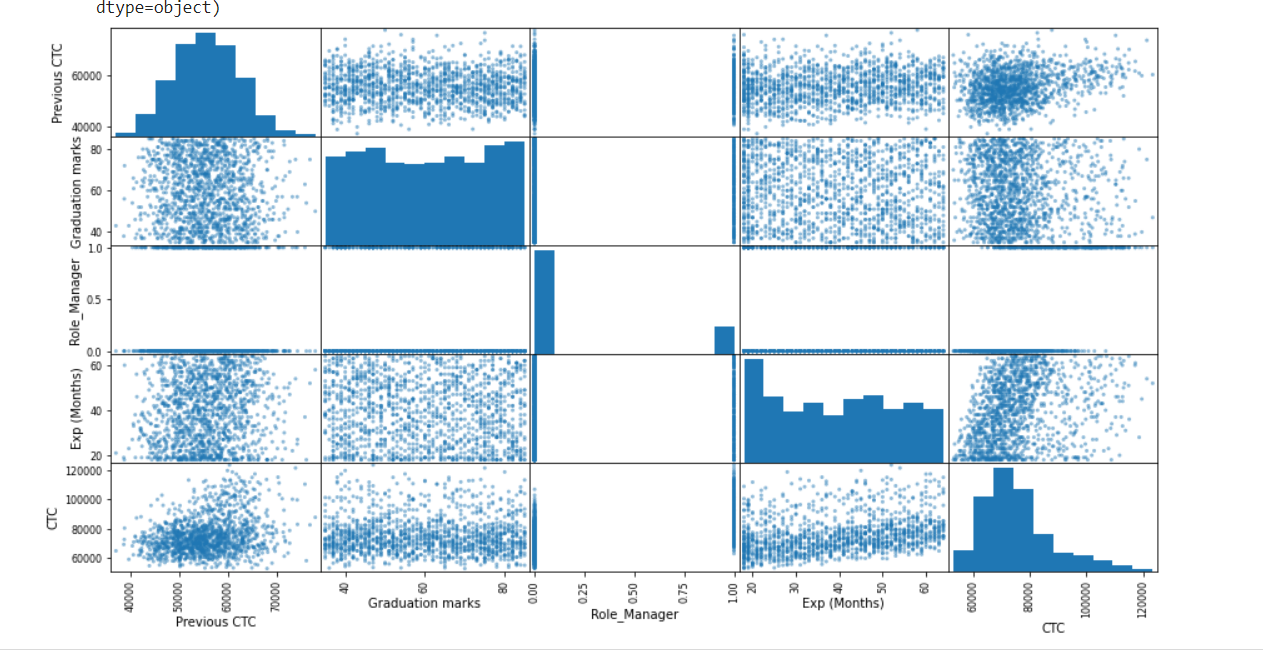
****

* **Experience and CTC**

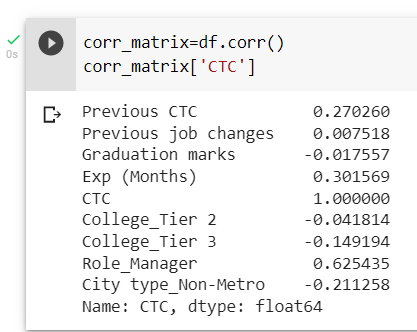
****

Scatter plots are made using matplotlib as follows:

****

****

Correlation among various features is studied using a correlation matrix as follows:

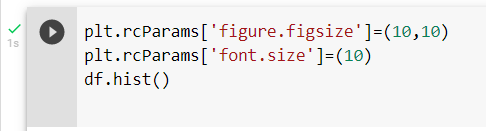
****

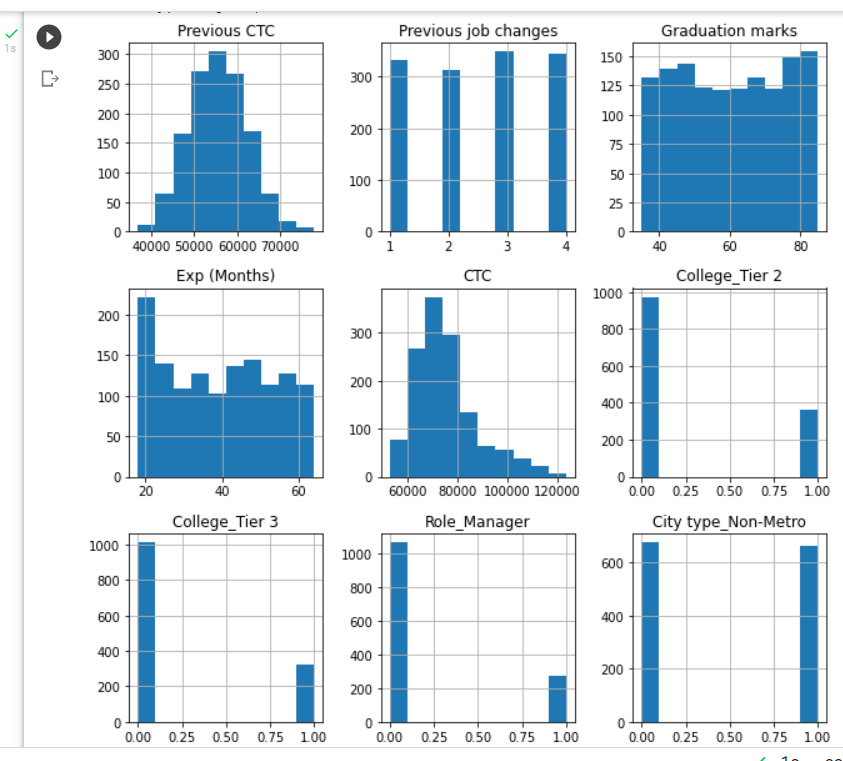
A heat map using the correlation matrix is generated as follows:

The dark colour indicated less correlation and the light colours indicate more correlation among the attributes.

****

Histograms are also created for all the features:

****

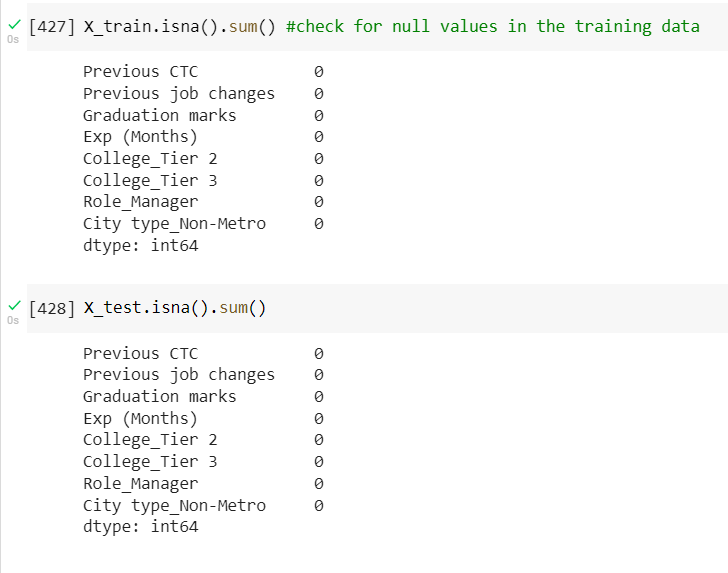
****

**3.5 SPLITTING THE DATASET INTO TRAIN AND TEST**

****

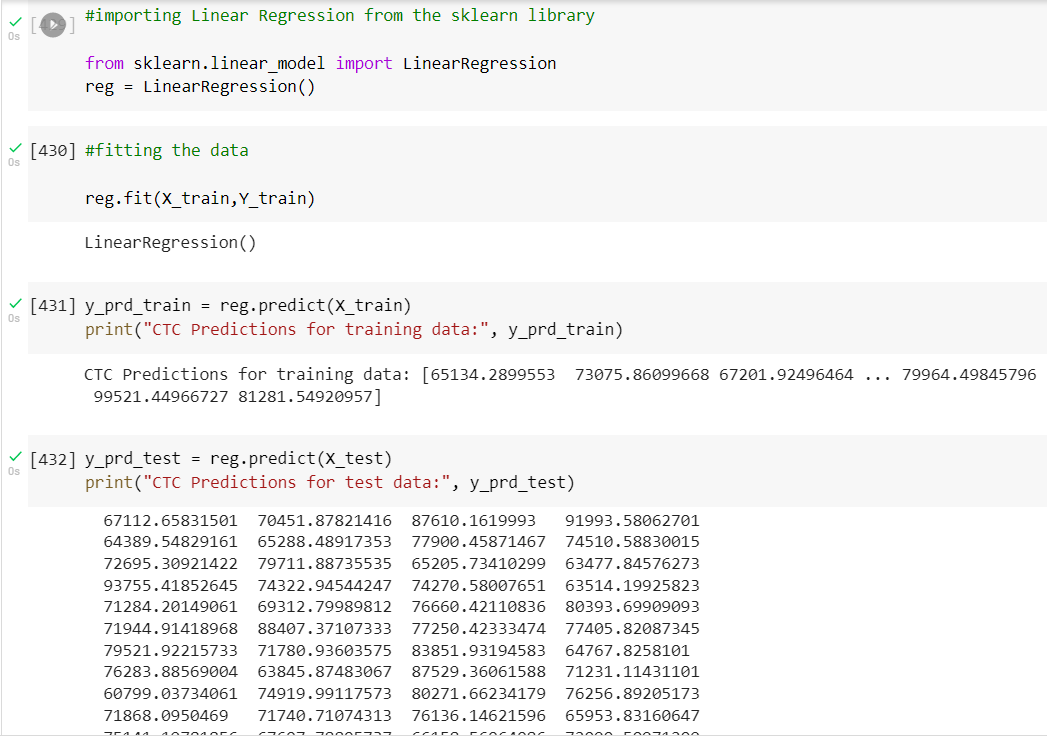
Random sampling is done to split the data intro training data and testing data by using train\_test\_split method of sklearn.model\_selection.

The training data has 1070 data points with 8 features.

****

The data set is checked of null values and it is observed that no null values are present in the dataset.

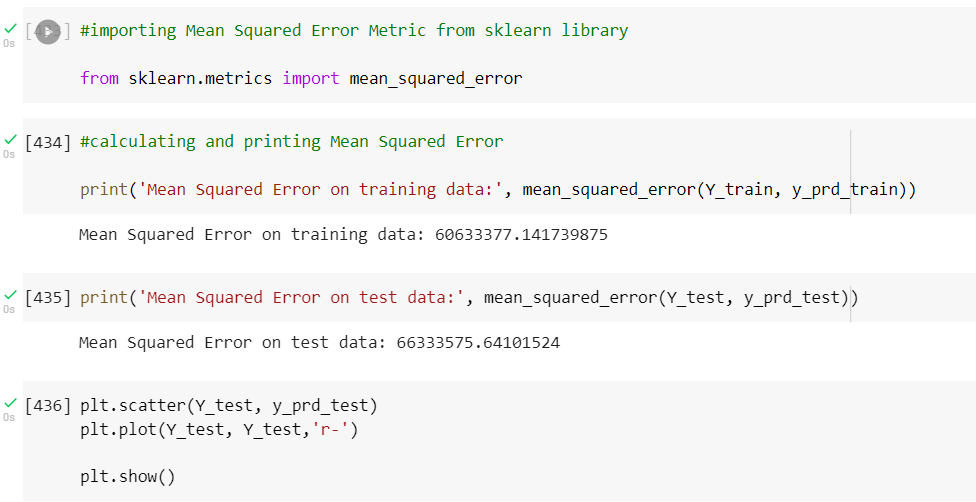
**3.6 TRAINING LINEAR REGRESSION MODEL**

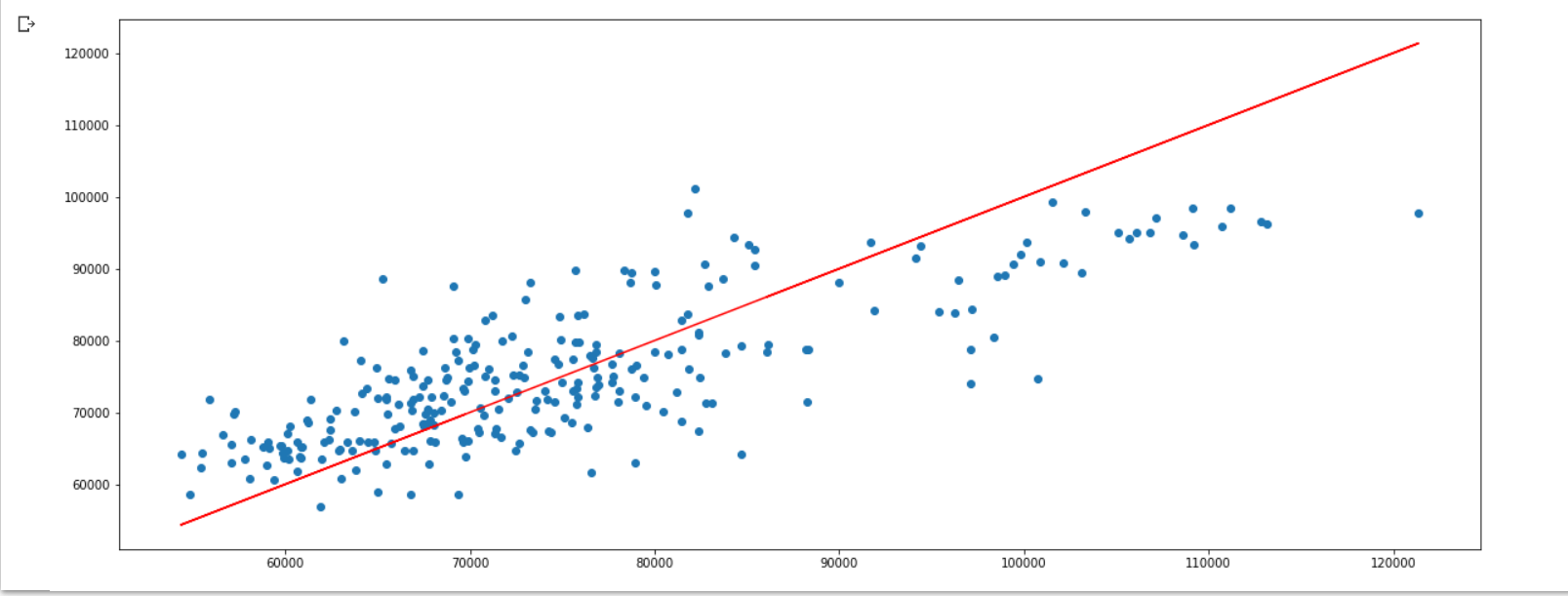
****

The linear regression model is selected from sklearn.model\_selection and the fit() method is used on a LinearRegression class object to train the model on the training data.

predict() method is used to get the predicted values of the CTC.

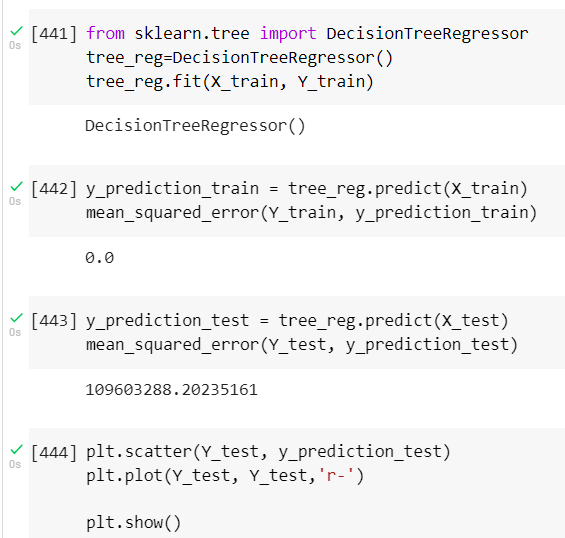
**3.7 EVALUATING RESULTS FOR LINEAR REGRESSION**

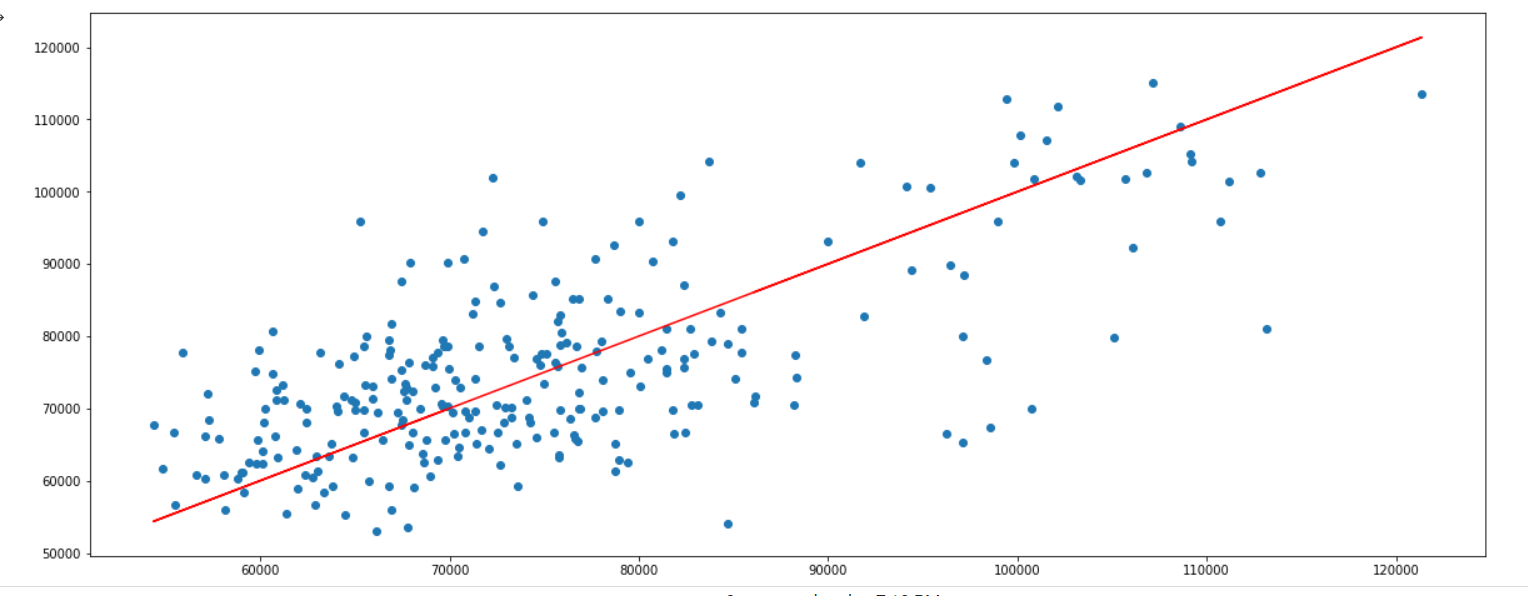
****

****

The mean\_square\_error() function is used to get the error made by the model. The MSE on training and test data is the same and as per the scatter plot, the line obtained is a good fit for the given dataset.

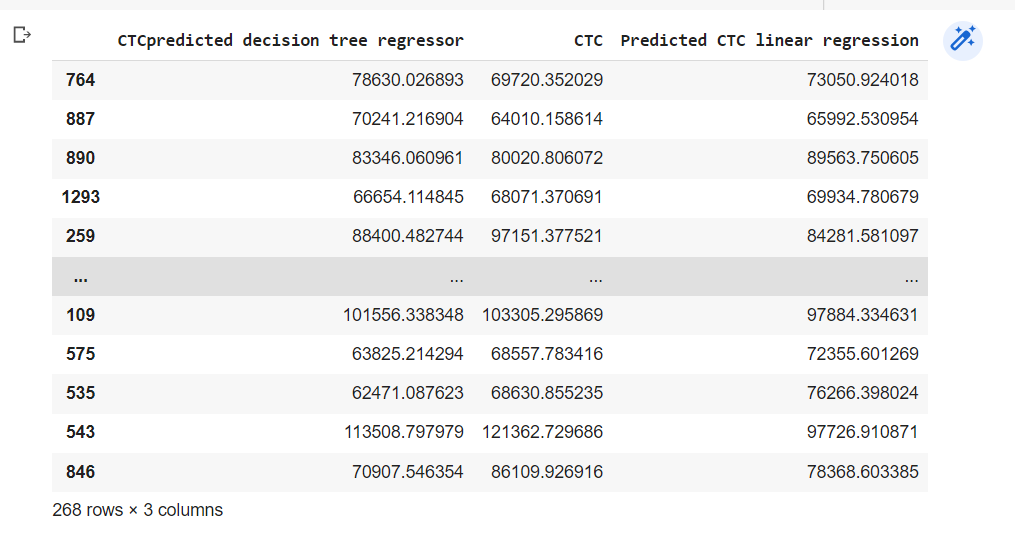
**3.8 TRAINING DESCISION TREE REGRESSION MODEL AND EVALUATING ERROR**

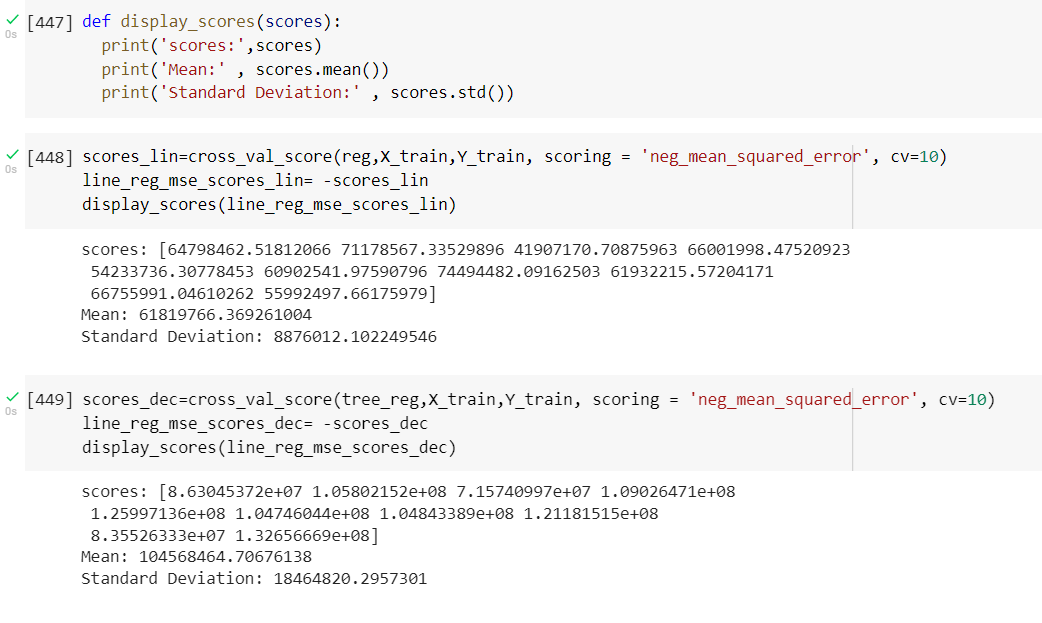
****

****

The mean squared error of this model is greater than the MSE of linear regression. Thus, Linear regression model is better than this model.

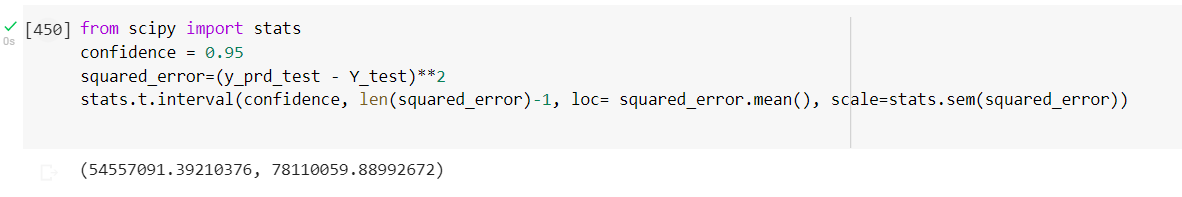
**3.8 COMPARING BOTH THE MODELS – CROSS VALIDATION**

****

****

The cross-validation scores obtained are compared and the one with the least mean is a model which has more accuracy. Here a 10 fold cross validation has been done and a separate MSE is calculated for each validation set, using which an estimated mean value of MSE is obtained.

**3.9 CONFIDENCE INTERVAL**

****

The MSE obtained via linear regression falls within the confidence interval of 95% and thus the model has a good accuracy over the data.

**CHAPTER 4: MERITS, DEMERITS AND APPLICATIONS**

**4.1 MERITS:**

There are several merits of using a salary prediction machine learning model, including:

1. Increased accuracy: Machine learning models can analyse large amounts of data and identify patterns and trends that would be difficult for humans to discern. This can lead to more accurate predictions of salaries.
2. Automation: With a salary prediction machine learning model, the process of predicting salaries can be automated, which can save time and resources.
3. Handling large data sets: Machine learning models can handle large data sets, which is beneficial in cases where there is a lot of data to analyse.
4. Personalization: Machine learning models can be trained to consider individual characteristics and qualifications, which can lead to more personalized salary predictions.
5. Speed: Machine learning models can process large amounts of data quickly, which can lead to faster and more efficient salary predictions.
6. Scalability: Machine learning models can be easily scaled to handle more data and to be used in different scenarios.
7. Cost-effective: In the long run, using a machine learning model for salary prediction can be more cost-effective than manual methods.
8. Elimination of human bias: Machine learning models can eliminate human bias by analysing data objectively, which can lead to more fair and accurate predictions.
9. Continual learning: Machine learning models can be updated and trained with new data, which can lead to improved predictions over time.

Overall, a salary prediction machine learning model can provide a more accurate, efficient, and cost-effective way to predict salaries, and also eliminating human bias and enabling continuous learning.

Top of Form

**4.2 DEMERITS:**

There are several demerits of using a salary prediction machine learning model, including:

1. Limited understanding of data: Machine learning models rely on patterns and correlations in the data, but they may not have a deep understanding of the underlying relationships and factors that influence salaries.
2. Lack of transparency: Machine learning models can be opaque and difficult to interpret, which can make it challenging to understand how the predictions are being made and to identify any potential issues.
3. Dependence on quality data: The accuracy of a salary prediction machine learning model is highly dependent on the quality of the data it is trained on. If the data is incomplete, biased, or contains errors, the predictions may be unreliable.
4. High complexity: Machine learning models can be complex and difficult to design and implement, which can make them challenging to use for organizations with limited resources or expertise.
5. Lack of domain knowledge: Machine learning models require domain knowledge to be able to select the right features, and to interpret and validate the results, which can be a challenge for organizations that don't have a deep understanding of the data.
6. Unforeseen circumstances: Machine learning models may not be able to handle unforeseen circumstances or new data that it has not seen before, leading to poor predictions.
7. High cost: Implementing and maintaining a machine learning model can be costly, particularly for small and medium-sized businesses.
8. Limited interpretability: Machine learning models can be difficult to interpret, making it difficult to understand why a certain prediction was made.
9. Ethical concerns: Machine learning models may perpetuate biases present in the data, leading to unfair and unjust predictions.

Overall, while a salary prediction machine learning model can provide accurate predictions, it also has its limitations, such as lack of transparency, dependency on data quality, complexity, cost, and ethical concerns. It is important to consider these limitations and to carefully evaluate the use of a salary prediction machine learning model in a specific context.

**4.3 APPLICATIONS**

Salary prediction machine learning models can be used in a variety of applications, including:

1. Human resources: Employers can use salary prediction models to determine fair and competitive salaries for new hires and existing employees.
2. Job posting: Job posting platforms can use salary prediction models to provide accurate salary ranges for job listings, which can help attract qualified candidates.
3. Recruitment: Recruitment agencies can use salary prediction models to predict the expected salary for a specific role and help match job seekers to appropriate job opportunities.
4. Salary negotiation: Employees can use salary prediction models to predict the expected salary for a specific role and negotiate a fair compensation package during job interviews.
5. Career development: Individuals can use salary prediction models to identify the potential salary for a specific role or career path and plan their professional development accordingly.
6. Business intelligence: Employers can use salary prediction models to identify patterns and trends in compensation and make data-driven decisions about their workforce.
7. Compensation benchmarking: Employers can use salary prediction models to compare their compensation packages to industry standards and adjust as needed.
8. Labour market analysis: Researchers and analysts can use salary prediction models to study the labour market and identify trends in compensation and employment.
9. Education: Students can use salary prediction models to identify potential salaries for specific majors and careers, which can help inform their college and career choices.
10. Government and policy: Government agencies and policy makers can use salary prediction models to study the labour market and develop policies that promote fair compensation and employment.

Overall, salary prediction machine learning models can be used in a variety of fields, including HR, recruitment, career development, business intelligence, labour market analysis, education, and government. They provide valuable insights and predictions that can be used to make informed decisions in various areas.

**CHAPTER 5: CONCLUSIONS AND FUTURE SCOPE**

**5.1 CONCLUSIONS**

In conclusion, salary prediction machine learning models are powerful tools that can provide accurate predictions of salaries based on large amounts of data. They can be used in a variety of fields, including human resources, recruitment, career development, and business intelligence. They can help employers determine fair and competitive salaries, job posting platforms provide accurate salary ranges for job listings, and individuals plan their professional development.

The model developed works with a confidence interval of 95%.

However, it is important to consider the limitations of these models, such as lack of transparency, dependency on data quality, complexity, cost, and ethical concerns. Additionally, while the future scope of salary prediction machine learning models is promising, with potential for increased accuracy, personalization, and automation, it is important to ensure that these models are transparent, fair, and unbiased.

It is important to understand the context and use cases of the predictions, and use the results in a responsible and ethical manner. Additionally, it is important to validate the predictions with real-world data and fine-tune the model accordingly. The integration of Explainable AI (XAI) and other techniques to understand and interpret the predictions can also be helpful in making fair and unbiased decisions.

**5.2 FUTURE SCOPE**

The future scope of salary prediction machine learning models is likely to be influenced by advances in technology and the increasing availability of data. Some potential developments include:

1. Improved accuracy: With advances in machine learning techniques, such as deep learning and reinforcement learning, salary prediction models will likely become more accurate in predicting salaries.
2. More robust data: With the growing availability of data from a variety of sources, such as social media and job postings, salary prediction models will have access to more robust data sets, which can lead to more accurate predictions.
3. Personalization: With the growing availability of data on individual qualifications and characteristics, salary prediction models will likely become more personalized, providing more accurate predictions for specific individuals.
4. Explainable AI: The development of Explainable AI (XAI) will make it possible to understand how the model is making predictions and identify any potential biases.
5. Integration with other data: Salary prediction models will likely be integrated with other data, such as performance data, to provide a more comprehensive view of an employee's potential salary.
6. Automation: With the integration of Natural Language Processing (NLP) and other technologies, the process of collecting and analysing data will become more automated, which can lead to more efficient and accurate predictions.
7. Predictive modelling for specific industries: Salary prediction models will be developed for specific industries, to better understand the salary trends in that industry.
8. Predictive modelling for emerging roles: As the job market evolves, new roles will emerge, and salary prediction models will adapt to provide predictions for these new roles.
9. Predictive modelling for remote work: With the increasing trend of remote work, salary prediction models will adapt to provide predictions for remote work specific roles.

Overall, the future scope of salary prediction machine learning models is promising, with potential for increased accuracy, personalization, and automation. However, it is important to consider the ethical implications of these models and ensure they are transparent, fair, and unbiased.



**REFERENCES**

[1] Internation journal of Advanced Scientific Research and Engineering trends – Volume 6, Issue 5 , May 2021, ISSN (Online) 2456-0774

[2] Andreas Mullar, Introduction to Machine Learning using Python: A guide for data Scientist, in OReilly Publisher, India.

[3] 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>

[4] 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.

[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](https://doi.org/10.10.1016/C2019-0-%2003019-1)

[7] 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.